A Swedish Grammar for Word Prediction

Ebba Gustavii and Eva Pettersson
{ebbag,evapet}@stp.ling.uu.se

Master's thesis in Computational Linguistics
Språkteknologiprogrammet
(Language Engineering Programme)
Uppsala University · Department of Linguistics

25th June 2003

Supervisor:
Anna Sågvall-Hein, Uppsala University

Abstract

Word prediction may be of great help in situations where text entering is cumbersome due to a physical or a cognitive disability, or due to the input device being small. Traditionally, word prediction has solely been based on statistic language modelling, but lately knowledge-based approaches, including the use of grammatical language descriptions, have entered the arena. By making use of grammar rules, the accuracy of the prediction suggestions is expected to increase, and the word predictor to give a more intelligent impression.

We have defined and implemented a Swedish grammar for the FASTY word predictor. The grammar rules, defined in terms of traditional grammatical functions, are expressed in the procedural UCP-formalism. The grammar functions as a grammar checking filter, reranking the suggestions proposed by a statistic n-gram model on the basis of both confirming and rejecting rules. What structures to cover has been decided in accordance with an investigation on what syntactic errors are most frequently produced by the statistic model. The investigation led us to prioritize rules for handling word order in the main clause, agreement within the noun phrase, verb inflection and prepositional phrases.

A preliminary evaluation of the grammar module was carried out, using Keystroke Saving Rate (KSR) estimations. The results showed only a slight improvement in KSR when adding the grammar module to the system, as compared to using only the statistic model based on word form bigrams and part-of-speech tag trigrams. A list length of one suggestion gave a larger improvement, than a list length of five, indicating that the strength of the grammar module lies in the reranking of already displayed suggestions, rather than the addition of new suggestions to a long list of suggestions.

Contents

List of Tables

List of Figures

Preface

The two authors behind this thesis have worked in close co-operation, as regards all practical aspects of developing the grammar. Concerning the report, Gustavii's main focus has been to present the FASTY word predictor (chapter 4), grammar checking in FASTY (chapter 5) and the error classification (chapter 6). Pettersson has focused on describing word prediction in general (chapter 2), language modelling (chapter 3) and the grammar rules for handling noun phrases (section 7.2). Both authors have however contributed to all chapters. Sections not mentioned above are the results of completely joint efforts.

First and foremost, we would like to thank our supervisor, Anna Sågvall-Hein, for great support and guidance throughout the completion of this thesis. We would also like to thank all participants of the FASTY project, in particular Mikael Wiberg, for his never ending stream of insightful comments. We are grateful to the Department of Linguistics at Uppsala University, for a pleasant working athmosphere and the thousand cups of coffee. We further wish to give our special thanks to Malin Wester, Gustav Öquist and Mikael Wiberg (again!), for coping with the many discussions held in our shared working room.

This work was supported by the European Union within the framework of the IST programme, project FASTY (IST-2000-25420).

Chapter 1

Introduction

Word prediction may be of great help in situations where text entering is cumbersome due to a physical or a cognitive disability, or due to the input device being small. Traditionally, word prediction has solely been based on statistic language modelling, but lately knowledge-based approaches, including the use of grammatical language descriptions, have entered the arena.

By making use of grammar rules, the accuracy of the prediction suggestions is expected to increase, and the word predictor to give a more intelligent impression. Sofar, no attempts have been made to make use of grammar in a Swedish word prediction system.

1.1 Purpose

The aim of this thesis is to define a Swedish grammar for partial parsing to be incorporated in the FASTY word predictor. The grammar is to function as a grammar checking filter, reranking the suggestions proposed by the statistic model on the basis of both confirming and rejecting rules. The grammar rules are to be expressed in a procedural formalism compatible with the Fasty version of the Uppsala Chart Parser (FastyUCP), and will be defined in terms of traditional grammatical functions. The aim is to cover contexts where the statistic model gives syntactically irrelevant suggestions.

A statistically based word prediction system may violate the syntax in an enormous number of ways and it is impossible to formulate a grammar that captures them all. In order to find out what kinds of syntactic constructions to prioritize in the development of the grammar, we will carry out a manual inspection of suggestions proposed by the statistic model, whereafter the syntactically inappropriate suggestions will be classified and the relative frequencies of the error types will be computed. The final grammar description will be evaluated by means of keystroke savings, estimated by simulating the production of texts both with and without the addition of the grammar checking module.

1.2 Outline of the thesis

The thesis is outlined as follows: In chapter 2, the concept of word prediction is introduced and prospective users are suggested. In chapter 3, we will discuss the basis for word prediction; statistic as well as knowledge-based language modelling. In chapter 4, the FASTY word predictor will be presented, and in chapter 5, we will give a description of how the FASTY grammar checking module works. In chapter 6, we will describe the manual inspection of prediction suggestions proposed by the statistic language model. The error classification will be explained, and the frequency order of the error types will be given. In chapter 7, the main chapter of the thesis, the grammar rules will be described. In chapter 8, a small evaluation of the grammar module will be presented, and in the final chapter the thesis is summarized and future developments are suggested.

Chapter 2

Word prediction

In this chapter the concept of word prediction is introduced and prospective users and domains are discussed.

2.1 What is word prediction?

Word prediction is about guessing what word the user intends to write for the purpose of facilitating the text production process. Sometimes a distinction is made between systems that require the initial letters of an upcoming word to make a prediction and systems that may predict a word regardless of whether the word has been initialized or not. The former systems are said to perform word completion while the latter perform proper word prediction. (Klund and Novak 2001)

Whenever the user types a letter or confirms a prediction, the system updates its guesses taking the extended context into account. The size and nature of the context on which the predictions are based, varies among different systems. While the simplest systems only take single word form frequencies into account, thus not at all making use of the previous context, more complex systems may consider the previous one or two word forms and/or the grammatic categories. Yet more complex systems combine these methods with other strategies such as topic guidance, recency promotion and grammatical structure. The language models behind the various techniques will be further examined in chapter 3.

There are two main strategies for displaying the prediction suggestions. One strategy is to present the most probable prediction directly in the text to be produced whereas the other is to present a set of probable word forms in the form of a ranked list. Both strategies have their advantages and the system designer's choice is ultimately depending on the purpose of the system. In existing word prediction systems, the direct-insert approach is however rarely used. (Carlberger and Hunnicutt n.d.)

2.2 Who benefits from word prediction?

The primary users of word prediction systems have traditionally been physically disabled persons. For people having motor impairments making it difficult to type, a word prediction system would optimally reduce both the time and effort needed for producing a text, as the number of keystrokes decreases. Further the selection of predictions typically requires a smaller set of keys than ordinary typing, which is useful for individuals whose motoric disabilities make it hard to manage a complete set of keys.

The usage of word prediction, however, imposes a high degree of perceptual and cognitive load since it calls for the user to shift focus from the text in progress and to read and choose from several alternatives (if the list strategy is chosen). Sometimes the total cognitive load is claimed to cost more in terms of communication rate [1] than is gained in keystroke savings. However, according to (Klund and Novak 2001), such conclusions are usually based on studies that have been carried out on first-time users

[1] Number of characters or words entered by the user per time-unit. (Fazly 2002)

and it is likely that frequent use of word prediction systems improves the user's performance and lessens the cognitive load. Further, regardless of whether there is a gain in communication rate, the advantage of less keystrokes may still save physical effort for the user. By enhancing the communication efficiency [2], the user may be able to engage in communicative activities for longer periods of time.

In recent years, word prediction has moreover proved beneficial for persons with dyslexia or other language impairments by helping with spelling and grammar structures. A study carried out by (Laine and Bristow 1999) shows that word prediction assisted students with impaired language skills in entering appropriate word forms along with enhancing the overall word fluency. The users engaged a larger vocabulary and produced more text. Some word prediction systems make use of speech synthesis, whereby the predictions are read to the user instead of being visually displayed. This may be of use for dyslectic persons that have difficulties in recognizing an intended word form among the prediction suggestions. (Carlberger 1997)

Lately, word prediction techniques have entered new domains. In particular, it has come to be an integrated part of many cellular phones as an assisting aid in the popular text message service, SMS. Since the essence of cellular phones lies in their mobility their size has to remain small. The text input device is therefore shared with the 9 standard keys for dialing numbers. That way, a key along with a number, represents three or more characters. Without predictive techniques the user manually has to disambiguate the keys by pressing the same key a different number of times. As a consequence, the user often has to make several more keystrokes than there are letters to be entered. To get around the cost in keystrokes Tegic's *T9 system* (Text on 9 keys) employs automatic letter disambiguation so that the user does not have to press a key more than once to select a character. The letters are then disambiguated in the context of each other. From all possible letter combinations, the system picks the one corresponding to the most frequent word. [3]. In a way T9 is not a word prediction system, since it does not predict words on the basis on previous context. Further, it does not make its guess until it has access to the fuzzy set of letters comprising the word. However, the objective to assist the user in text production by guessing her intentions, as well as the strategy to base the guesses on a statistic language model, is perfectly in line with traditional word prediction systems. There are attempts to enter truly predictive strategies into the domain. *POBox* (Predictive cOmposition Based On eXample) is a word prediction technique that has been developed with the explicit goal to be integrated in small hand held devices; cellular phones as well as PDA's. (Masui 1999)

Besides from being useful in contexts where text entering is particularly cumbersome (due to the user being physically disabled or the keyboard being small-sized), word prediction may save effort for a non-disabled person using a full-sized keyboard. This has been noticed by the developers of the open source word processor *OpenOffice*, which along with standard word processing features provides word completion. [4].

Further, word prediction techniques may be part of other NLP applications, such as part-of-speech taggers, context-sensitive spelling correction and systems making use of word-sense disambiguation. (Fazly 2002)

[2]Number of movements needed to enter a character (Fazly 2002)

[3]For further informatin on Tegic, please visit: http://www.t9.com

[4]For further information on OpenOffice, please vivit: http://www.openoffice.org

Chapter 3

Language modeling

For a word prediction system to make predictions it needs access to some kind of language model. A language model is an approximative description that captures patterns and regularities present in natural language and is used for making assumptions on previously unseen language fragments. There are basically two kinds of language modeling; statistic and knowledge-based. A statistic language model is based on data automatically extracted from large corpora. A knowledge-based language model, on the other hand, relies on manually defined linguistic generalizations, such as grammar rules. (Rosenfeld 1994)

Generally, word prediction systems make use of statistic language models, as described in section 3.1 below. As will come clear in section 3.2 though, some systems supplement these models with a knowledge-based source of information.

3.1 Statistic language modeling

Statistic language models have generally been the main source of information used in word prediction systems. The statistics included in the systems varies from single word frequencies to part-of-speech tag n-grams.

3.1.1 Word frequencies

Most of the earliest word prediction systems, developed in the beginning of the 1980s, are based on unigram statistics, taking only the frequency of single words into account when trying to predict the current or next word. A system that is solely based on unigram statistics will always come up with the same prediction suggestions after a particular sequence of letters or after a blank space, irrespective of the preceding context. The lack of context considered entails that the suggestions often will be syntactically and semantically inappropriate. (Fazly 2002) Still this language model has proved useful in systems, such as the *Predictive Adaptive Lexicon* (PAL) (Fazly 2002).

3.1.2 Word sequence frequencies

In practice, the probability of an upcoming word is influenced by the previous context, something which has been explored and incorporated in later systems, such as *WordQ*. (Fazly 2002) When using n-grams (with $n>1$) as a basis for the prediction process, the probability of a certain word given the previous word(s) is computed. Incorporating an n-gram model, some syntactic and semantic dependencies are captured, since words that frequently co-occur are promoted. This is in particular true for languages like the European ones, with a rather fixed word order. (Jelinek 1991)

Still, some syntactically and semantically inappropriate suggestions are likely to be displayed as well, since the n-grams are limited to a local context consisting of a fixed number of words. Usually no more

than two or three words are considered, as a great amount of text is needed to be able to find reliable enough n-grams of a larger size. (Kronlid and Nilsson 2000)

No matter how much data is used as a basis for the n-gram extraction there will always be a problem of sparse data - most n-grams will have low frequencies and many possible word sequences will not be attested at all in the training material. *Smoothing* is the process of improving the reliability of the n-gram frequencies by reducing the number of zero probabilities. A common smoothing method is the linear interpolation algorithm, taking into account not only the frequencies of the largest n-grams used, but also the frequencies of the smaller ones. For instance, for a system that relies on trigrams as the largest context, the probability of a word to appear is computed as the sum of the weighted probabilities of that word to occur according to the trigram, the bigram and the unigram statistics. This way a certain word may occur in the prediction list even though it is not part of an attested trigram, provided that it is present in the lexicon and is given a unigram and/or bigram value. (Rosenfeld 1994)

3.1.3 Part-of-speech sequence frequencies

In order to generalize the n-gram model, some systems make use of n-grams of part-of-speech tags instead of word forms. Using this method the probability of the current tag given the previous tag(s) is computed. Since part-of-speech tags capture a lot of different word forms in one single formula it it possible to represent contextual dependencies in a smaller set of n-grams. One major advantage of making use of part-of-speech frequencies is thus that a larger context may be taken into consideration. There is a loss in semantics though, since the tags only tell, for example, that a verb is likely to follow a noun and not what particular verbs are typically connected with what nouns.

Another drawback is the problem that arises when new words are encountered and ought to be added to the lexicon, since the system needs to know the syntactic category of a word to incorporate it in the statistics. A way to solve this is to have the system interpret the prefixes and suffixes of the new word in order to find out the most probable part-of-speech. Still, there will probably occur words without any prefixes or suffixes fitting the morphological heuristics. These words may then simply be given the most probable tag according to the overall statistics. (Cagigas 2001)

The later version of *Prophet*, developed by (Carlberger 1997), uses both word bigrams and part-of-speech trigrams in the prediction process.

3.2 Knowledge-based language modeling

While n-gram based systems never consider contexts exceeding a limited number of words, a knowledge-based system may take an arbitrarily large sentence fragment into consideration. Thus a well-defined knowledge-based language model would have the advantage of being able to capture long-term dependencies not covered by any of the statistic language models described above. If grammar rules are used, the number of syntactically inappropriate suggestions would be reduced, leaving more room for possibly intended word forms to appear in the prediction list. Further, a low rate of inappropriate prediction suggestions imposes a lower cognitive load, enhancing the comfort of the user. (Wood 1996) It has also been stated that dyslectics would benefit from a system where no confusing suggestions are displayed. (Carlberger and Hunnicutt n.d.)

3.2.1 Existing prediction systems using grammar rules

Some word prediction systems have incorporated grammar rules in the prediction process to enhance the accuracy of the prediction suggestions. In the following sections, some of these systems will be described.

The Intelligent Word Prediction Project

The *Intelligent Word Prediction Project* has delivered a prototype word predictor for English that bases its predictions on syntax. The grammar is implemented as an augmented transition network (ATN) in which the transitions are marked with probabilities. Each word form in the lexicon is attributed with information on its part-of-speech distribution in some training corpus. By combining these statistics the most probable next words are calculated and presented to the user. During the parsing all possible analyses are kept in separate branches, so as to cope with lexical and syntactic ambiguity. As far as the authors are aware there are unfortunately no detailed description of the system nor any published test results.

Windmill

Windmill is another word prediction system for English that incorporates grammar rules with simple word form frequencies. The system has a modular architecture, separating different knowledge sources from each other and from the system engine, whereby the portability to other languages becomes easier and the effect of different prediction strategies is possible to assess.

The grammar formalism used is an augmented context free grammar and the parser attempts to make a complete syntactic analysis of the sentence fragment entered so far using a top-down strategy. Windmill does not attach probabilities to the rules, as was done in the Intelligent Word Prediction Project. Rather Windmill tries to find all possibly appropriate syntactic categories for the next word, and then divides the lexicon into one subset containing appropriate word forms and a second subset of word forms not confirmed by the grammar. The unigram statistics is then used to rank the subset of approved words. The grammar checking mechanism in Windmill is hence defined so as to regulate what constructions are acceptable (as opposed to negatively defined grammar checkers) whereby only word forms confirmed by the grammar may be presented to the user. In order to make the system more robust there is a fall-back strategy implemented which makes the system solely rely on statistic predictions whenever there is no possible syntactic parse of the input sentence.

Windmill has been evaluated in relation to three test texts; an article, a scientific paper (about Windmill) and a set of transcripts of short conversations. The evaluation is based on keystroke saving rate (KSR), hit rate[1] and the position of the desired word in the prediction list. All measures confirmed that the best results were obtained by using statistic data in combination with the grammar rules. When using only word form frequencies the average KSR obtained was 50.31% whereas the addition of the grammar module enhanced the figure to 55.10%. Windmill was however further tested on an informally written write-up, known to expose syntactic structures out of scope of the grammar rules. When the grammar module was added the KSR did not increase but instead marginally dropped. When evaluating this test text the fail-safe strategy turned out to make a crucial difference; if it was turned off the grammar module made the KSR drop from 53.3% to 28.8%. (Wood 1996)

A word prediction system for Spanish

A Spanish word prediction system has been developed by Palazuelos-Cagigas. The system includes several modules in the prediction process. In the first stage the most probable syntactic categories of the upcoming word are estimated with respect to part-of-speech bigram and trigram statistics. In the next stage the most probable word forms belonging to any of these categories are selected. The final step incorporates probabilistic context free grammar rules, which output the probabilities of the next syntactic category, whereafter all measures are weighted together. The system thereby functions in a way similar to the Intelligent Word Predictor, albeit with an extended n-gram model.

As with the Windmill system the modules within the system are separated from each other, making it easier to adapt the system to other languages and also to test the modules one at the time or in different combinations. Evaluation tests showed that the grammar rules slightly reduced the percentage of

[1] The frequency with which the intended word appears in the prediction list before any letters have been typed (Wood 1996)

keystroke savings compared to the use of part-of-speech trigram statistics. According to (Cagigas 2001), this was due to the test text containing a lot of references and very long sentences, that was out of scope of the grammar, thus causing the parser to stop. New evaluation tests were carried out where the test text was somewhat adapted to the grammar. The adaptation included eliminating references, abbreviations containing a full stop (otherwise interpreted as the end of a sentence), text between parenthesis, lists and enumerations. Further, sentences longer than 50 words were rewritten. Tests on the adapted text showed that incorporating grammar rules in the prediction process gave somewhat better results than only using the full n-gram model, improving the KSR with approximately 0.5 percent units. (Cagigas 2001)

Compansion

The prototype system developed within the *Compansion Project* is not a word prediction in the traditional sense but share several features with the systems previously described. Compansion is a syntactically and semantically based system with the purpose of reducing text entry effort for disabled users. But instead of trying to predict an upcoming word it tries to infer the syntactic structure from input uninflected content words and to generate a corresponding well-formed sentence. For instance, if the user enters APPLE EAT JOHN an attempted output would be THE APPLE IS EATEN BY JOHN. The system works in several steps: first it gives the input a syntactic analysis, second it assigns semantic roles to the analyzed words and finally it generates a well-formed English sentence based on the semantic representation. The system is still at a prototypic stage and a long term goal is to integrate the Compansion system with the Intelligent Word Prediction System described above.[2](Demasco and McCoy 1992)

3.3 Adaptation heuristics

Irrespective of the language model used, there are several techniques that can be added to the system in order to make it more suitable for a specific user. One way is to invoke *short-term learning*, whereby the system adapts to the current text, e.g. by recency promotion or topic guidance. *Recency promotion* means that a word that has already occurred in the text will be given a higher probability and thus more likely appear in an upcoming prediction list than before. Usually such algorithms dynamically assign higher probabilities to words more recently entered in the text, so that it does not only take into account what words have been typed but further, how recent. (Carlberger and Hunnicutt n.d.) *Topic guidance* is a way of adapting the predictions to the overall subject of the current text. To do so, the general lexicon is complemented with domain specific lexicons containing words that are frequently occurring within certain domains, though not very common in general. (Lesher and Rinkus 2001)

Another technique is to make use of *long-term learning*, whereby the system adapts to the user, considering not only the current text but also previous texts produced by the user. Long-term learning may for instance involve adding new words to the lexicon, whenever the user types words unknown to the system. This way words that are rarely occurring in general texts, thus not covered by the lexicon, but which are frequently used by a certain user, are given the possibility to be predicted in the future. However, this method involves the risk of adding misspellings and ungrammatical sequences to the lexicon. To avoid this, one strategy is to only add a word to the lexicon if it has been typed several times. A more refined method is to use a spell checker or a grammar checker in the process. Another problem with adding words is to decide what values to assign to them. If the system relies on frequency information, an appropriate probability value in relation to the other words in the lexicon is called for and if grammatical information is included, the system needs to decide on the grammatical category of the word. (Sjöberg 2001)

[2]For further information on the Compansion Project, please visit: http://www.asel.udel.edu/natlang/nlp/compansion.html

Chapter 4

FASTY word predictor

The Swedish grammar, that is the main topic of this thesis, has been developed within the framework of the FASTY project. FASTY is an EU-funded project within the Information Society Technologies. The project aims at developing a complete word prediction environment for supporting disabled people in the text generation process. The FASTY word prediction system is developed for four languages; German, French, Dutch and Swedish, but an explicit goal has been to make future inclusion of other languages possible, imposing a strictly modular architecture of the system.

In section 4.1 below we will present the language model used within the system. In section 4.2 the collection of the Swedish language data required will be described and in section 4.3 the lexicon used is introduced.

4.1 FASTY language model

The FASTY language model is based on several sources of information, implemented as clearly separated modules, some of which are state-of-the-art and thereby guarantees a basic performance and some of which are carrying innovative features.

4.1.1 Basic modules

The basic modules in the FASTY language component rely on a statistic language model that is defined in close resemblance to the language model described in (Carlberger 1997). Some adaptive heuristics have further been implemented: the user is provided the posssibility to create user specific dictionaries and to combine these with the general dictionary. There is however no automatic means to attribute unknown words with the morpho-syntactic information required by the grammar module and the part-of-speech tag model.

Word form bigrams are used as a base, and as is customary, the bigrams are supplemented with word form unigrams. The probabilities obtained from both uni- and bigrams are weighted together using a standard linear interpolation formula.

The statistic language model is further based on trigrams of part-of-speech tags. The same interpolation formula as for the word form n-grams, is used to include probabilities from smaller part-of-speech tag n-grams. This information is added to the word model to estimate a combined probability of the next word, given both the previous word and the two previous tags. The probability distribution for different part-of-speech tags given a word form is further included, since a word form may be ambiguous and adhere to more than one part-of-speech.

4.1.2 Innovative modules

There are two innovative modules within the FASTY language component that interact with the basic modules in order to enhance the prediction performance; *grammar checking* and *compound prediction*. As far as the authors are aware, there are no existing word prediction systems involving grammar rules for the target languages, whereby the grammar checking module may be regarded an innovation. Since the grammar component is the main topic of this thesis there are several chapters describing its function and development. Thus only the compound prediction module will be touched upon in this section.

In three of the FASTY target languages: German, Dutch and Swedish, compounds constitute a group of words that is particularly hard to predict within a word prediction system. In these languages compounds can be productively formed to fill a contextual need. It is of course impossible to predict such a word formation by means of traditional n-gram frequency counts, since lots of productively formed compounds will not have been attested in the training material. On the other hand, compounds tend to be long words, which means that successful prediction would save a great deal of keystrokes. Within the FASTY language model, compounds have hence been given a special treatment. Instead of trying to predict a productively formed compound as a whole, its parts will be predicted separately. More specifically, the current implementation supports the prediction of right-headed nominal compounds, since these, according to a corpus study of German corpus data, are by far the most common.(FASTY 2003)

The split compound model provides two quite different mechanisms for predicting the respective parts of a compound, i.e. *modifier prediction* and *head prediction*. Since the system has no means of knowing when a user wants to initiate a compound, the prediction of modifiers is integrated with the prediction of ordinary words. If the user selects a noun that has higher probability of being a compound modifier, the system assumes this use is intended and starts the prediction of the head part instead of inserting the default white space after the selected word.

The head of a compound determines the syntactic behavior and the basic meaning of the compound as a whole. Hence, we may expect a productively formed compound to appear in the same type of contexts as the head would if functioning as an independent word. When predicting the head, the system therefore makes use of the word preceding the modifier, as if the modifier were not there. Let us assume that the user has written *en god äppel* (a tasty apple), and intends tocompound write *en god äppelpaj* (a tasty apple pie). When searching for possible compound continuations, the system will then search for bigrams with the first position held by *god*, and if the training corpora contains instances enough of the sequence *god paj*, *paj* is suggested as a possible head of the compound. The head prediction model gives precedence to words that functioned as heads in many compounds in the training material. According to studies of German and Swedish compounds (carried out within the FASTy project), some words occur much more often in compounds as heads, than do other words. A secondary feature that has been implemented, is the semantic relation between the modifier and the head. It is assumed that the semantic class of the modifier may give hints on the semantic class of the head.

4.2 Data collection

The Swedish word form n-grams have been extracted from a training corpus comprising approximately 19 million word tokens. The corpus constitutes 90% of the UNT-corpus, which was compiled at the Uppsala University and contains newspaper articles published in Upsala Nya Tidning during the period of 1995 to 1996. (Dahlqvist 1998) The remaining 10% were left aside to form potential test material. Before extraction, headings, author signatures and articles in foreign languages were removed from the training corpus. From the extracted data, n-grams containing non-alphabetical characters were eliminated (except for punctuation marks) as well as all n-grams with a frequency below two. The n-gram frequency tables were further modified to suit the split compound prediction module. For instance, all unigrams containing a compound were split into two unigrams representing the modifier and the head respectively.

This was done by means of shared project software in conjunction with the Scarrie compound analyzer.[1] The semantic classes, also contributing to the compound prediction, have been approximated by automatically extracting information on how words co-occur in the training corpus, and the probabilities of pairs of semantic classes to join in compounds have been estimated on the basis of how they co-occur as modifiers and heads in the attested compounds. These statistics have also been collected using shared project software in accordance with a model proposed by (Baroni, Matiasek and Trost 2002).

The part-of-speech statistics were extracted from the hand-tagged, balanced Stockholm-Umeå Corpus (SUC) comprising approximately one million word tokens.[2] In order to make the tags compatible with the part-of-speech information provided by the lexicon, some conversions had to be made and the fine-grained subcategorization were made more general in both models.

4.3 Lexicon

The part-of-speech tag model requires information about the possible part-of-speech tags of each word form. Further, the grammar module, soon to be described, bases its analyses on a morpho-syntactic description of the input word forms. Thus, the FASTY language model needs some kind of lexicon, providing for all relevant information. The Swedish lexicon used has been extracted from the Scarrie lexical database (Olsson 1999) which is corpus based and contains word forms attested in the UNT-corpus or the SVD-corpus. The derived lexicon contains approximately 200,000 word forms annotated with morpho-syntactic features.

[1] For more information on the Scarrie project, please visit : www.hltcentral.org/projects/detail.php?acronym=SCARRIE

[2] For further information on SUC, please look at: www.ling.se.su/DaLi/projects/SUC/Index.html

Chapter 5

Grammar checking in FASTY

The overall motivation for the grammar module is to enhance the accuracy of the prediction suggestions. The grammar module does not by itself generate any prediction suggestions but filters the suggestions produced by the n-gram model so that the grammatically correct word forms will be presented to the user prior to any ungrammatical ones.

5.1 Grammar checking as a filter of suggestions

Input to the grammar module is a ranked list of the most probable word forms according to the n-gram model. The grammar module parses the sentence fragment entered so far, and assigns a value to each word form in the input list based on whether it is confirmed by (grammatical), turned down by (ungrammatical) or out of scope of the grammar. Based on those values, the word forms are then reranked whereby the grammatical suggestions are ranked the highest and the ungrammatical the lowest. Since only a subset of the reranked prediction suggestions will be presented to the user, the lowest ranked word forms will generally not be displayed. This way the ungrammatical suggestions will hopefully not be presented and the user will not be disturbed by grammatically impossible suggestions. In addition, there will be more room for possibly intended words in the list to be presented.

5.2 Parsing engine and chart scanning

The grammar module, using a modified version of the Uppsala Chart Parser (FastyUCP), is based on partial parsing, which means that the parser does not need the context of a whole sentence to interpret a constituent. Instead sentence fragments of any length may be analyzed. This is of course an indispensable requirement when a prediction suggestion is analyzed, since the parser will only have access to the left context of the sentence. It also makes it possible to analyze a specific constituent, such as a noun phrase, irrespective of its surrounding context. [1] UCP employs a procedural formalism, quite similar to imperative programming languages, and all chart manipulations are made explicitly from the grammar rules. The basic data structure is the feature structure which mainly is operated upon by means of the unification operator.(Weijnitz 1999).

Chart parsing can be described in terms of a directed, acyclic graph in which the vertexes represent the order between the constituents and the edges represent the constituents themselves. There are two types of edges; passive and active. A passive edge spans a completely analyzed constituent, while an active edge only spans the first part of a constituent (which may be empty) and is searching for the appropriate continuation to make it complete. For the purpose of illustration, assume there is an active edge holding a preposition and that it is striving to form a prepositional phrase by adjoining a nominal

[1] The use of UCP for partial parsing has previously been explored within the grammar checking project SCARRIE, for more information see (Sågvall Hein and Starbäck 1999).

constituent. If the active edge then meets a passive edge holding a nominal constituent a new passive edge will be created spanning both edges.

The information held by a passive edge is a feature structure representing the spanned constituent. An active edge, on the other hand, carries information not only about the constituent to be built but also restrictions on the passive edges to be searched for. An active edge may be defined to search for any kind of segment, so even erroneous ones, whereby a passive edge may be created containing a feature indicating some sort of syntactic error. (Weijnitz 1999)

When all combinations of active edges meeting passive ones have been exhausted, the execution terminates and the chart is scanned. In particular the edges spanning the prediction suggestion, the rightmost word, are examined. If the edge contains an error feature the prediction suggestion is turned down by the grammar. If, on the other hand, it does not contain any error features and the suggested prediction is at least the second constituent of the spanned segment, the suggestion is confirmed by the grammar. The latter requirement is called for since the parser is partial and words that initiate phrases would be confirmed everywhere if also edges containing only one constituent were included. The requirement is fulfilled irrespective of whether the suggested word form is the second constituent at a basic level or forms a higher level segment which in turn is a second constituent. For the sake of clarity, suppose we are to evaluate a single noun, that by itself constitutes a noun phrase. The noun then, being the first constituent of the noun phrase will not be considered confirmed. If the noun however, is preceded by a preposition, the noun phrase formed will be the second constituent of a prepositional phrase and the noun itself will be confirmed.

For suggested word forms spanned by several edges, the longest edge takes precedence. If there is more than one edge of the same length, where at least one of them does not contain an error feature, the suggested prediction will not be turned down by the grammar. In other words, the confirming edges take precedence over the rejecting ones. This way an error rule may be quite generally defined taking only a smaller context into consideration, while an accepting rule may be invoked taking a larger context into consideration confirming the same segment. For example an error rule permitting two finite verbs to occur after each other would be suitable in many contexts but an accepting rule could be used to allow two finite verbs at a clause border.[2]

5.3 Manipulating the chart

Since the prediction suggestions confirmed by the grammar will be placed at the top of the ranked prediction list while pushing down other suggestions, it is of vital importance that the sentence fragments covered by the accepting rules have a thorough grammar description. Accordingly great care must be taken when writing an accepting rule. As already stated an accepting rule may sometimes be defined for the purpose of neutralizing the effect of an error rule. It is however not always the case that the context of the neutralizing rule is thoroughly covered. For instance, a consequence of the neutralizing rule allowing two consecutive finite verbs in the context of a clause border is that a finite verb will always be promoted after a finite verb supposedly ending a clause. This means that other possible continuations, like adverbials and conjunctions, will be pushed down, something which may not at all be desired. To come to terms with this, a kind of truly neutralizing rule is called for; that is, a rule that neutralizes the effect of an error rule while at the same time, not confirming the suggested word form. Since the system does not provide this possibility we have worked out a strategy simulating the behavior of a neutralizing rule. Pinning it down, an edge containing an error is disregarded as soon as there is another edge, as long or longer, that does not contain an error feature. For a suggested prediction to be confirmed though, it is not only required for it to be spanned by a long enough edge without the error feature, but also that the word form is constituent number two (or higher). The solution is thus, to write a grammar rule that somehow creates long enough edges without making the suggested prediction constituent number two. At first

[2]The general ideas behind the chart scanning are taken from (Starbäck 1999), whereas the relevance of constituent number, proposed by Anna Sågvall-Hein, is new for the FASTY implementation.

glance this seems to be an impossible task. As it happens, the constituent number is not automatically enumerated, but is controlled by the grammar rules by a specific operator, and is thus at the hands of the grammar witer. Whenever we want to neutralize an error without confirming a constituent, we therefore omit to invoke the enumerating operator.

Summing up, a suggested word form is treated according to the following:

- A word form is promoted when it is (at least) the second constituent in a spanning edge with no error features, and the word is not spanned by a longer edge containing an error feature.
- A word form is suppressed when the longest edge by which it is spanned, contains an error feature.
- A word form is neither promoted nor suppressed, if there is no edge spanning the word, or if the longest edge contains no error feature and does not assign the word in question (or a higher level phrase of which it is part) a constituent number above 1.

Another difficulty encountered is to confirm the first constituent of a segment within a larger structure, such as a noun phrase initiator in a prepositional phrase. As already mentioned a noun phrase is to be confirmed after a preposition. Since the active edge holding the preposition is searching for a complete noun phrase it will not expand to span only an initiator of a noun phrase. The initiator will therefore not be constituent number two in any edge and thus not promoted. In order to get such words promoted supplementary rules are called for, that create edges spanning, for instance, the preposition and a definite article. Those rules hence function as a kind of bridges between constituents.

Chapter 6

Error classification and class frequencies

A word prediction system based on an n-gram model may violate the syntax in an enormous number of ways and it is impossible to formulate a grammar that captures them all. It is therefore desirable to give priority to the most common kind of syntactic errors produced by the n-gram model. Unfortunately there is, as far as the authors are aware, no previous investigation on this matter. It is plausible, however, that the n-gram model tends to make certain kinds of errors more often than others. As stated by (Rosenfeld 1996), the n-gram model suffers from two particular shortcomings:

1. The n-gram model depicts order relations and disregards the linguistic relations between constituents.

2. The n-grams only capture the local context and are insensitive to features outside their limited scope.

As follows from the first shortcoming, the n-gram model may encounter problems whenever an optional constituent comes in between two syntactically dependent units, such as a clause adverbial intermediating an auxiliary and a head verb; the contexts *Han skulle* (He should)and *Han skulle aldrig mer*(He should never again are very different from an n-gram point-of-view, yet the expectations on the next word are approximately the same.

As follows from the second shortcoming, the n-gram model may give high probabilities to suggestions that violate global constraints as long as they are in accordance with the local context. This entails that the n-gram model will have problems whenever a certain word order is frequent in one global context but erroneous in another. One such globally dependent word order in Swedish concerns the clause adverb and the finite verb. In a main clause the adverb is to succeed the verb whereas it is to precede it in a subordinate clause. We may therefore expect the n-gram model to produce erroneous suggestions in the context of finite verbs and clause adverbials.

It is however, impossible to locate all potentially problematic contexts and estimate their actual effect on the prediction suggestions without empirical data. Thus, in order to find out what kind of syntactic errors the n-gram model most frequently gives rise to, we have carried out a manual inspection of suggestions produced by the system. The suggestions were produced while a simulated user entered the first 18 sentences of *Passionsspelet*, a novel written by the Swedish author Jonas Gardell. From the collection of suggestions we singled out those that, one way or the other, violate the syntactic regulations of Swedish. The erroneous suggestions were then classified in accordance with what kind of syntactic regulations they violate.

Due to the inspection being manual, the number of sentences is fairly limited, which of course negatively influences the generality of the results. We do however, believe that the sample is large enough to expose the most frequent error types even though some error types may be overly represented and others not covered at all.

During the experiment, weights were set in accordance with the optimal results so far and the prediction list contained 5 prediction suggestions. The logged prediction lists comprise a sample of 3,883 suggestions, out of which we have judged 994 suggestions syntactically erroneous.

Below, the first two sections are devoted to the principles by which we have judged and classified the suggestions. In the third section we describe the classification results.

6.1 What is considered erroneous?

It may seem straightforward to identify suggestions that are syntactically erroneous, but in the context of a word prediction system, several decisions must be made.

When judging a suggestion we have only taken the left context into consideration, i.e. the text entered so far. From this strategy two things follow: first, we have not taken the actual text as key for what is correct, rather we have relied upon our own linguistic feeling. Secondly, we have not used the right context to resolve lexically or syntactically ambiguous left contexts, instead we have kept all possibilities open and considered a suggestion correct as long as it is in accordance with at least one interpretation common in use.

This implies that a suggested word form, being appropriate according to one interpretation, is considered syntactically correct even though it turns out that the actually intended interpretation renders it erroneous. For instance, the verb *har* (have) may both function as a transitive head verb and a temporal auxiliary. If it functions as an auxiliary it imposes a supine head verb to immediately follow (in some contexts), thus making all other suggestions erroneous. If, on the other hand, the verb functions as a transitive head verb, a noun phrase suggestion is expected. According to our strategy both supine verbs and noun phrase initiators will thus be considered correct.

In a more compact way, the strategies are:

- A suggestion is judged exclusively in relation to its left context
- If there are competing interpretations, a suggestion is considered syntactically correct if it is appropriate according to at least one of them - that is common in use.

The notion of *common in use* is intuitively defined and somewhat arbitrary. We have however found it plausible to ignore some improbable interpretations. Even though Swedish word order is fairly regulated there is a wide range of counterexamples constituted by unusual, though acceptable structures. For instance, there is a major rule stating that once the potential verb complements are given, no nominal constituents are to be expected which makes the following suggestion wrong: *Maria donade *världen* (Maria pottered about *world). There are however possible continuations which would make the same suggestion correct, such as *Maria donade världen över* (Maria pottered about all over the world). The choice to exclude such interpretations is motivated by the assumption that the resulting suggestions, most of the time, would seem wrong to the user.

An alternative approach would be to make use of the right context when judging a suggestion, either to resolve ambiguities or to exclude all constituent types but the actually intended. Such a strategy is however more dependent on the representativity of the input sentences, and is probably better suited when the n-gram model is trimmed in relation to a specific user or when there are automatic means to evaluate the system responses in relation to a considerably larger set of input sentences.

6.2 Classification strategies

There are previously defined error typologies for Swedish, such as (Rambell 1998). These are however defining errors made by human writers in full sentence context, and are thereby not directly applicable to the errors produced by an n-gram model. For instance, many of the error types in (Rambell 1998) are

defined in relation to more than one constituent, such as the word order error labeled inversion: *På det här utrymmet vi kommer ge plats åt...* (In this area we will give room for...) [1]. In the context of a word predictor such definitions do not make sense, since we are not to judge full sentences but one word at a time and only in relation to the left context. If the above sentence were to be produced, *vi*(we are better considered erroneous with respect to the expectations emanating from the left context than in relation to the sofar not accessible *kommer*. We have thus found it necessary to form our own set of error classes.

The error classes have been defined incrementally as instances of previously unseen error types have been encountered. This entails that the error classification is not to be regarded as a complete typology but rather as means to get an overview of the actual error sample. The overall structure of the error classification and the principles by which the classes have been defined, were however settled in advance and will be described in the following.

The error classes are hierarchically organized. At top level, eight major classes are differentiated based on the context needed for the error descriptions. There are three classes comprising errors at phrasal level (noun phrase errors, prepositional phrase errors and adverb phrase errors), three classes containing word order violations at sentential level (main clause word order errors, subordinate clause word order errors and infinitive phrase word order errors), one class comprising verb inflection errors and verb valency errors (verb errors) and finally, one minor class accounting for a heterogeneous set of errors that do not neatly fit into any of the other classes (miscellaneous). At lower levels, the classes have in principle been defined as either of the following:

Inflectional errors are word forms suggested in contexts where their constituent type (part of speech) is expected but the inflection of the word form is inappropriate. A typical class defined this way captures violations of noun phrase agreement.

Missing constituents are suggestions belonging to syntactic categories others than the category to be expected. For natural reasons this kind of definition is only applicable when the expected constituent is compulsory and the word order is regulated. For instance, there are classes defined for missing finite verbs.

Constituents out of position are suggestions given at inappropriate positions. This way of defining has been useful in contexts in which no particular constituent must compulsory follow, as for instance, at the position succeeding potential verb complements.

The three ways of defining sometimes lead to intersecting classes. If, for instance, a pronoun is suggested on the position that is supposed to hold the finite verb, it can either be classified as a missing verb or a nominal constituent out of position. In order to make the classes non-overlapping we have defined a precedence order; classes that are defined to hold inflectional errors take precedence over all other classes, whereas classes expressing that a constituent is missing, take precedence over the out of position classes. Thus, if an erroneous suggestion may be classified according to more than one principle we choose to exclusively account for it in the class of highest precedence. The precedence order is defined so as to mirror the level of specificity by which the corresponding classes are defined.

Besides from classificational ambiguity due to intersecting classes, there may be syntactic or lexical ambiguity which makes the classification procedure less clear-cut. Sometimes a suggestion, in a syntactically or lexically ambiguous context, turns out to be erroneous according to all interpretations. For instance, the sentence initial token *Niklas* may be interpreted as a proper noun either in its base form or in its genitive form. Regardless of the interpretation chosen, the suggestion *hur* (how) is erroneous, either since it holds the position of the finite verb (succeeding the base form) or since the expected head noun is missing. Under these circumstances we have chosen to consider the suggestion as an instance of all affected classes. This was the case in 79 instances, whereby the 994 tokens in the error sample correspond to 1,073 error instances.

[1] In Swedish the finite verb is to precede the subject when the clause is initiated by an adverbial.

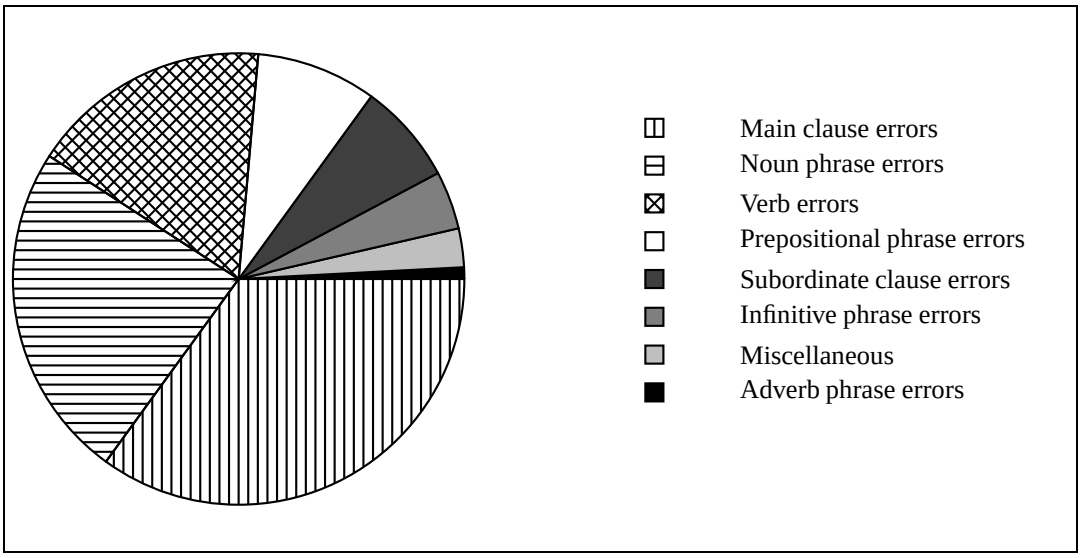

Figure 6.1: Distribution of error frequencies

6.3 Results

At top level the erroneous suggestions are distributed as shown in figure 6.1 below. The largest error class is main clause word order errors, which accounts for more than one third of the errors. The second largest error class, noun phrase errors, corresponds to approximately one fourth of the erroneous suggestions and errors having to do with the verb group and its complements (verb errors) amount to approximately one sixth of the total error sample. Thus, the three largest error classes account for three fourths of the errors. The remaining fourth is mainly constituted by prepositional phrase errors and subordinate clause word order errors.

Below we will give a more detailed description of the subclasses that comprise the eight top level classes. The classes will be presented in the order of frequency, both on top and subclass level.

6.3.1 Main clause word order errors

Swedish has a rather fixed word order even at clausal level. Before we proceed we would like to give a short summary of Swedish word order in terms of a topographical field scheme, a word order description first proposed by the Danish linguist Paul Diderichsen.

The field scheme is comprised of two major fields, the nexus field, which holds the subject, tensed verb and optional clause adverbial and the content field holding all other constituents. The scheme is somewhat different for main and subordinate clauses, as illustrated in tables 6.1 and 6.2 respectively.

Fundament	**Nexus field**			**Content field**		
	V1	N1	A1	V2	N2	A2
	Tensed Verb	Subject	Clause adverbial	Verb particles and infinite verbs	Verb complements and predicative attributes	Adverbials
-	*ska* (will)	*du* (you)	*inte* (not)	*köpa* (buy)	*boken* (the book)	*i morgon* (tomorrow)
du	*ska*	-	*inte*	*köpa*	*boken*	*i morgon*
i morgon	*ska*	*du*	*inte*	*köpa*	*boken*	-

Table 6.1: Main clause word order scheme

Subordinating field	**Nexus field**			**Content field**		
	N1	A1	V1	V2	N2	A2
	Subject	Clause Adverbial	Tensed verb	Verb particles and infinite verbs	Verb complements and predicative attributes	Adverbials
att (that)	*du* (you)	*inte* (not)	*ska* (will)	*köpa* (buy)	*boken* (the book)	*i morgon* (tomorrow)

Table 6.2: Subordinate clause word order scheme

The main clause scheme provides an extra field, referred to as the fundament field, which precedes the actual nexus field. This field can hold almost any constituent, which then leaves its ordinary position. Since Swedish is a verb second language, the fundament actually has to be filled in all main clause types except yes-no questions. (Jörgensen and Svensson 1995)

The rather strict word order is especially apparent in the nexus field since it holds the mandatory tensed verb and subject. As a consequence, the most obvious and indisputable violations of word order at sentence level are suggested by the system at clause beginning. This is reflected in the way by which the error classes are defined; whereas most problems in the nexus field are defined in terms of missing constituents, the errors in the rest of the clause are mainly quite vaguely defined as various constituents out of position. Some errors in the content field are however defined as missing verbal complements under the top level class verb errors.

In the following section we will describe the three subclasses comprising the top level class main clause word order errors: constituents out of position, missing verb and missing subject.

Constituents out of position

At positions where no particular constituent must compulsory follow we have classified erroneous suggestions as constituents out of position. This error class is hence applicable to suggestions succeeding syntactically complete clauses, i.e. in contexts where there may only come modifying adverbials or conjunctions. The erroneous word forms are mainly initiators of nominal constituents or verbs, both of which will be examplified below.

Nominal constituent out of position

The most frequent nominal constituent classified as out of position is the noun. Mostly these are suggested after other nouns, as in the following example:

- *Maria satte fram en rykande varm kastrull spagetti : personer*
  (Maria put out a smokingly hot saucepan spaghetti : people)

These suggestions are probably not supported by a large amount of higher order n-grams, which entails that the n-gram model may handle this context better if more weight were put on context sensitive n-grams. There are however other inappropriate suggestions that are in accordance with the bigram context, such as neuter nouns succeeding an adverb coinciding with a neuter adjective, e.g.:

- *ta det lugnt : underlag*
  (take it easy : foundation)

Some possible noun phrase initiators have, in many contexts, not been considered erroneous since they may constitute the beginning of an adverbial noun phrase, such as *en lång tid* (a long time) or *några månader* (some months).

Verbs out of position

Various verb forms are quite often suggested at the position for optional adverbial modifiers. We have distinguished finite and infinite verbs out of position, where the former are the most frequent. Finite verbs out of position are mostly suggested after pronouns and nouns, such as:

- *Maria donade vid diskbänken : blir*
  (Maria pottered about by the sink : becomes)

This may be due to that the subject usually is preceding the finite verb in Swedish. Consequently, we can expect the n-gram model to be based on a large amount of word form bigrams with the first position held by a pronoun or a noun and the second by a tensed verb; bigrams that will apply irrespective of the global context.

Besides from being suggested after nouns or pronouns, finite verbs out of position are mainly suggested after other verbs, e.g.:

- *Ljuset tändes : tändes*
  (The candle was-lit: was-lit)

The considerably smaller subclass holding infinitive verbs out of position have approximately the same distribution of preceding items as the subclass holding finite verbs out of position. It is either suggested after another verb or after a noun or a pronoun:

- *Det hade regnat: regnat*
  (It had rained : rained)
- *Maria satte fram en rykande varm kastrull: stanna*
  (Maria put out a smokingly hot sauce-pan : stay)

Missing verb in main clause

An instance of a missing verb may be of two types; missing verb in the nexus field or missing verb in the content field.

Missing verb in the Nexus field

Suggestions not being verbs, have been considered erroneous at the position succeeding the fundament field, such as:

- *Aron: att*
  (Aron : to)

Caution has to be taken though. First, if the fundament position holds a nominal constituent we must allow for possible post-attributes, e.g. a prepositional phrase, a relative subordinate clause or an apposition. This entails that even though a noun phrase may seem complete and the verb ought to be expected, there may still come other words modifying the noun. The possible noun phrase continuations are quite diverse. For instance, the relative pronoun in a subordinate clause may be left out whereby the noun phrase head noun is directly succeeded by another nominal constituent, i.e. the subject of the subordinate clause. Such suggestions are often semantically inappropriate, but since we are not to consider semantics we have not judged instances as the following wrong:

- *Maria : diskbänken*
  (Maria : sink+DEF)

Moreover, if the fundament holds an adverbial, this may in turn be further modified by another adverbial, such as *I går på morgonen* (Yesterday morning) and if the adverbial is comprised by a prepositional phrase the governed noun phrase may be modified, just as the nominal constituents above. The latter naturally also holds for adverbial noun phrases.

Having taken these cautions we have encountered 73 instances of missing verbs, amounting to approximately one fifth of the main clause word order errors.

Missing verb in the Content field

The second subclass accounts for non-verb suggestions in the second verb field in clauses where an infinite head verb is expected:

- *Nu kunde hon : sig*
  (Now could she : herself)

Actually, the above example constitutes the only indisputable instance of a missing verb in the content field. This may of course be interpreted as the n-gram model being particularly well suited for handling such constructions; often the auxiliary is directly succeeded by its head verb whereby the dependencies are local enough to be handled by the limited context provided by the n-grams. The low frequency is however further dependent on three things. First, some of the most frequently used auxiliaries may also function as transitive head verbs, e.g. *hava* (have) and *få* (get), whereby suggestions initiating appropriate verbal complements are considered as possibly correct continuations, i.e. they are not marked as syntactic errors. Second, if there is a nominal constituent in the fundament field not being marked for case, there is no way of knowing whether that constituent is the subject of the clause or a possible verbal complement of an upcoming head verb. If the latter is true, the subject is to hold the position directly after the auxiliary, making initiators of a nominal constituent syntactically correct. Finally, there may be clause adverbials between the auxiliary and the infinite verb, meaning that possible clause adverbials must be considered correct.

Missing subject in the main clause

As previously mentioned, when the subject is not holding the fundament position it is to follow the finite verb, which renders suggestions like the following wrong:

- *Utanför fortsatte : att*
  (Outside kept-on : to/that)

At this position, appropriate nominal constituents are mostly noun phrases. Nominal subordinate clauses and infinitive phrases are only rarely acceptable subjects at the position succeeding the finite verb. (Teleman, Hellberg and Andersson 1995) In the example above for instance, *att* is erroneously suggested.

6.3.2 Noun phrase errors

The top level class noun phrase errors subdivides into three lower level classes: naked nouns/adjectives, agreement errors and missing head.

Naked nouns/adjectives

An indefinite noun phrase in the singular must, as a rule, be quantified by some kind of determiner, unless the head noun is uncountable or used with a non-referential meaning (see further section 7.2.1). Therefore it will be considered syntactically wrong whenever the system suggests an indefinite, singular noun that neither is uncountable nor can be interpreted as non-referential, in a context not preceded by a determiner, e.g.:

- *Det regnade mot : fönsterruta*
  (It rained against : windowpane)

In general, this error type is only applicable to nouns even though an erroneously naked noun phrase may be initiated by an indefinite singular adjective. This is due to that while the attribute is suggested it is not known whether an uncountable head noun will follow. There is however one circumstance under which we have considered an attribute being naked, namely when it is in the definite form and carries masculine inflection, such as:

- *Hampus såg : nye*
  (Hampus saw : new+MASC+DEF)

Albeit not being naked in the strict sense, since that would require them to be in the indefinite form, these instances are naked in the sense of lacking a determiner. A definite noun may not be modified by an attribute without also having a determiner (see further section 7.2.4). In general though the inflection of definite attributes coincides with that of indefinite plurals, whereby they may not unambiguously be considered erroneous. Since this ambiguity does not adhere to definite attributes in the masculine, these may be considered erroneous in the above context.

The class naked nouns/adjectives accounts for 106 of the encountered errors, out of which only 4 are further subclassified as naked adjectives. Together the naked nouns and adjectives share approximately 40% of the total noun phrase error class.

Agreement errors

In Swedish, a noun and its modifiers are to agree in number, definiteness and gender (see further section 7.2.2). When a noun phrase constituent does not agree with the preceding items it is therefore considered erroneous. We have however chosen not to label indefinite nouns in definite noun phrases erroneous, since these are appropriate if the right context reveals a restrictive subordinate clause, as in *den natt han försvann* (the night+INDEF he disappeared), see further section 7.2.2.

The agreement error class is for natural reasons only applicable on constituents that are not at phrase initial position, leaving all kind of determiners out of the error description.

The erroneous word form may be an attributive modifier, such as:

- *Ett : ljusa*
  (A+SING/+INDEF: light+PLUR/+DEF)

or a noun, such as:

- *de andra : året*
  (the+PLUR other : year+SING)

Most of these, about three fourths, are violating the agreement restrictions even though the preceding word (usually a determiner) unambiguously exposes all features. For these instances, the n-gram model might work better if the weights were set differently i.e. giving a higher weight to word form bigrams. For the remaining fourth, the previous word is still a too limited context. Definite adjectives and pronouns are not marked for gender nor number in attributive context and further, they coincide with the indefinite plural. The inflection of a noun succeeding an ambiguous attribute, is therefore dependent on the determiner that is out of bigram context. Even if the n-gram model were extended to word form trigrams or a fuller part-of-speech tag description it is easy to imagine constructions for which the n-gram model would still provide a too limited context; a noun modified by more than one attribute or an attribute itself modified by an adverb. This implies that the error type exhibits a genuine shortcoming of mere n-gram techniques.

The agreement class amounts to approximately 30% of the total noun phrase error class, out of which a majority concerns non-agreeing nouns.

Missing head in the noun phrase

A determiner initiates a noun phrase, a property which further may hold for attributes, such as plural adjectives. This means that once a determiner (or an attribute) has been entered, suggestions that cannot be interpreted as continuations of a noun phrase will be considered syntactically wrong, e.g.:

- *Hampus och Arons : att*
  (Hampus' and Aron's: to/that)

Suggestions directly succeeding potential articles coinciding with pronouns are not assigned errors, since these may by themselves constitute noun phrases. We have generally chosen not to allow for adjectives used as the head of the noun phrase, but in accordance with (Teleman et al. 1995) we have accepted adjectives denoting human properties in this use.

6.3.3 Verb errors

The top level class verb errors is subdivided into two approximately equally sized lower level classes: verb inflection errors and missing verb complements, and one smaller class: Non-agreeing predicatives.

Inflection errors

The class verb inflection errors covers erroneously inflected verb suggestions. We differentiate between two kinds of verb inflection errors: verb inflection errors *in the nexus field* and verb inflection errors *in the content field*.

Verb inflection errors in the nexus field
Approximately half of the verb inflection errors adheres to the nexus field and are constituted by infinite verb suggestions at the position of the finite verb, e.g.:

- *Ljuset : dröja*
  (The light : stay+INFINITIVE)

Usually these errors concern the main clause, but some instances are suggested in a subordinate clause, e.g.:

- *de som : stanna*
  (they that : stay+INFINITIVE)

Verb inflection errors in the content field
Erroneously inflected verb forms have further been suggested in the content field of the main clause and the content field of the infinitive phrase. In the main clause it has been the expectations emanating from an auxiliary verb that have been violated, as in:

- *Det hade : regna*
  (It had: rain+INFINITIVE)

In the infinitive phrase, the nexus field (see section 6.3.6) may hold nothing but a clause adverbial, whereas the second verb field compulsory must hold an infinitive verb, a restriction that quite often has been violated, e.g.

- *Utanför fortsatte det att : regnade*
  (Outside it kept on to : rained)

The infinitive verb may further be an auxiliary verb, thus imposing an appropriate infinite head verb to follow. At a few occasions the system has given prediction suggestions in conflict with this constraint, e.g.:

- *ändå hispade hon runt (...) utan att sluta : arbetade*
  (yet she fiddled about (...) without to stop : worked)

Missing verb complements

The subclass missing verb complements accounts for suggestions not initiating an expected verb complement. These may be further subdivided according to what kind of verb complement that is expected. The largest such subclass contains suggestions violating the valency frame of monotransitive verbs, such as:

- *Han hade börjat skaffa sig : att*
  (He had started to get himself : to/that)

The other subclasses are rather small and covers instances such as missing predicatives in copula constructions and missing verb particles:

- *Hans huvud var : fortsatte*
  (His head was : continued)
- *Hampus såg nyvaken ut när han steg : inte*
  (Hampus looked newly awake when he stood : not)

Non-agreeing predicatives

In copula constructions an adjectival predicative is to agree with its correlating noun phrase. At a few occasions a non-agreeing predicative modifier has been suggested, such as:

- *Hans huvud var : färdig*
  (His head+NEUTR was : finished+UTR)

A predicative modifier is sometimes constituted by a noun phrase, which does not have to agree with its correlating noun phrase. There are however, often semantic restrictions imposing the modifying noun phrase to share number with its correlate. Since this is a semantically dependent restriction we have chosen not to account for such violations.

6.3.4 Prepositional phrase errors

Under the top level class prepositional phrase errors, all errors have been described in terms of missing government.

Missing government

Prepositions modify nominal constituents, such as noun phrases and nominal subordinate clauses. As a rule, the nominal item is directly succeeding the preposition, whereby any prediction suggestion other than initiators of appropriate nominal constituents will be classified as wrong after en entered preposition, e.g;

- *allt som hördes från : vandrade*
  (all that was heard from : wandered)

Some prepositions tend to take nominal constituents of certain subtypes, sometimes due to inherit features of the preposition. For instance, *bland* (among) restricts its complements to noun phrases headed by a plural or an uncountable noun. At other occasions restrictions emanates from a verb that heads the prepositional phrase as a whole. Regardless of whether a suggestion is not initiating a nominal constituent at all or initiating a nominal constituent of the wrong subtype, we have accounted for it under the same class.

The vast majority of erroneous suggestions succeeding a preposition consists of other prepositions or verbs. Only at a few occasions the error involves a nominal constituent of the wrong subtype and then it is always *att* suggested in a questionable context:

- *Varje dag lovade hon sig själv att sitta ner med : att*
  (Every day she promised herself to sit down with : to)

Prepositional phrases are sometimes split whereby the nominal constituent holds the fundament field whereas the preposition is in the content field, as exemplified in the following sentence: *henne tror jag verkligen på* (her I really believe in). (Jörgensen and Svensson 1995) Under such circumstances the government is of course not expected at the position succeeding the preposition and we may expect the n-gram model to do the opposite of missing a government - to produce nominal constituents out of position. The input sentences did however not include any split prepositional phrases.

6.3.5 Subordinate clause word order errors

The top level class subordinate clause word order errors subdivides into the subclasses missing subject and missing verb. Since a subordinate clause is a component part of the main clause, superfluous suggestions given after its completion are accounted for as out of position errors in the main clause. There is hence no need for such an error class under the top level class for subordinate clauses.

Missing verb in the subordinate clause

All instances of this error class adhere to missing verbs in the nexus field. This is due to that the limited set of input sentences was free from subordinate clauses with auxiliary verbs imposing a verb in the second verb field.

In a subordinate clause the subject is either to be succeeded by a clause adverbial or a finite verb. Suggestions that can not be interpreted as any of these are therefore considered erroneous and have been classified as instances of this class, e.g.:

- *innan det : sig*
  (before it : oneself)

Taking similar cautions as for missing verbs in the main clause, we have classified 55 instances as erroneous, amounting to more than 70% of the word order errors in the subordinate clause.

Missing subject in the subordinate clause

In a subordinate clause, the subject must compulsory hold the first position in the nexus field unless the subordinate clause is initiated by a relative pronoun that may then function as the subject. We have therefore considered non-initiators of nominal constituents as erroneous when these directly succeed a subjunction, e.g.:

- *utan att : slängde*
  (without: threw)

- *Sedan ställde hon sig i dörren och ropade att : minska*
  (Then she placed herself in the door and called out that: decrease+INFINITIVE)

The subjunction *att* is coinciding with the infinitive marker and if it functions as the latter, it is to be immediately followed by an infinitive verb. To a great extent, infinitive phrases and nominal subordinate clauses appear in the same syntactic contexts, making both infinitive verb forms and nominal constituents possibly correct after *att*. (Jörgensen and Svensson 1995) Sometimes though, the governing verb calls for a more restricted set of complements, as is the case in the second example above, in which the infinitive verb *minska* is inappropriate.

6.3.6 Infinitive phrase word order errors

All errors covered by the top level class infinitive phrase word order errors may be accounted for under the same lower level class missing verb. In principle, the word order in the infinitive phrase can be described by the same topographical field scheme as the subordinate clause if the fields holding the subject and the finite verb are left out. (Jörgensen and Svensson 1995) As follows, there are no classes accounting for missing verbs in the first verb field nor missing subjects.

Missing verb in the infinitive phrase

Even though the infinitive phrase has no finite verb there may be complex verb groups in the second verb field, since the infinitive verb may be an auxiliary. We have therefore distinguished missing first verb and missing second verb.

Missing first verb
Since the infinitive phrase lacks the fields holding the subject and the finite verb, the infinitive is to directly succeed either the infinitive marker or an optional clause adverbial. Hence, we have labeled suggestions that neither are possible clause adverbials nor verbs, at the position succeeding the infinitive marker, as missing verbs, e.g.:

- *Utanför fortsatte det att : det*
  (Outside it kept on to : it)

Missing second verb
If the infinitive verb is an auxiliary, the head verb of the phrase is to immediately follow. We have encountered a small set of suggestions violating this restriction, all succeeding the same left context e.g.:

- *utan att sluta : arbete*
  (without to stop : job)

6.3.7 Miscellaneous

The heterogeneous top level class miscellaneous covers three small subclasses; *adverbial noun phrase errors*, *pronoun case errors* and *conjunction errors*.

Adverbial noun phrase errors

A noun phrase, being a nominal constituent, is usually not to succeed a syntactically complete clause. Under some circumstances though, a noun phrase may function as an adverbial. This is often the case when the heading noun is denoting some kind of measure or a time reference point. (Teleman et al. 1995) When other nouns have been suggested we have considered them erroneous, e.g.:

- *Det hade regnat hela: landet*
  (It had rained the whole : country+DEF)

Pronoun case errors

In Swedish, personal pronouns are marked for subject and object case and sometimes the system has suggested pronouns in the wrong case, e.g.:

- *Varje dag lovade : honom*
  (Every day promised : him)

Conjunction errors

Conjunctions are used both to coordinate main clauses and components of clauses and may thus be used relatively freely. The coordinated items must however share the same syntactic function. (Teleman et al. 1995) In general it is not possible to judge a suggestion succeeding a conjunction wrong, since there may be several potential first coordinators compatible with the initiated second one. At a few occasions, all in the beginning of a sentence, the n-gram model has produced suggestions that unequivocally initiate a second coordinator not compatible with the first one, e.g.:

- *Hampus och : med*
  (Hampus and : with)

6.3.8 Adverb phrase errors

The few encountered errors concerning the adverb phrase may all be viewed as *missing head*. Some adverbs may not by themselves form an adverbial, but rather grade other adverbs (or adjectives). Hence, suggestions interfering with the completion of the adverb (or adjective) phrase, are seen as syntactically inappropriate, e.g.:

- *Hampus och Arons steg lät väldigt : att*
  (Hampus' and Aron's steps sounded very different: to/that)

Chapter 7

Defining the grammar

The constructions that had the highest error frequencies, according to the above described error classification, have been prioritized in the development of grammar rules. Thus, we have mainly striven to implement rules for handling word order in the main clause, agreement in the noun phrase, verb inflection and prepositional phrases. Some other constructions have also been covered by the rules, since they are essential constituents within larger segments.

An error may be handled by rules promoting expected constituents or by rules suppressing inappropriate suggestions, or by a combination of both. The choice of strategy has been dependent on the context in question, i.e. how rigid the word order is and whether the next constituent is compulsory. In general, the approach has been to handle errors labeled missing constituent by promoting the expected constituent and to cope with inflectional errors and constituents out of position by means of suppressing rules.

In the following sections, the constituents covered by the grammar, either by promoting or suppressing rules, will be described.

7.1 Rules for word order in the main clause

The most frequent errors occurring in the test material concerns word order in the main clause. Below we will describe the rules implemented to handle constituents out of position, missing verb and missing subject of the clause.

7.1.1 Rules for constituents out of position

According to the error classification, the most frequent word order error in the main clause adheres to the class constituents out of position. This class subdivides into two main categories; nominal constituents out of position and verbs out of position.

Nominal constituents

To be able to handle nominal constituents out of position, knowledge of the valency frame of the verb must be accessible. Without this information it is not possible to judge whether a nominal constituent is a legitimate part of the clause or not. In the lexicon used, the verbs are not annotated with valency information, leaving nominal constituents out of position out of scope of the grammar.

Verbs

At top level, a main clause may generally contain one head verb only. Thereby, a second head verb may be judged out of position as long as we can be certain to be in the same clause and that no subordinate clause has been initiated. The crucial question is thus to determine whether a second head verb is part

of a new clause or not. To be able to find definite clause borders, a thorough analysis imposing valency information is required. The absence of valency information in the lexicon thus leaves us with a cruder border detection.

To somewhat restrict the problem, the current grammar only copes with main clauses at sentence beginning, whereby the starting point of the clause, i.e. its left border, is known. However, the problem with finding the right border still remains. Often a verb complement or an adverbial is expressed by a subordinate clause, which then is part of the overall main clause. When the subordinate clause is initiated by a subordinate clause initiator (a subjunction, an adverb or a relative pronoun) the clause border is apparent. Sometimes though, the subordinate clause initiator is left out, which may be the case in attributive subordinate clauses and in subordinate *att*-clauses.

As a consequence, second head verbs are suppressed by the grammar in quite limited contexts, more specifically at positions immediately succeeding the head verb of the clause, possibly intermediated by a clause adverb and/or a noun phrase, as illustrated in figure 7.1.

1. NP ⇒ FINITE HEAD VERB ⇒ (CLAUSE ADVERB) ⇒ (NP) : *HEAD VERB
   *Flickan gillade (inte) (boken) *springer*
   (The girl did (not) like (the book) *runs)

2. NP ⇒ FINITE AUX ⇒ (CLAUSE ADVERB) ⇒ (INFINITE AUXS) ⇒INFINITE HEAD VERB ⇒ (NP): *HEAD VERB
   *Flickan skulle (inte) (kunna) gilla (boken) *springer*
   (The girl would (not) (be able to) like (the book) *runs)

3. ADVP ⇒ FINITE HEAD VERB ⇒ NP ⇒ CLAUSE ADVERB ⇒ NP : *HEAD VERB
   *Igår gillade flickan inte boken *springer*
   (Yesterday, the girl did not like the book *runs)

4. ADVP ⇒ FINITE HEAD VERB ⇒ (CLAUSE ADVERB) ⇒ (NP) ⇒ (CLAUSE ADVERB) ⇒ (NP+OBJ) : *HEAD VERB
   *Igår gillade (inte) (flickan) (honom) *springer*
   *Igår gillade (flickan) (inte) (honom) *springer*
   (Yesterday, (the girl) did (not) like (him) *runs)

5. ADVP ⇒ FINITE AUX ⇒ NP ⇒ (CLAUSE ADVERB) ⇒ (INFINITE AUXS) ⇒ INFINITE HEAD VERB ⇒ (NP) : *HEAD VERB
   *Igår skulle flickan (inte) (kunna) köpa (boken) *springer*
   (Yesterday, the girl would (not) (be able to) buy (the book) *runs)

Figure 7.1: Scheme of suppressed head verbs in the main clause

To make sure that we do not cross the border to an attributive subordinate clause, the rules are not applicable when there is more than one consecutive noun phrase between the two head verbs, unless the second one is in the object case. This is due to that the second noun phrase may otherwise constitute the potential subject of an attributive subordinate clause, as in the following example: *Flickan köpte boken läraren sa var bra* (The girl bought the book the teacher said were good).

Unfortunately, this means that misplaced head verbs succeeding two verb complements or a subject and a verb complement will pass unnoticed, as in *Flickan gav läraren boken springer* (The girl gave the

teacher the book runs) and *Igår köpte flickan boken springer* (Yesterday, the girl bought the book runs) respectively.

If however, the noun phrases are separated by a clause adverbial (as in context 3 in figure 7.1), a second verb is suppressed since an attributive subordinate clause may not be separated from its correlate.

For certain verbs a second head verb may be appropriate, and should not be suppressed, in the above listed contexts. This is the case for verbs taking as complements clauses or clause-like constructions where the subjunction is left out. With the guidance from various listings and descriptions in (Teleman et al. 1995) and (Thorell 1973), we have singled out verb types taking these kinds of complements and annotated approximately 500 verb forms with the corresponding features. Below these verb types, and the exceptions they impose on the scheme in 7.1, will be described.

Verbs taking *att*-clauses with the subjunction possibly left out

Verbs denoting the process of thinking, perceiving or uttering, often take as complement an *att*-clause with the subjunction left out (Thorell 1973), as in: *Flickan tror läraren kommer* (The girl thinks the teacher will come). If a potential subject of an upcoming subordinate clause is entered after such a verb, we may have crossed a clause border, and a second verb in the finite or supine[1] form is therefore not suppressed, e.g.:

- *Flickan tror (inte) pojken :* {*dansar/dansat*} [UNKNOWN]
  (The girl does (not) think the boy : dances)

If the intermediating noun phrase is marked for object case, we may however be sure not to have crossed an *att*-clause border. In the following sequence a second verb may therefore be suppressed:

- *Flickan tror (inte) honom : dansar* [REJECTED]
  (The girl does (not) think him : dances)

Verbs taking object+infinitive

Some verbs may take an object+infinitive as a complement, e.g.: *Jag hjälpte honom dansa* (I helped him to dance). If a potential object is entered after such a verb, a second verb in the infinitive is not suppressed:

- *Flickan tvingade (inte) honom : dansa* [UNKNOWN]
  (The girl did (not) force him : to-dance)

Verbs adhering to this class denote the process of helping, forcing[2], uttering[3] or perceiving.

Evidently, utterance and perception verbs, may either take an *att*-clause or an object+infinitive as a complement. If such a verb is succeeded by a noun phrase that may either be interpreted as the subject of an *att*-clause or as the object of an object+infinitive construction, no verb forms are suppressed:

- *Pojken hörde (inte) flickan :* {*komma/kom/kommit*} [UNKNOWN]
  (The boy did (not) hear the girl : {come+INF/came/come+SUP})

Verbs taking infinitive phrases with the infinitive marker left out

Yet another exception to the error rule concerns verbs that may take an infinitive phrase with a left out infinitive marker as a complement: *Jag råkade se mina vänner pussas.* (I happened to see my friends kiss+INFINITIVE) (Teleman et al. 1995). Therefore, infinitive verb forms are not suppressed immediately succeeding these:

[1]The supine form may appear as the first verb of a subordinate clause, since the temporal auxiliary *hava* (have) is often omitted.(Teleman et al. 1995)

[2]Some grammarians, as for instance (Jörgensen and Svensson 1995), do not label constructions following helping or forcing verbs object+infintive, since the infinitive marker may optionally be inserted.

[3]Most uttering verbs may only take certain object+infinitive constructions, such as *Hon sa sig vara trött* (She claimed herself to be tired.)

- *Jag råkade : se*[UNKNOWN]
  (I happened to : see+INFINITIVE)

Verbs taking a main clause as a complement

The same verbs that may take an *att*-clause with an omitted subjunction, may also function as quotation verbs taking a main clause as a complement. The main clause, functioning as a complement, often precedes the quotation verb, as in: *Han kommer, hörde jag* (He will come, I heard). As a rule, the complement and the quotation verb are separated by a comma, but as a precaution, quotation verbs are never suppressed as second verbs:

- *Han kommer : hör* [UNKNOWN]
  (He will come : hear+PRESENT)

Neither auxiliary verbs are suppressed, since they may initiate a verb group with a quotation verb as head, e.g. *han kommer har jag hört* (he will come I have heard).

7.1.2 Rules for missing verb

Verb in the nexus field

Since Swedish is a verb second language, the finite verb in the declarative main clause is to immediately follow the fundament and it would, of course, be desirable to have it promoted there. The crucial problem is to determine when the fundament constituent is complete. In SUC, the most frequent constituents at sentence initial position are noun phrases, prepositional phrases and adverb phrases, and below we will describe the promotion of finite verbs succeeding these.

Noun phrases

On word level, pronouns are the most frequent initiators of sentences in SUC. Pronouns rarely take any post-attributes and it is reasonable to assume the fundament border to immediately follow the pronoun. The promotion of finite verbs is hence straight forward, e.g.:

- *Han : spelar* [PROMOTED]
  (He : plays)

If the user however enters a post-attributing prepositional phrase, it is parsed as such and a finite verb is promoted once the prepositional phrase is completed:

- *Han med den röda jackan : spelar* [PROMOTED]
  (He with the red jacket : plays)

Due to nouns being more inclined to take post-attributes, the situation becomes more complicated when turning to noun phrases with nouns as heads. In SUC, only 60% of the head nouns are immediately succeeded by the finite verb [4]. The most common post-attribute is the prepositional phrase, and prepositions and finite verbs together amount to 83% of the words succeeding a head noun. We have therefore found it reasonable to promote not only verbs but also prepositions at this position:

- *Den gamla mannen : spelar* [PROMOTED]
  (The old man : plays)
- *Den gamla mannen : med* [PROMOTED]
  (The old man : with)

[4]This is an approximate figure, based on the distribution of simple noun phrase constructions which at most are modified by two premodifying attributes.

Proper nouns are also quite common at sentence initial position and are generally followed by verbs or other proper nouns. Since the succeeding word almost as often is a proper noun as it is a verb, it is not feasible to promote verbs without also promoting proper nouns. The promotion of proper nouns is however problematic. Consecutive proper nouns are mostly a combination of a first name and a surname, and given that the lexicon does not distinguish different types of proper nouns, we have chosen not to promote anything at all after proper nouns.

Prepositional phrases

In SUC, almost 60% of the sentence initial prepositional phrases are succeeded by finite verb forms. [5] Apart from verbs, the most common succeeding constituent is another preposition, and together the verbs and prepositions account for approximately 84% of the words following prepositional phrases. We have therefore chosen to promote prepositions as well as finite verbs at this position:

- *På morgnarna : spelar* [PROMOTED]
  (In the mornings : plays)
- *På morgnarna: under* [PROMOTED]
  (In the mornings : during)

Adverb phrases

Adverbs constitute a heterogeneous set of words displaying quite diverse behavior. In the context of a word prediction system, the subset that is easiest to handle comprises adverbs that may not modify other adverbs or adjectives, but may only function as adverbials on their own. Succeeding these, finite verbs are promoted:

- *Ofta : spelar* [PROMOTED]
  (Often : plays)

Some of these adverbs may initiate a subordinate clause, e.g. *När han ringde* (As he phoned), and after such adverbs both finite verbs and noun phrase initiators (constituting the beginning of the potential subject) are promoted:

- *När: spelar* [PROMOTED]
  (As : plays)
- *När: han* [PROMOTED]
  (As: he)

A second set of adverbs may not by themselves constitute an adverb phrase, but always modify other adverbs or adjectives. e.g. *ganska* (rather). At the position succeeding such an adverb a verb is consequently not to be expected but rather a second adverb or an adjective. If the adverb is succeeded by a second adverb, it is checked whether the latter adverb in turn, may modify adjectives or other adverbs, and if that is not so finite verbs are promoted. This process is iterative and proceeds as long as adverbs are entered:

- *Ganska : spelar* [REJECTED]
  (Rather : plays)
- *Ganska ofta : spelar* [PROMOTED]
  (Rather often : plays)

[5] This is an approximate figure, based on the distribution of prepositional phrases with very simple governments.

A third set of adverbs may either function independently or modify adjectives or other adverbs. In order not to overly promote verb suggestions, finite verbs are not promoted directly succeeding these.

Since the initial lexicon did not provide information on whether an adverb may modify adjectives or other adverbs, nor if they may initiate subordinate clauses, we have manually examined and annotated them. The annotation was guided by how the adverbs are distributed in SUC.

Verb in the content field

If the finite verb is an auxiliary it is feasible to promote an appropriately inflected head verb in the content field. The problem is to determine exactly when to promote it, since there are several constituents that may come between the auxiliary and the head verb. If the subject is not holding the fundament position it is mandatory at the position between the auxiliary and the head verb and further, there may come an intermediary clause adverbial. Many auxiliaries are moreover homographs to transitive head verbs, whereby no second verb may be expected at all. Altogether, this has led us not to promote verbs in the content field, since other, possibly intended suggestions may be pushed down in the prediction list if the promotion is done too early. However, we do reject erroneously inflected verb forms, as described in section 7.3.2 below.

7.1.3 Rules for missing subject

Noun phrases are promoted after the finite verb in contexts where we may be certain that the subject has not yet been entered, i.e. when the fundament is holding a pronoun in the object case or an adverbial:

- *Honom pratade :* {*jag/flickan*} [PROMOTED]
  (Him talked : I/girl+DEF)
- *Igår pratade :* {*jag/flickan*} [PROMOTED]
  (Yesterday talked : I/girl+DEF)

7.2 Rules for noun phrases

Approximately one fourth of the errors found in the error classification process were noun phrase errors. Further, noun phrases form cardinal constituents in larger constructions, such as the subject in a clause. Hence, we have focused on handling various noun phrase constructions. In the following we will describe the rules implemented for handling naked noun phrases, agreement within the noun phrase and missing head. [6]

7.2.1 Naked noun phrases

A naked noun phrase is an indefinite noun phrase with a singular countable noun as head, occurring without any quantitative attribute. The use of naked noun phrases is more limited to certain contexts than the use of ordinary noun phrases. We have not yet fully investigated the distribution of naked noun phrases, whereby the current grammar never promotes constituents within the naked noun phrase.

- *ganska ful : padda* [UNKNOWN]
  (quite ugly : toad)

Once a complete naked noun phrase has been entered though, it is interpreted as a noun phrase. This means, for instance, that finite verbs are promoted after a sentence initial naked noun phrase as well as after a sentence initial non-naked noun phrase (see further 7.1.2).

[6] Throughout the chapter describing rules for noun phrases, (Teleman et al. 1995) is the default reference.

- *Ganska ful padda : hoppar* [PROMOTED]
  (Quite ugly toad : jumps)

The most common kind of naked noun phrase consists of a single noun, but postmodifying and premodifying attributes may also be included (except for a quantitative attribute in accordance with the definition of naked noun phrases). The current grammar handles the same attributes for naked noun phrases as for ordinary noun phrases (see below sections 7.2.4 and 7.2.5).

It would have been desirable to make a distinction between countable and uncountable nouns, since only uncountable nouns may form non-naked noun phrases without a quantitative attribute. Since this information is absent in the current lexicon, all indefinite noun phrases occurring without a quantitative attribute are accounted for as being naked.

7.2.2 Agreement

Within the noun phrase the attributes and the head word are to agree in gender, number and definiteness. Since a grammar used for prediction only has access to the left context of a clause, one must assume that the attributes already written contains the correct features. The current grammar thereby rejects prediction suggestions that do not agree with the preceding attributes of the noun phrase:

- *de gröna : paddan* [REJECTED]
  (the+PLUR green+PLUR : toad+SING)

An exception to the agreement restriction is the definiteness agreement in the definite noun phrase, where the head noun normally is to be in the definite form (if the phrase is not initiated by a definite attribute that requires an indefinite head noun, see section 7.2.4). However, if the noun phrase is followed by a restrictive relative subordinate clause the head noun of the noun phrase may be in the indefinite form. Since the right context of the current sentence is unknown to the grammar, there is no way of knowing whether a restrictive relative subordinate clause will follow or not. Therefore we have chosen neither to suppress nor promote the indefinite forms in this case.

- *de gröna : paddorna* [PROMOTED]
  (the green : toads+DEF)
- *de gröna : paddor* [UNKNOWN]
  (the green : toads+INDEF)

7.2.3 Rules for missing head

To avoid prediction suggestions interfering with the completion of the noun phrase, rules have been defined to promote constituents allowed within the noun phrase. These rules are activated once a noun phrase initiator has been typed and are active until another word has been entered that either may be interpreted as the head of the noun phrase or as a constituent not belonging to a noun phrase at all, e.g. a verb.

A noun phrase may have a noun, a proper noun or a pronoun as its head. In noun phrases headed by a proper noun or a pronoun, attributes are only rarely occurring and we have chosen not to parse any attributes in such noun phrases. [7] A noun phrase with a noun as head may take both premodifying and postmodifying attributes and in the remaining part of this chapter it is this kind of noun phrase we refer to by the term noun phrase.

For both definite and indefinite noun phrases, the possible premodifying attributes are the same (apart from the definite attribute exclusively used in definite noun phrases). (Teleman et al. 1995) points out

[7] Note, section 7.1.2, that prepositional post-attributes are parsed after pronouns at sentence initial position, albeit not promoted.

three main types of premodifying attributes; definite attributes, quantitative attributes and adjective attributes. In (Teleman et al. 1995):s description, the adjective attribute comprises both pronouns and adjectives, even though the pronouns have a slightly different function from the adjectives and are to precede the latter. For the sake of simplicity we have chosen to incorporate yet another attribute type, a pronoun attribute. The preferred word order within the noun phrase is then:
definite attribute ⇒ quantitative attribute ⇒ pronoun attribute ⇒ adjective attribute ⇒ head noun

Even though the word order is rather strict, there is an exception regarding the quantitative attribute in the definite noun phrase. It may leave its ordinary position and move into the adjective attribute instead, occurring after any adjectives in the superlative, but before adjectives in the positive:

- *de tre finaste gröna paddorna*
  (the three finest green toads)
- *de finaste tre gröna paddorna*
  (the finest three green toads)

The grammar suppresses quantitative attributes on the latter position though, if a quantitative attribute has already been entered:

- *de tre finaste : tre* [REJECTED]
  (the three finest : three)

Even though we consider the above described word order the most appropriate one, the user may enter a noun phrase with a less acceptable word order, for example a noun phrase where the adjective attribute is preceding the pronoun attribute, as in *tre gröna sådana paddor* (three green such toads). We do not want to promote this word order, since it is more or less inappropriate, but it would nevertheless be desirable to interpret the constituents as part of a noun phrase in order to give qualified guesses about the following constituents. Hence, the word order rules promote constituents following the defined word order, so that these suggestions are displayed at the top of the prediction list. Simultaneously, constituents that may be part of a noun phrase, but that are out of position, are neither promoted nor suppressed (provided that they follow the agreement restrictions). The only attributes that must strictly be at their defined positions are the definite attribute in the definite noun phrase and the quantitative attribute in the indefinite noun phrase, since these are the initiators of the phrase and ought always to be at first position.

- *en annan : sådan* [PROMOTED]
  (another : such)
- *en sådan : annan* [UNKNOWN]
  (a such : other)
- *en sådan annan : hund* [PROMOTED]
  (a such other : dog)

Occasionally the head of the noun phrase may be left out, whereby an attribute functions as head of the phrase. There are restrictions on what attributes may function as the head of a noun phrase. We have not yet accomplished a thorough investigation on this matter. Consequently, the current grammar always promotes a head noun to follow any premodifying attributes. Further, a noun phrase with the head left out will not be interpreted as a noun phrase by the grammar, which may sometimes lead to less appropriate prediction suggestions:

- *De gamla : vårdas* [UNKNOWN]
  (the old : are nursed)

- *De gamla männen : vårdas* [PROMOTED]
  (the old men : are nursed)

7.2.4 Premodifying attributes

In the succeeding sections we will give a detailed description of the various attributes in their order of appearance within the noun phrase. The associated rules simultaneously account for the promotion and suppression of words based on the attested errors labeled missing head in the noun phrase and agreement errors.

Definite attribute

The definite attribute initializes a definite noun phrase and may consist of a definite article, a demonstrative pronoun or a genitive construction.

Definite article
The definite article should only be used when at least one more premodifying attribute is entered before the head noun occurs, or else when the noun phrase is followed by a restrictive relative or narrative subordinate clause. However, since all the definite articles are homographs to demonstrative pronouns, which are allowed to immediately precede the head noun, this restriction has not been taken into account in the development of our grammar rules. Hence, head nouns immediately succeeding a definite article are not suppressed. Since the definite article coincides with a personal pronoun, we have chosen not to promote possible noun phrase continuations, unless we are at sentence initial position or another attribute has been entered. As described in section 7.1.2 finite verbs are promoted after a sentence initial pronoun and if we were not to promote possible noun phrase continuations these would be superseded by verb suggestions.

- *den : paddan* [UNKNOWN]
  (the : toad+DEF)
- *Den : paddan* [PROMOTED]
  (The : toad+DEF)
- *den gröna : paddan* [PROMOTED]
  (the green : toad+DEF)

Demonstrative pronouns
A demonstrative pronoun may either immediately precede the head noun of the noun phrase or be succeeded by other noun phrase attributes. We have chosen not to promote adjectives in the comparative or superlative in a noun phrase initiated by a demonstrative pronoun, since these only at exceptional circumstances appear together. We believe their reluctance to co-occur to be based on the demonstratives pointing out an object uniquely, whereby further selection provided by a superlative or comparative adjective appears redundant. Therefore only adjectives in the positive (rather being descriptive than selective) are promoted at this position. However, adjectives in the comparative or superlative are not suppressed, since it is rather a question of semantics than syntax, and a potential user may intend to write a noun phrase constructed this way.

Since *denna* may by itself constitute a noun phrase, the promotion of noun phrase continuations is restricted to the same contexts as for the definite article.

- *dessa : paddor* [UNKNOWN]
  (these : toads)

- *Dessa : paddor* [PROMOTED]
  (These : toads)

- *dessa fina : paddor* [PROMOTED]
  (these fine : toads)

The head noun ought to be in the definite form when the noun phrase is initialized by a demonstrative pronoun, except if the demonstrative pronoun in question is *denna* (this) or any of its inflected forms. Adjective attributes and head nouns with the wrong definiteness are suppressed, in accordance with the agreement restrictions.

- *dessa : paddorna* [REJECTED]
  (these : toads+[DEF)

Genitive constructions

A genitive construction may be a possessive pronoun or a noun phrase in the genitive. As for the demonstrative pronoun *denna*, a head noun following a genitive construction is to be in the indefinite form. Any premodifying attributes between the genitive construction and the head noun should however be in the definite form. Prediction suggestions not following these agreement restrictions are suppressed and appropriate suggestions are promoted.

- *den söta flickans : gröna* [PROMOTED]
  (the cute girl+GEN : green+DEF)

- *den söta flickans : grön* [REJECTED]
  (the cute girl+GEN : green+INDEF)

- *min : padda* [PROMOTED]
  (my : toad+INDEF)

- *min : paddan* [REJECTED]
  (my : toad+DEF)

Immediately succeeding the genitive construction, preceding any other attributes, the pronoun *egen* (own) or any of its inflected forms may occur. There is a norm saying that *egen* should be in its indefinite form if occurring at this position. This is however not a very strict rule, and in common use the definite form is also occurring immediately after the genitive construction. Therefore the current grammar promotes both the definite and the indefinite form of *egen* at this position. If on the other hand, *egen* occurs at the adjective attribute position, i.e. is preceded by other attributes, *egen* should always be in the definite form and accordingly, the grammar suppresses indefinite forms.

- *flickans : egen* [PROMOTED]
  (girl+DEF+GEN own+INDEF)

- *flickans : egna* [PROMOTED]
  (girl+DEF+GEN own+DEF)

- *flickans finaste : egna* [PROMOTED]
  (girl+DEF+GEN finest : own+DEF)

- *flickans finaste : egen* [REJECTED]
  (girl+DEF+GEN finest : own+INDEF)

Quantitative attribute

The quantitative attribute forms the initial item of the indefinite noun phrase and may also function as a premodifying attribute within the definite noun phrase. The quantitative attribute includes cardinal numbers, numerical phrases and quantitative pronouns. In the indefinite noun phrase the quantitative attribute may also be an indefinite article or a quantitative noun phrase.

Cardinal numbers

Only cardinal numbers that agree in number with preceding attributes are promoted. Non-agreeing cardinal numbers are not suppressed though, since the cardinal number may constitute the beginning of a descriptive noun phrase and may hence not be grammatically inappropriate, as in *den tre meter höga byggnaden* (the three meters high building).

- *De : tre* [PROMOTED]
  (The+PLUR : three)
- *Den : tre* [UNKNOWN]
  (The+SING : three)

Numerical phrases

We have defined a numerical phrase as consisting of any of the possibly independent numerical nouns *hundra* (hundred) or *tusen* (thousand), or a cardinal number followed by a numerical noun, as in *tre hundra* (three hundred) and *tre miljoner* (three millions). The adverbs *cirka* (roughly) and *ungefär* (approximately) may further modify the numerical phrase. Syntactically the numerical phrase functions the same way as ordinary cardinal numbers and once it is initiated possible continuations are promoted.

- *de tre : hundra* [PROMOTED]
  (the three : hundred)

Quantitative pronouns

Quantitative pronouns may either be at phrase initial position in an indefinite noun phrase or constitute an attribute within a definite noun phrase. Some quantitative pronouns are inherently indefinite or definite though, and are only applicable in one or the other of these contexts. In a definite noun phrase, only quantitative pronouns that are either not specified for definiteness or that are inherently definite are promoted whereas indefinite pronouns are suppressed, e.g.:

- *grodans : ingen* [REJECTED]
  (frog+GEN : no+INDEF)
- *grodans : båda* [PROMOTED]
  (frog +GEN : both+DEF)

Indefinite article

The indefinite article *en* (or its neuter counterpart *ett*) may initialize an indefinite noun phrase in the singular. The indefinite article in its uter form is homographic to an indefinite personal pronoun, whereby the promotion of noun phrase continuations succeeding these follows the same restrictions as for the definite article. Plural or definite continuations are however rejected in all contexts.

- *En : padda* [PROMOTED]
  (A : toad+INDEF)
- *En : gröna* [REJECTED]
  (A : green+PLUR/DEF)

Quantitative noun phrases

The quantitative noun phrase has a measure denoting or a quantity denoting noun as a head, e.g. *meter* (meter) and *kopp* (cup), and is used when the head noun of the overall noun phrase is uncountable or denotes collective plural. Quantitative noun phrases are generally used only in indefinite noun phrases.

To decide which nouns to include in the set of measure or quantity denoting nouns, we searched SUC for n-grams consisting of nouns in the nominative followed by a new noun or an adjective. The list of such n-grams was manually inspected and nouns frequently occurring at this position that further seemed reasonable in the function as quantitative nouns were annotated as such. The inspection led to a list of 113 word forms (corresponding to 96 lemmas).

Some of the listed nouns, such as *skiva* (slice), may only modify uncountable nouns. Since the distinction between countable and uncountable nouns is missing in the current lexicon, any noun in the singular is promoted as head of the overall noun phrase (at sentence beginning or after another attribute), whereas plural nouns are suppressed:

- *En skiva : ost* [PROMOTED]
  (A slice of: cheese)
- *En skiva: padda* [PROMOTED]
  (A slice of : toad)
- *En skiva: paddor* [REJECTED]
  (A slice of : toads)

Others, such as *flock* (flock), may only modify plural nouns:

- *En flock: paddor* [PROMOTED]
  (A flock of: toads)
- *En flock: padda* [REJECTED]
  (A flock of : toad)

The largest set of measure or quantity denoting nouns may modify both uncountable and plural nouns and in the context of these we promote all nouns, regardless of number, e.g.:

- *Ett paket : tuggummi* [PROMOTED]
  (A packet of: chewing-gum)
- *Ett paket: cigarett* [PROMOTED]
  (A packet of : cigarette)
- *Ett paket: cigaretter* [PROMOTED]
  (A packet of : cigarettes)

The corpus study confirmed the statement made in (Teleman et al. 1995), that quantitative noun phrases are only rarely occurring in definite noun phrases. We have therefore chosen to implement rules for handling quantitative noun phrases as attributes exclusively in indefinite noun phrases.

There is also a special kind of genitive construction that functions very much the same as quantitative noun phrases. The construction in question is built with *sorts* (kind of), *sorters* (kinds of) or *slags* (kind of) and is similar to the measure and quantity denoting noun phrases in that it modifies indefinite noun phrases, while ordinary genitive constructions constitute the definite attribute in definite noun phrases. Accordingly, any attributes (as well as the head noun) succeeding this special kind of genitive construction is to be in the indefinite form. Hence, the current grammar promotes indefinite attributes and nouns succeeding this kind of construction and also suppresses definite forms.

- *en sorts : hund* [PROMOTED]
  (a kind of : dog+INDEF)

- *en sorts : hunden* [REJECTED]
  (a kind of : dog+DEF)

Pronoun attribute

The pronoun attribute is not part of the definition of noun phrases made in (Teleman et al. 1995). We have chosen to treat it as a separate attribute, since the pronouns that (Teleman et al. 1995) includes within the adjective attribute actually normally precede any adjectives and further function in a slightly different way. In our grammar the pronoun attribute of the definite noun phrase includes ordinal numbers, ordinative pronouns, partitive pronouns and the word form *enda* (only), whereas the pronoun attribute of the indefinite noun phrase may consist of either a relational or a comparative pronoun.

Ordinal numbers

The ordinal numbers may hold either phrase initial position or the position succeeding the quantitative attribute. When it is the initial item of the noun phrase we find it unlikely that another premodifying element come in between the ordinal number and the head noun, such as ?*tredje pratiga sidan* (?third chatty page). Thus only nouns are promoted after an ordinal number occurring without a definite attribute. If the definite attribute is present, both nouns and adjective attributes are promoted.

Whatever position the ordinal number holds, succeeding plural constituents are suppressed, since ordinal numbers may only occur in singular noun phrases.

- *tredje : paddorna* [REJECTED]
  (third : toads)

Succeeding constituents are only promoted if they are both singular and definite. According to (Teleman et al. 1995) though, the head noun may be in the indefinite form when the ordinal number is phrase initial. We believe the indefinite form, only occurring in certain lexicalized expressions, like *tredje plats* (third place), to be rare enough not to be covered by the promoting rules.

- *tredje : paddan* [PROMOTED]
  (third : toad+DEF)

- *den tredje : paddan* [PROMOTED]
  (the third : toad+DEF)

Ordinative pronouns

Ordinative pronouns point out one or more referents within an ordered set of referents. The ordinative pronouns are *första* (first), *sista* (last), *nästa* (next) and *förra* (former). Pronouns functioning similar to ordinative pronouns, although traditionally viewed as other kinds of pronouns, will also be accounted for here. That is the reason why the comparative pronoun *samma* (same) and the perspective pronoun *ena* (one of) will be discussed below.

Förra and *ena* function the same way as ordinal numbers, as do the pronouns *första* and *sista*, except that *första* and *sista* may also have a plural meaning. Promoting and rejecting rules are implemented accordingly.

- *första : gången* [PROMOTED]
  (first : time+DEF)

- *den första : gången* [PROMOTED]
  (the+SING first : time+DEF+SING)

- *de första : gångerna* [PROMOTED]
  (the+PLUR first : times+DEF+PLUR)

Nästa and *samma* differ in that they may only be at noun phrase initial position and that the head noun is to be in the indefinite form. Succeeding *nästa* only attributes in the positive form are promoted, since we believe *nästa* to point out an object in the same way as an adjective in the superlative or comparative does (compare *Demonstrative pronouns* above). Indefinite adjectives and definite nouns are suppressed by the grammar as agreement violations.

- *samma/nästa: padda* [PROMOTED]
  (the same/the next : toad+INDEF)
- *samma/nästa : paddan* [REJECTED]
  (the same/the next : toad+DEF)

Partitive pronouns

We have introduced the notion partitive pronouns for pronouns that in (Teleman et al. 1995) are said to have a partitive meaning. The pronouns of this kind mentioned in (Teleman et al. 1995) are *endera* (one of), *ingendera* (neither of), *någondera* (one or the other of), *vardera* (each) and *bådadera* (both) along with their inflectional forms. We have also included *vilkendera* (which of) and the synonyms to *bådadera*: *bäggedera*, *båda* and *bägge*, since they function likewise.

We have defined the partitive pronouns to function as noun phrase initial items in the definite noun phrase with the head noun in its definite form. Nouns in the indefinite form are thus suppressed by the grammar. We have also chosen not to promote any attributes between the pronoun and its head noun.

- *bådadera : barnen* [PROMOTED]
  (both : children+DEF)
- *bådadera : snälla* [UNKNOWN]
  (both : kind+DEF)
- *bådadera : barn* [REJECTED]
  (both : children+INDEF)

enda

The pronoun *enda* (only) may occur either as a noun phrase initial item or at the position succeeding the quantitative attribute. In our grammar *enda* may only hold phrase initial position in the definite noun phrase. According to (Teleman et al. 1995) *enda* may also initiate an indefinite noun phrase, as in *enda orsak* (only reason). We find this a rare construction, that might prevent other more appropriate suggestions from being displayed in the prediction list if promoted. Accordingly only attributes and head nouns in the definite form are promoted succeeding a phrase initial *enda*, whereas both indefinite and definite forms are promoted elsewhere.

- *enda : möjliga* [PROMOTED]
  (only : possible+DEF)
- *enda : orsaken* [PROMOTED]
  (only : reason+DEF)
- *den enda : orsaken* [PROMOTED]
  (the only : reason+DEF)
- *en enda : orsak* [PROMOTED]
  (a single : reason+INDEF)

Relational pronouns

The relational pronouns *höger* (right), *vänster* (left), *rätt* (correct) and *fel* (wrong) are only to occur at noun phrase initial position without any succeeding adjective attributes. There are also relational pronouns that may or may not be at noun phrase initial position, such as *egen* (own). *Annan* (other) and *enda* (only) on the other hand ought to be preceded by a quantitative attribute, if occuring in an indefinite noun phrase.

- *vänster : dörr* [PROMOTED]
  (left : door)
- *ett : annat* [PROMOTED]
  (an : other)
- *eget : rum* [PROMOTED]
  (own : room)
- *ett : eget* [PROMOTED]
  (an : own)

Comparative pronouns

We have chosen to treat the pronouns *dylik* (suchlike), *slik* (such), *sådan* (such) and *likadan* (similar) as a separate category, hence comparative pronouns. These may only occur with a preceding quantitative attribute. Succeeding a comparative pronoun, adjective attributes and head nouns are promoted.

- *en sådan : padda* [PROMOTED]
  (such a : toad)

In case the pronoun attribute of an indefinite noun phrase includes both a relational pronoun and a comparative one, the relational pronoun is to precede the comparative pronoun. Accordingly the current grammar promotes comparative pronouns after a relational pronoun. The opposite case is neither promoted nor suppressed.

- *en annan : sådan* [PROMOTED]
  (another : such)
- *en sådan : annan* [UNKNOWN]
  (such an : other)

Adjective attribute

The adjective attribute includes adjective phrases with adjectives or pronouns as head and simple participle phrases.

Within the adjective attribute only a certain word order is promoted, where adjective phrases in the superlative or comparative are to precede adjective phrases in the positive in accordance with (Teleman et al. 1995).

- *den finaste : gröna* [PROMOTED]
  (the finest : green)
- *den gröna : finaste* [UNKNOWN]
  (the green : finest)

Concerning the participles we have chosen to let the so called pronounlike participles (see below) belong to the same category as the superlatives and comparatives, since we believe that these should also precede the adjectives in the positive, while ordinary participles have the same position as the adjectives in the positive.

Adjective phrases

An adjective phrase, as defined in our grammar, consists of a single adjective or an adjective modified by a preceding adverb. This entails that an adverb ought to be promoted at positions in which an adjective phrase is expected. Adverbs constitute a quite heterogeneous group of words, where not all are likely to occur within a noun phrase. To find out which adverbs are common in this use, we searched SUC for n-grams consisting of an adverb immediately followed by an adjective and a noun. From the list of adverbs (with a frequency above 10) we excluded most of those that are homographs with adjectives in the neuter gender, since the coinciding adjective interpretation may seem erroneous to the user in an uter noun phrase. Therefore words like *starkt* (strong/strongly) have been excluded from the group of attributive adverbs. Some of the homographs were not excluded though, since they are most commonly used with the adverbial interpretation. This is the case for words like *oerhört* (enormous/enormously). The decisions made on this matter are solely based on our linguistic intuition.

- *grodans : oerhört* [PROMOTED]
  (the : enormous(ly))
- *grodans : starkt* [UNKNOWN]
  (the : strong(ly))

The investigation led to a list of 24 adverbs, that were further divided on the basis of the grade of comparison they impose on the succeeding adjective, as described in 7.5.

When no other adjective attribute has yet been entered in the noun phrase, adverbs as well as adjectives and participles are promoted. When the first adjective attribute has been typed though, adverbs are no longer promoted, since we believe that it is not very common to use an adverbially modified adjective phrase after another adjective attribute. If the user however chooses to type yet another adverb, appropriate succeeding adjectives are still promoted.

Adjectives with the wrong grade of comparison are suppressed, as are uter adjectives in neuter noun phrases. The reverse, a neuter adjective in an uter noun phrase, is however not suppressed since it may coincide with an adverb. The lexicon only lists these homographs as adjectives, whereby this approach is necessary to make sure that no appropriate suggestions are suppressed.

- *den ganska : fula* [PROMOTED]
  (the quite : ugly)
- *den ganska : fulare* [REJECTED]
  (the quite : uglier)
- *ett : ful* [REJECTED]
  (an+NEUTR : ugly+UTR)
- *en : fult* [UNKNOWN]
  (an+UTR : ugly+NEUTR)

Regardless of whether the adjective phrase consists of a single adjective or an adjective modified by an adverb, there are different agreement restrictions depending on the grade of comparison of the adjective. Adjectives in the comparative are not inflected regarding definiteness and are, accordingly, appropriate in both definite and indefinite noun phrases. Positive adjectives are also applicable in both

kinds of phrases, but must be appropriately inflected. Superlative adjectives are only to occur in definite noun phrases, even though there are both definite and indefinite superlative forms, due to that the indefinite form is generally used in predicative position or in naked noun phrases. There are some exceptions where the superlative form is applicable also in an indefinite noun phrase, but these are rare and have not been covered by our grammar. Accordingly, any adjective in the superlative form is suppressed in indefinite noun phrases, while only those in the indefinite form are suppressed in definite noun phrases.

- *den : finaste* [PROMOTED]
  (the : finest+DEF)
- *den : finast* [REJECTED]
  (the : finest+INDEF)
- *en : finast* [REJECTED]
  (a : finest)

Participles

Both past and present participles are promoted at the adjective attribute position. We have made a distinction between ordinary participles and pronounlike participles, such as *föregående* (previous), where pronounlike participles are defined as participles functioning syntactically similar to relational pronouns. The main difference between ordinary and pronounlike participles is that the pronounlike participles have the ability to initiate noun phrases, as in *föregående två dagar* (previous two days, and further that the quantitative attribute may succeed the pronounlike participle, as in *de föregående två dagarna* (the previous two days).

To find out which participles should be considered pronounlike we did a corpus study, using SUC. From the corpus we extracted those participles that were sentence initial or immediately preceded by a verb or a preposition and immediately succeeded by a noun (thus constituting a noun phrase initial item). The list was manually inspected and the participles that functioned in a pronounlike way were gathered together with the examples of pronounlike participles mentioned in (Teleman et al. 1995). This gave us a list of 86 word forms (60 lemmas) to be used in the grammar.

The pronounlike participles may constitute a noun phrase initial item either in a definite or an indefinite noun phrase. Since present participles are not inflected regarding gender, number or definiteness, all forms of the adjective attribute are promoted after an initial pronounlike present participle. If, on the other hand, the noun phrase initial pronounlike participle is in the past tense, succeeding attributes with non-agreeing features are suppressed.

- *nuvarande : svenska* [PROMOTED]
  (present : Swedish+DEF)
- *nuvarande : svensk*[PROMOTED]
  (present : Swedish+INDEF)
- *ovannämnt : svensk* [REJECTED]
  (above-mentioned+NEUTER : Swedish+UTR)

7.2.5 Postmodifying attributes

Apart from premodifying attributes, postmodifying attributes may occur as part of the noun phrase. Possible postmodifying attributes are: relative subordinate clauses, adjectival phrases, participle phrases, prepositional phrases, adverb phrases, adverbial subordinate clauses, noun phrases, infinitive phrases, narrative/interrogative subordinate clauses and certain constructions modifying any of the premodifiers. (Teleman et al. 1995)

The current grammar generally does not promote postmodifying attributes after noun phrases. This is due to that postmodifying attributes are not compulsory. Thus the user may intend to write for example an adverbial after the head noun of the noun phrase and not insert any postmodifying attributes. If the grammar was to promote postmodifying attributes each time a noun phrase had been written, other possible next constituents would be pushed down in the prediction list and in some cases not be displayed to the user at all.

However, after head nouns at sentence initial position, finite verbs are promoted and possible post-attributes may in turn be pushed down. Since prepositional post-attributes turned out to be quite frequent at this position (see further 7.1.2) prepositions are, along with the finite verbs, promoted. Further, complete prepositional phrases following a noun phrase are interpreted as postmodifying attributes once they have been written, irrespective of where in the sentence the noun phrase appears.

7.2.6 Noun phrases consisting of single nouns

A noun may sometimes by itself constitute a noun phrase. Actually, a noun in the definite form without any attributes is the most common kind of noun phrase in the Swedish language. Definite nouns are thus promoted at positions where a noun phrase is expected, e.g. as the government within a prepositional phrase.

- *mot : paddan* [PROMOTED]
  (towards : toad+DEF)

Further, certain nouns denoting close relationship, such as *mamma* (mum), *mormor* (grandmother) and *faster* (aunt), as well as names of months may constitute a definite noun phrase without any premodifying attributes, and are then functioning almost as proper nouns. Thus the grammar promotes these nouns where noun phrases are expected, e.g.:

- *i : januari* [PROMOTED]
  (in : January)
- *utan : mamma* [PROMOTED]
  (without : mum)

An indefinite noun phrase consisting of a single noun is most often considered a naked noun phrase with restricted use [see section 7.2.1]. Indefinite nouns may only constitute a non-naked indefinite noun phrase on their own if they are uncountable or in the plural form. Since there is no distinction made between countable and uncountable nouns in the current lexicon all singular indefinite nouns are regarded as naked and thus not promoted. Indefinite nouns in the plural are however distinguishable and thus promoted.

- *med : mjölk* [UNKNOWN]
  (with : milk)
- *på : paddor* [PROMOTED]
  (on : toads+INDEF)

7.3 Rules for verb valency and inflection

In the following sections, we will describe rules for handling verb inflection and verb valency.

7.3.1 Verb inflection in the nexus field

The kernel constituent of a clause is the finite verb, which normally occurs at second position in a Swedish main clause. Since we know that the verb must be finite at this position, infinite verb forms may be suppressed. It is, however, not a trivial matter to determine when the fundament constituent is complete and the finite verb ought to follow. This is the same problem as arose when defining rules for promoting missing verbs (see section 7.1.2) and a similar strategy has been used for the suppression of infinite verb forms. Consequently, we have restricted ourselves to handle this kind of verb inflection at sentence beginning and only when the fundament is holding a noun phrase, a prepositional phrase or an adverb phrase:

- *Han : spela/spelat* [REJECTED]
  (He : play+INF/play+SUP)
- *I morgon : spela/spelat* [REJECTED]
  (Tomorrow : play+INF/play+SUP)
- *Ofta : spela/spelat* [REJECTED]
  (Often : play+INF/play+SUP)

7.3.2 Verb inflection in the content field

As stated in section 7.1.2, the promotion of head verbs in clauses holding auxiliaries, is problematic. The rejection of erroneously inflected verb forms is, however, less risky, since these do not coincide with any of the words that may intermediate the auxiliary and the head verb. Therefore, a rejecting set of rules is implemented to suppress erroneously inflected verb forms in the content field, regardless of whether there is an intermediate subject or clause adverb:

- *Han ska: sjunger/sjungit* [REJECTED]
  (He will : sings/sung)
- *I morgon ska han inte: sjunger/sjungit* [REJECTED]
  (Tomorrow he will not : sings/sung)

The rejecting rules keep track on what verb forms to reject in an arbitrarily long chain of auxiliaries, e.g.:

- *Han skulle vilja ha kunnat: sjunger/sjungit* [REJECTED]
  (He would have wanted to be able to : sings/sung)

Turning to infinitive phrases, the content field must hold an infinitive verb form and it is feasible to suppress finite and supine verbs. The infinitive marker (*att*) and the infinitive verb may only be separated by a clause adverbial, whereby the localisation of the verb forms to suppress, may seem straight-forward. However, the infinitive marker has a homograph constituting a subjunction, and it is often not possible to disambiguate it from only the left context and no detailed valency information. If it is a subjunction, the subject of the subordinate clause usually follows, but sometimes a finite verb occurs, and should not be suppressed, as in: *Han skrek att springer du så slår jag dig* (He yelled that if you run I will hit you). Either way, a supine verb is highly unlikely and is therefore suppressed.

- *Det skrämde mig att : sjungit* [REJECTED]
  (It frightened me that : sung)

7.3.3 Verb valency

Since valency information is not available in the lexicon used, verb complements are currently neither promoted in the context of transitive verbs nor suppressed after intransitive verbs. The limited set of copula verbs are however marked as such, enabling us to implement some controlling mechanisms in connection to the prediction of mandatory predicatives.

There are mainly two kinds of predicatives in copula constructions: adjectival and nominal predicatives (Jörgensen and Svensson 1995). An adjectival predicative is to agree in gender and number with the subject of the heading copula verb, as in *Hunden är brun* (the dog+UTR+SING is brown+UTR+SING) . This restriction does, however, not hold for nominal predicatives though they usually share number with the correlating noun phrase: *Mina vänner är ?doktor* (my friends are doctor). Since we do not employ full parsing and it is necessary to isolate the subject for the copula restrictions to be applicable, we only handle predicatives in sentence initial main clauses. In connection to these, singular non-agreeing adjectival predicatives are suppressed:

- *I morgon blir huset: såld* [REJECTED]
  (Tomorrow the house +NEUTR will be : sold+UTR)
- *Huset ska bli: såld* [REJECTED]
  (The house+NEUTR will be : sold+UTR)

Lots of neuter adjectives are homographs to adverbs that may modify an adjective, e.g. *Polisen var onödigt försiktig* (The police was unnecessarily/unnecessary+NEUTR careful). In the current lexicon these word forms are not separately listed as adverbs and there is no feature marking these adjectives as homographs. In order not to suppress any appropriate words we have chosen to let all neuter adjectives pass, so even those that are quite unlikely as adverbs, such as *ledset* (sadly/sad+NEUTR).

Plural adjectives are not being checked for agreement, since these may initiate a noun phrase, that in turn, does not have to agree with its correlate. Singular adjectives may initiate a naked noun phrase or a noun phrase with an uncountable head noun. However, since these noun phrases preferably are constituted by a single noun (Teleman et al. 1995), we believe a singular adjective, at the position in question, to only rarely function as the initiator of a noun phrase.

Even though there are only a few words that may unambiguously be judged erroneous in copula constructions, there is a secondary reason to parse adjectival predicatives. If the fundament holds an adverbial and the subject is succeeding the copula, the rules for agreement within the noun phrase may erroneously interpret the adjectival predicative as an inappropriate noun phrase continuation. Compare the noun phrase fragment: *det *roligt trollet* (the+DEF fun+INDEF troll+DEF) with *Ofta är det roligt* (Often it is fun). In the latter example the adjective ought not to be suppressed. Therefore possible adjectival predicatives are given a neutralizing parse, so as to neutralize the effect of the agreement error rules.

Parallel to the suppression of non-agreeing singular adjectives, it may be of use to promote correct ones. As already implied a predicative in a copula construction may be of various forms. Apart from being an adjective or a noun phrase, it may be a prepositional phrase or a subordinate clause, and, as always, it is risky to promote words in such indeterminate contexts.

7.4 Rules for prepositional phrases

Errors in the prepositional phrase comprise less than one tenth of the attested error sample and is the smallest error class that we have dealt with on the basis of its frequency. As stated, the errors in the prepositional phrase either involves a suggestion not initiating a nominal constituent or a suggestion initiating a nominal constituent of the wrong subtype. So far, we have only approached the former of those. To cope with the latter, detailed valency information of verbs heading the prepositional phrases is

required, as well as some subcategorization of the prepositions themselves, something which the current lexicon does not provide for.

Prepositional phrases are constructed by a preposition followed by a government, that in general is a noun phrase, a nominal subordinate clause or an infinitive phrase [8]. At positions succeeding a preposition we hence promote words that may initiate any of the above constituents:

- *Jag tror på : tomten* [PROMOTED]
  (I believe in : Santa Claus)
- *Jag tror på : att* [PROMOTED]
  (I believe in : that)

The only words that are currently suppressed after a preposition are pronouns in the subject case [9]:

- *Jag tror på : han* [REJECTED]
  (I believe in: he)

In the context of split prepositional phrases e.g. *honom tror jag på* (him I believe in) the promotion of governments will be overly applied, the effect of which will have to be further investigated.

7.5 Rules for adverb phrases

Even though adverb phrases were not a prominent error class, we have had reason to implement some rules to parse them at sentence beginning for the purpose of promoting/rejecting constituents in their near context. While doing that we added a restriction, that must be fulfilled for the adverbs not to be suppressed.

There are intricate rules for what other adverbs a given adverb may modify, most of which belong to semantics and are hence difficult to formalize. One restriction is however quite clear; the modifying adverb often imposes certain grades of comparison on the head adverb. For instance *ganska* (quite) may not modify words in the comparative or superlative and *allra* (the most) may only modify words in the superlative. Therefore we have chosen to suppress adverbs in the inappropriate grade of comparison when these are part of a (sentence initial) adverb phrase:

- *Ganska :* {*hellre/helst*} [REJECTED]
  (Quite : rather+COMP/rather+SUP)
- *Allra :* {*bra/bättre*} [UNKNOWN]
  (the most : good/better)

Caution has been taken though, so as not to suppress actually wanted adverbs. Let us assume that the user has entered *ännu* (yet). *ännu* imposes its head adverb to be in the comparative, thus making positive adverbs, such as *mycket* (much) erroneous. However, *mycket* may in turn modify a comparative adverb, which will then be the head adverb, e.g.: *ännu mycket bättre* (yet much better). The problem is thus, that we do not know the hierarchical structure of the complete adverb phrase: (ännu mycket) or (ännu (mycket ...)). When the latest inserted adverb does not have the property of modifying other adverbs, we may however exclude the latter structure and it is possible to suppress inappropriate grades of comparison. Thus we do not suppress *mycket* in the above context, but adverbs such as *sällan* (seldom):

[8] Rarely occurring as governments within a prepositional phrase are also adjective/participle phrases, prepositional phrases, adverb phrases and adverbial subordinate clauses. (Teleman et al. 1995)

[9] The only prepositions that may take governments explicitly in the nominative, are *utom* and *förutom*, both meaning except for, as in *alla utom jag* (all except for me+NOM). (Teleman et al. 1995)

- *Ännu: sällan* [REJECTED]
  (Yet : seldom)
- *Ännu: mycket* [UNKNOWN]
  (Yet : much)

Complicating matters further, a third adverb may be restricted by both preceding adverbs, as is the case in a structure like: (ännu (mycket (bättre))) or only by the immediately preceding adverb, as in ((ganska mycket) bättre). As we do not know what structure is intended, we never suppress an adverb on the basis of adverbs other than the immediately preceding one. This entails that there will be less coverage of the rule. For instance, adverbs in the positive, like *bra* (good) will pass unnoticed in the context of *ännu mycket*, since they are in accordance with the closest adverb *mycket*.

Chapter 8

Evaluation

To what extent a word prediction system actually facilitates the process of text production is probably best measured in relation to a specific user. One user may emphasize the physical effort saved when choosing from a list, whereas another user may have difficulties scanning a list and thereby finds it of vital importance that the intended words are presented at top of the list. For people with language impairments the most important factor may be that no ungrammatical suggestions are displayed. Most evaluations are however relying on crude, statistic measurements, partly due to these being easier to estimate, and partly due to subjective experiences being less general and harder to assess in relation to other system evaluations.

In line with the project specification, we have evaluated the FASTY grammar model by means of the most commonly used evaluation metric for word prediction systems: Keystroke Saving Rate (KSR). KSR estimates the physical effort saved, which is only one aspect influencing user satisfication. Another relevant aspect might be the level of distraction imposed by grammatically incorrect suggestions. A larger evaluation in relation to specific users will be carried out in a later stage of the project, the results of which will be published during the autumn of 2003.

The KSR metric will be described in the following section, whereafter our test results will be presented.

8.1 Keystroke Saving Rate

The Keystroke Saving Rate is defined as the number of keystrokes saved when a text is entered using a word prediction system as compared to the keystrokes needed to enter the same text without any word prediction facilities.(Wood 1996) KSR estimations are usually made automatically by means of software simulating the behaviour of a user entering a text using a standard keyboard.[1] This way large test texts may be used and the simulation may be rerun with different settings.

Generally, the KSR is computed with a formula like the following:

$$KSR = 1 - \frac{NmbrOfKeystrokes}{NmbrOfChars}$$

where the numerator accounts for the keystrokes needed to enter the test text and the denominator represents the number of characters comprising the test text. On a detailed level, simulators may define the terms differently (Wood 1996)(Cagigas 2001). Often the relation between characters and keystrokes is simplified so that one character is assumed to always cost one keystroke, even though some characters, such as capital letters, require two keystrokes.

An important factor when computing the KSR is the assumed list length, as the figure normally increases with its size. Evaluation reports generally includes estimations with a list length of five. This is

[1] Even though potential users may use other input devices, such as a single switch, the simulated effect of using a standard keyboard will probably reflect general tendencies.

assumed to be the maximum number of items an average user may overview at a glance.(Cagigas 2001) Estimations with a list length of one are further of particular interest, since this reflects how often the intended word form is ranked the highest.

8.2 Test results

We have evaluated the Swedish FASTY grammar in relation to three test texts. The test texts belong to different genres and consist of a review, a news paper article and an extract from a short story (see Appendix B). Except for removing headings, the texts have been left unmodified.

The KSR was esimated using the Simple Word Predictor (SWP)(FASTY 2003), a test platform developed within the FASTY project. We used a simplified mapping between characters and keystrokes (one character equals one keystroke) and assumed the user to utilize implemented functions for auto-capitalisation at sentence beginning and the insertion of a white space after punctuation marks.

To assess the effect of the Swedish grammar, we first run the test texts with a language model solely based on n-gram statistics, whereafter we added the grammar checking module to the same statistic model. The relative weights used for the different types of n-grams are given in Appendix A. As is customary, we ran the tests with a list length of one, as well as of five. The results are listed in tables 8.1 and 8.2 respectively.

Test text	**I suggestion**			
	n-grams	n-grams + grammar	diff (points)	diff (percent)
Review	34.55	34.89	+0.34	+0.98
Article	31.09	31.82	+0.73	+2.35
Short story	28.17	28.35	+0.18	+0.64
Average	31.27	31.69	+0.42	+1.34

Table 8.1: Evaluation results for a list length of one

Test text	**5 suggestion**			
	n-grams	n-grams + grammar	diff (points)	diff (percent)
Review	47.78	48.04	+0.26	+0.54
Article	44.31	44.33	+0.02	+0.05
Short story	41.60	41.87	+0.27	+0.65
Average	44.56	44.75	+0.19	+0.43

Table 8.2: Evaluation results for a list length of five

As can be seen, there is a slight improvement when the grammar module is added. The difference is however too small to be regarded significant. The improvement is somewhat larger, both in terms of percent points and percents, when the suggestion list is limited to one suggestion. This indicates that the strength of the grammar module lies in the reranking of displayed suggestions, so that the intended suggestion is at top of the list, rather than actually adding the suggestion to a long list of suggestions. Since the grammar module does not produce any suggestions, but reranks a limited list of statistically based suggestions, this is to be expected. The number of suggestions input to the grammar module has been set to 10 plus the number of suggestions to be displayed, a figure that may be to small and is yet to be optimized.

The somewhat discouraging results are not completely unexpected. The results for Spanish, as presented in (Cagigas 2001), only shows an enhancement of 0.5 percentage points when adding grammar to a full-blown n-gram model. For English, (Wood 1996) has reported an improvement of 5 percentage points. The statistic model, to which the English grammar was added, was however solely based on word form unigrams, rendering the basis for improvement uncomparable to ours.

As should be evident from the error classification and the grammar description, there are quite few contexts in which we may be certain what syntactic structure is under construction, and therefore what grammatical categories to either suppress or promote. This is partly due to limitations inherent to word prediction, i.e. only the left context is accessible, and partly due to the lack of valency information in the lexicon used. The lack of valency information further imposes the grammar rules to be rather locally defined, capturing approximately the same contexts as the n-gram model does.

Another explanation to the low figures, may be that a word suggestion often may not be judged totally incorrect, but rather more or less unlikely, whereby crude suppression is a too blunt tool. Promotion is, similarly, not granular enough to give a true picture of language variation. Since the promotion of one word form imposes that other word forms are pushed down, it is of vital importance that all, or none, of the appropriate word forms are promoted. An appropriate suggestion may however be more or less likely, and a relative weighting of the promotion rules may give a better result.

In its current state, the FASTY grammar module provides no means for handling multi-word units nor homographs, which may be further reasons to the results. A multi-word unit often imposes other expectations on the context than does the corresponding compositional unit. For instance, the last token in the adverbial multi-word unit *till och med* (even) normally functions as a preposition, but in the adverbial meaning no government is expected.

As regards the homograph problem, all interpretations of a word form are parsed simultaneously, and the resulting charts are scanned together. In case a word form has one common interpretation, not covered by any grammar rules, and a less common interpretation that is suppressed, the total effect will be that the word form is suppressed. As for illustration, the word form *för* has several interpretations, out of which one is an adverb and one is a present tensed verb. In accordance with the rules for verbs out of position (section 7.1.1) the verb interpretation is suppressed in the following context: *Hon gick: för (långt)*, even though the adverb interpretation, not covered by any rule, makes the sequence perfectly in order.

The large size of the lexicon aggravates the homograph problem. Even highly infrequent interpretations are covered, sometimes competing with more common ones. The word form *flicka*, is for instance listed both as a noun, meaning girl, and as a very unusal verb, meaning to patch. In its current state the grammar module and the split compound prediction module are not integrated. During our evaluation the compound prediction module has thus been turned off. In order to make the grammar module and the compound module compatible, strategies will have to be worked out for handling the agreement conflict imposed by the prediction of the first part of the compound, the modifier. The base form of a lexeme often coincides with its compounding form, and whereas the former has to agree with potential pre-modifiers, the latter does not.

Chapter 9

Concluding remarks and future development

In this thesis we have defined and implemented a Swedish grammar for word prediction. What structures to cover has been decided on the basis of error type frequencies estimated from logged prediction suggestions.

A preliminary evaluation of the grammar module only showed a slight improvement in keystroke saving rate. The improvement was somewhat larger when the suggestion list was limited to one suggestion only, as compared to a list length of five. This indicates that the strength of the grammar module lies in the reranking of already displayed suggestions, rather than the addition of new suggestions to a long list of suggestions. The grammar module does not produce any suggestions by itself, but filters a limited set of statistically based suggestions and the optimal size of this set is yet to be determined.

The left context has shown often not to be sufficient for determining what syntactic structure is under construction, and thereby what constituents to either suppress or promote. Aggravating the problem, the reranking is based on a coarse-grained trinary classification of the prediction suggestions, that may be a too blunt tool to capture the variation of language. Further explanations to the disappointing figures may lie in the absence of verb valency information in the lexicon and the inability of the system to handle multi-word units and homographs.

As regards future developments, there are several matters to be explored, possibly improving the performance of the grammar module. First, the promoting and rejecting sets of rules ought to be evaluated separately, so as to judge whether either set influences the results in a negative way. In particular, this may hold for the promoting set of rules, since these may have side-effects not foreseen, as the promotion of one constituent simultaneously imposes other constituents to be pushed down.

Secondly, the number of suggestions accessible to the grammar module (the prediction list to be reranked), determines the impact of the grammar module, and the optimal number is yet to be estimated.

Further matters to develop involves the already mentioned lack of verb valency information in the lexicon. If the valency frame of the verbs were provided, the grammar rules could be expanded to cover significantly larger structures. KSR only estimates the physical effort saved, and not user satisfication in general. Clues on the latter, may be provided at a later stage of the project, as a large user-evaluation will be carried out.

Bibliography

Baroni, M., Matiasek, J. and Trost, H. (2002). Wordform- and classbased prediction of the components of german nominal compounds in an aac system, *Proceedings of COLING 2002*.

Cagigas, S. P. (2001). *Contribution to word prediction in Spanish and its integration in technical aids for people with physical disabilities*, Phd dissertation, Madrid University, Madrid.

Carlberger, J. (1997). *Wordpredict: Design and implementation of a probabilistic word prediction program*, Master's thesis, Royal Institute of Technology, Stockholm.

Carlberger, J. and Hunnicutt, S. (n.d.). Improving word prediction using markov models and heuristic methods, In press.

Dahlqvist, B. (1998). The scarrie swedish newspaper corpus, *in* A. Sågvall Hein (ed.), *Working Papers in Computational Linguistics & Language Engineering*, Vol. 6, Uppsala University, Department of Linguistics.

Demasco, P. and McCoy, K. (1992). Generating text from compressed input: An intelligent interface for people with severe motor impairments, *Communications of the ACM*.

FASTY, Primary author Department of Linguistics, U. U. (2003). Second edited annual report for publication. Deliverable D2.6, FASTY IST-2000-25420.

Fazly, A. (2002). *The use of syntax in word completion utilities*, Master's thesis, University of Toronto, Toronto.

Jelinek, F. (1991). Up from trigrams!, *Eurospeech '91*.

Jörgensen, N. and Svensson, J. (1995). *Nusvensk Grammatik*, Gleerups.

Klund, J. and Novak, M. (2001). If word prediction can help, which program do you choose? Available at: http://trace.wisc.edu/docs/wordprediction2001/index.htm.

Kronlid, F. and Nilsson, V. (2000). *Treepredict*, Master's thesis, Göteborg University, Göteborg.

Laine, C. and Bristow, T. (1999). Using manual word-prediction technology to cue students' writing: Does it really help?, *Proceedings of the Fourteenth International Conference on Technology and Persons with Disabilities*, Los Angeles, CA: California State University, Northridge.

Lesher, G. and Rinkus, G. (2001). Domain-specific word prediction for augmentative communication, *Proceedings of the RESNA 2001 Annual Conference*, Reno.

Masui, T. (1999). Pobox: An efficient text input method for handheld and ubiquitous computers, *Computer Science* **1707**: 289–300.

Olsson, L.-J. (1999). A swedish wordform dictionary. Project Report 3.2.3, EC-project Scarrie.

Rambell, O. W. (1998). Error typology for automatic proof-reading, *in* A. Sågvall Hein (ed.), *Working Papers in Computational Linguistics & Language Engineering*, Vol. 4, Uppsala University, Department of Linguistics.

Rosenfeld, R. (1994). *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*, Phd dissertation, Carnegie Mellon University, Pittsburgh.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modeling, *Computational Lingustics* **22**(1): 39–71.

Sågvall Hein, A. and Starbäck, P. (1999). A test version of the grammar checker for swedish, *in* A. Sågvall Hein (ed.), *Working Papers in Computational Linguistics and Language Engineering*, Vol. 12, Uppsala University, Department of Linguistics.

Sjöberg, P. (2001). *Word prediction in an internet chat*, Master's thesis, Uppsala University, Uppsala.

Starbäck, P. (1999). Scarcheck - a software for word and grammar checking, *in* A. Sågvall Hein (ed.), *Working Papers in Computational Linguistics and Language Engineering*, Vol. 12, Uppsala University, Department of Linguistics.

Teleman, U., Hellberg, S. and Andersson, E. (1995). *Svenska Akademiens Grammatik*, NorstedtsOrdbok.

Thorell, O. (1973). *Svensk Grammatik*, Esselte Studium.

Weijnitz, P. (1999). Uppsala chart parser light: System documentation, *in* A. Sågvall Hein (ed.), *Working Papers in Computational Linguistics and Language Engineering*, Vol. 12, Uppsala University, Department of Linguistics.

Wood, M. (1996). *Syntactic Pre-processing in Single-word Prediction for Disabled People*, Phd dissertation, University of Bristol, Bristol.

Appendix A

Relative weights used in the statistic language model

```
#------------------------------------------------------------
#  [extract from] Configuration file for SWEDISH
#------------------------------------------------------------
#------------------------------------------------------------
# Prediction Weights
#------------------------------------------------------------
# --- interpolation toplevel
word_unigram_weight .2 # weight of word unigrams
word_bigram_weight .6 # weight of word bigrams
word_tag_weight .2 # weight of P(w|t)


# --- relative weights of general and user dictionary
gdict_weight .8 # weight of general dictionary
udict_weight .2 # weight of user dictionary

# --- P(w|t) interpolation
tag_uni_weight .1 # weight of word|tag from word unigrams
tag_bi_weight .9 # weight of word|tag from word bigrams

# --- tag model
q1 .1 # weight of tag bigrams
q2 .9 # weight of tag trigrams
```

Appendix B

Test texts

B.1 Review

[*Amelie från Montmartre* Downloaded from www.bio.nu, 250503]
Amelie bestämmer sig för att charma och göra gott. Amelie arbetar på ett litet café i Paris. Hon känner sig ensam. Förhållanden har inte gett henne någon lycka. Men en dag när hon upptäcker ett skrin med en mans gamla barndomsminnen förändras allting. Hon bestämmer sig för att försöka hitta mannen och återlämna skrinet. Mannen blir överlycklig av att få återse det borttappade skrinet. Amelie bestämmer sig för att göra människor i sin omgivning lika glada.

Amelie från Montmartre är en rakt igenom helskön och sympatisk film. Melodramisk, varm, romantisk och mycket humoristisk. Tempot är från första början högt uppskruvat och man får koncentrera sig för att uppfatta alla små genialiska detaljer. Amelie från Montmartre är en film som inte påminner om något annat. Där finns också dess styrka. Det mycket visuella, estetiska, kitchiga och lite sagolika över den här filmen gör den till något helt underbart.

Regissören Jean-Pierre Jeunet är mest känd för att ha regisserat filmen Alien: Resurrection från 1997. Men redan innan han blev tillfrågad att göra Alien av Fox, hade han börjat spåna på några lösa idéer som senare skulle bli Amelie från Montmartre. Han hade redan då en mängd idéer för scener, situationer, karaktärer men inte någon gemensam nämnare för filmen.

Sockersöta huvudrollen Amelie spelas imponerande av Audrey Tautou. Hennes motspelare Nino spelas av Mathieu Kassovitz. Filmen innehåller mängder med skruvade karaktärer. En svartsjuk man som spelar in allt som sägs i Amelie's café tex.

Amelie från Montmartre är en historia skriven nästan som en saga som berör samtidigt som den inte på något sätt kan beskyllas för att vara djup. Trivsam och godhjärtad är den däremot utan tvekan.

En del scener etsar sig fast i minnet. Framförallt Amelie's lite fega försök att få kontakt med den hon intresserar sig för. Han samlar på misslyckade foton från fotoautomater som folk kastat ifrån sig. Hon tar kort av sig själv i fotoautomater som hon river sönder och lämnar vid automaten för honom att upptäcka. Hon skriver vykort till honom där hon stämmer träff med honom. Men så fort hon kommer i närheten av att avslöja sig för honom blir hon skrämd och vågar inte stå för sin invit.

Alla små detaljer och finurliga scener väcker definitivt ens intresse och beundran inför regissören. Filmeffekterna, de snabba vändningarna, de varma färgerna och en rapp humor bidrar säkert till en ordentlig kultstatus även utanför hemlandet Frankrike.

1930-50-talsestetiken är passande i den här filmen och man kan inte annat än gilla denna mycket visuella film.

Amelie från Montmartre är en färgsprakande glädjespridare som bara måste ses.

B.2 Article

[*Bosättare tror inte att marken återlämnas*, by Lotta Shüllerqvist, 03062003, Downloaded from www.dn.se 030603]

Från berget Herodion sydöst om Betlehem ser man berg och dalar, och i fjärran skymtar Jordaniens violetta berg. Här ligger palestinska byar sida vid sida med israeliska bosättningar och bosättarutposter av baracker och husvagnar. Vi åker ner mot en av utposterna men blir inte insläppta. I stället följer vi med en ung man vi möter, Noam Lowy, hem till hans hyrda hus i Nokdim, en av de bosättningar som Sharon lovat utrymma i utbyte mot fred med palestinierna.

- Det är svårt att förklara för utomstående vad som pågår här. Vi bosättare beskrivs som våldsamma och oresonliga, men så enkelt är det inte, säger Noam.

Han talar om palestinierna, om deras kamp för frihet och ett riktigt liv, som han kan förstå.

- Men vi judar har också starka religiösa och ideologiska band till det här landet, som vi har återvänt till efter två tusen år. Det tillhörde våra förfäder, och att ge bort det skulle vara ett brott mot Guds vilja. Palestinierna hatar oss inte för att vi tar deras land, utan för att vi är judar. De vill fördriva oss härifrån, inte leva i fred med oss.

Noams beskrivning av bosättarlivet är full av rädsla och oro:

- Jag är skräckslagen när jag åker härifrån, min fru är rädd när jag är borta och jobbar på nätterna. Barnen som går i skola i Jerusalem åker i pansarbussar med militäreskort. Bara under förra året dödades tolv av mina vänner, av bussbomber, under bilresor på Västbanken eller under militärtjänsten.

Hur uppfattar han då Sharons besked om utrymning av Nokdim?

- Jag är inte överraskad. Sharon är utsatt för hårt tryck nu. Jag tror inte att det kommer att förverkligas, men om han låter det ske förlåter jag honom aldrig.

Noam är dock inte beredd att försvara sin tillvaro med våld:

- Väpnat våld mellan judar är otänkbart för mig, men det finns de som har annan uppfattning.

Försedda med en flaska kallt vatten åker vi vidare till den närbelägna bosättningen Tekoa där vi träffar rabbinen Menachem Frohman, som är en av bosättarrörelsens grundare. Han föddes i Israel strax efter andra världskriget och deltog som fallskärmssoldat i erövringen av Jerusalem:

- Erövringen, eller befrielsen, det beror på hur man vill se det, säger han och skrattar i sitt röriga vardagsrum som är fullt av leksaker. Han har tio barn och barnbarn i huset.

Rabbi Frohman har bott 25 år i Tekoa, och han tar Sharons besked om utrymning med lugn:

- Avveckling har alltid funnits som bakgrundsmelodi i Tekoa. Det har ju kommit många fredsplaner från olika presidenter genom åren. Och även den här gången har vi äran att stå på listan över avvecklingsobjekt.

Men rabbi Frohman har en helt annan vision av hur konflikten ska lösas - genom kommunikation och förnuft. Han har under många år fört samtal med Arafat och andra arabiska ledare - inte på arabiska, men med ”hjärtats språk”, säger han och förklarar:

- Fred på jorden är människornas ansvar, och vägen dit går genom våra religioner - där finns grunden för ömsesidig förståelse och respekt.

- Med Guds hjälp skulle vi i detta lilla land kunna bygga en bro mellan europeisk och arabisk kultur och göra Jerusalem till en gemensam fredens huvudstad. Tid och människans förmåga till öppenhet är de resurser som behövs, säger rabbi Frohman innan han hastar i väg till en tv-intervju.

B.3 Short story

[*Jag rider ut i världen på min ko* in *Håll svansen högt* by Ulf Nilsson, Downloaded from www.eboken.nu, 250503]

När jag vaknade mitt i natten hemma hos farmor, kände jag mig liten och ensam. Mamma låg på BB sedan fem dagar tillbaka och jag bodde ensam hos farmor. Jag låg i kökssoffan, jag frös, jag hörde mössen

knapra på en ostbit inne i skafferiet, jag hörde hur det knakade i golvbrädorna på vinden som om stora spöken gick fram och tillbaka och stånkade. Min lilla katt Esmeralda var inte hos mig. Hon tyckte nätter var så spännande, då gick hon ut och smög efter små tassemöss. Jag längtade så efter mamma. Hum hum, sa jag. Mamma brukade alltid höra minsta lilla pip från mig och genast svara med ett ”hum hum”. Det betydde: ”Det är ingen fara, du kan komma hit och sova i min säng.” HUM HUM, sa jag mycket högre. Ingen svarade. Mamma låg på BB. Och pappa arbetade natten på tullen. Det är så konstigt, man kan vara en modig en som ensam seglar ut i stormiga natten, en våglig en som kan prata med troll, men så är man bara en liten pipsill när man vaknar ensam i en säng. Jag skyndade mig upp och tassade in till farmor. Men jag blev rädd när jag såg henne. Min farmor var stor och tjock när hon satt i korgstolen. Jag trodde alltid att hennes tjocka rumpa skulle fastna i stolen, så när hon reste sig satt den kvar på baken. Men det hände aldrig. Min farmor var stor och tjock och modig och kunde jaga bort arga hundar med sin käpp. Men nu hade hon tagit av sig alla skjortor, blusar, klänningsliv, tröjor, koftor, kjolar, klänningar och förkläden, alla dessa konstiga sorters kläder som min farmor hade. Utan kläder var hon nog inte så tjock egentligen. Hon låg alldeles stilla och platt under täcket. Hennes ansikte såg också annorlunda ut nu. Hon hade tagit av sig glasögonen och hörapparaten och löständerna. Löständerna låg i ett glas med vatten på nattduksbordet och skrämdes. Hennes käpp stod lutad mot väggen. Jag ropade på henne, men hon hörde inte för hörapparaten låg ju på bordet. Och om hon hade hört mig, så hade hon inte sett mig, för hon hade ju inga glasögon. Och inte hade hon kunnat skrämma bort spöken nu när hon inte hade några tänder. Jag stod stilla på det kalla golvet och visste inte vad jag skulle göra. Jag kände mig ensam och svag. Jag tog farmors käpp i handen, så jag hade något att försvara mig med om spökena kom. Jag kom på att jag kunde stoppa in hennes löständer i min mun. Då skulle jag få ett riktigt rovdjursgap som kunde skrämma de flesta. Fast jag vågade inte röra vid hennes löständer. Jag tog på mig hennes glasögon och allt blev stort och suddigt. Så stoppade jag den lilla öronmusslan i örat och skruvade på hörapparaten. Nu skulle ingen kunna överraska mig. Jag såg bättre och hördebättre än alla. Käppen hade jag i näven. Hörapparaten tjöt och jag skruvade ner volymen lite. Nu hörde jag hur det knäppte och knakade i hela huset. Överallt fanns det små liv, och det var tusentals små musfötter som tassade fram över vindsgolvet, tusen små mushjärtan som bultade och tusen hungriga små magar som knorrade. Jag satte ifrån mig käppen. Jag frös och ville krypa ner till farmor, men vågade inte. Hon var säkert inte så mjuk som mamma. Hon luktade nog helt annorlunda. Farmor låg alldeles stilla. Tänk förresten om farmor hade dött under natten. Hon rörde sig inte det minsta och andades inte. Jag försökte höra om hennes hjärta bultade och skruvade upp hörapparaten tills den började tjuta igen. Jag stod där och väntade och väntade. Spökena stånkade runt på vinden och mössen prasslade i väggarna och åt ost i skafferiet. Till sist bestämde jag mig för att farmor faktiskt var död. Det var fruktansvärt! Då plötsligt snarkade hon till som en stor gris. Jag sprang tillbaka till min kökssoffa. Jag var så ensam och kall och längtade så. På morgonen gick jag ut i stallet. Mjölkmaskinen var redan igång, korna stod och mumsade och grisarna skrek när farbror Gustav skoffade utfoder till dem. De var som vilda när de fick mat. De trängdes över mathon och fodret rann in i ögonen och näsan på dem. Men de brydde sig inte, de älskade foder. En nös och det stod ett moln av mjöl runt dem. Vet du vem jag skulle vilja vara allra bästa vän med? frågade jag farbror Gustav. Nej. Mamma, sa jag. Är du inte det då? Jo, men hon är ju inte här! sa jag. Ska en allra bästa vän bara försvinna? Mamma är på BB och min katt vill bara ut och smyga på möss. Om hon ska ha ett litet barn på BB så måste hon väl vara borta. Och katter går som dom vill... Jag svarade inte. Men vem skulle annars vilja vara bästa vän med, frågade farbror Gustav, om du inte får välja din mor? Jag gick runt i stallet och funderade. Farbror Gustav hade åtta kor. Sju stora och en liten. De stora hette Rosa, Broka, Flora, Vera och andra vackra namn. Den lilla kon hette Knaggen. Knaggen var liten till växten och skulle aldrig bli någon stor ko som de andra. Jag klappade Knaggen på halsen. Hon var en annorlunda ko och den ende jag vågade gå in till. De andra blev rädda och trampade runt och slängde med sina stora huvuden. Men Knaggen var lugn och varm. När jag såg Knaggen i ögonen, så tittade hon alltid tillbaka, de andra korna flackade med blicken och drog sig bakåt. Jag lade kinden mot Knaggen, hon luktade gott, av mjölk och sommargräs. Vem skulle du välja? frågade farbror Gustav igen. Knaggen, sa jag. Man kan väl inte vara vän med en ko! svarade han och skrattade. Varför inte? Kor ska bara finnas, man har ingen nytta av att vara

vän med dem. Så är det. Fast om du vill kan du få Knaggen av mig. Bara ta henne du! Farbror Gustav skrattade. Skulle jag verkligen få Knaggen? Skulle Knaggen bli min alldeles egna ko. Jag strök henne över hennes ludna och varma öra Du är den finaste ko som finns! viskade jag. Vi ska visa farbror Gustav att en ko är det bästa som finns att vara vän med ... Jag såg henne i ögonen. Ingen ko hade så snälla ögon som Knaggen. Inga andra kor hade ögonfransar som en fin dam. Jag borrade min näsa mot hennes kind och hon slickade mig. Hennes tunga var sträv sompappas haka. Måste hon vara kvar hos de andra korna? frågade jag allvarligt. Du kan väl ha henne här tills du hittat något bättre, hehe, svarade han. Jaja, tosigheter finns det gott om... Jag var åtta år gammal och lycklig ägare till en liten ko. Jag tyckte faktiskt inte att Knaggen skulle vara kvar i stallet. De andra korna var så feta och präktiga, de tyckte nog att hon var en eländig spink. Man vet aldrig vad de sa när de brölade till varandra därinne istallet. Ute var det kallt och fruset. Jag gick runt och letade efter ett bra ställe där jag kunde ha min ko. I hönshuset, nej, därinne var ett evigt kackel och tuppen gol i ett. I vedbon var det hårt och pinnigt. Och i traktorgaraget luktade det bensin. På fältet bakom stallet fann jag en hög med höbalar under en presenning. Balarna var fyrkantiga som jättelika byggklotsar. Som byggklotsar... Vad skulle man kunna göra med jättestora byggklotsar av hö... Jag skulle ju kunna bygga mig ett hus, ett hus av hö på fältet! En höbal var inte särskilt tung. Runt den fanns ett snöre och om jag högg tag i det så kunde jag dra den dit jag ville. Jag fick hålla på hela förmiddagen innan jag hade fått upp väggar som på en igloo och en öppning som skulle vara dörr. På eftermiddagen lade jag ut några brädor på taket och baxade sedan upp höbalar ovanpå dem. Ett vackert runt höhus hade jag nu på fältet! Jag tog itu en bal och spred ut höet på golvet så det skulle bli varmt och skönt. Så släpade jag fram två extra höbalar som skulle kunna passas in i dörrhålet. Åh, det skulle bli så varmt och fint i Knaggens lilla hus. Här skulle hon få det bra! Jag stötte på farbror Gustav när jag skulle gå in i stallet igen. Hehe, hur går det för koägaren? sa han. Jag bara nickade. Kor är liksom ingenting, sa han då. Bara en mun som äter och en stor rumpa som skiter. Det bästa med dom är att man kan få mjölk från mitten. Jag gick in och hämtade Knaggen. Bredvid mig var hon jättestor, jag var rädd att bli trampad på tårna av hennes stora klövar. Och jag fick akta mig för att inte få hornen i sidan. Knaggen blev förvånad när jag tog loss de två kättingarna som satt med en karbinhake i halsremmen. Farbror Gustav hade sagt att man måste vara bestämd mot kor. Man måste säga till på skarpen, annars springer de sin väg. Så jag drog i halsremmen och Knaggen spjärnade emot så det skulle behövts tio man för att få henne ur fläcken. Jag satte mig ner och såg henne i ögonen. Jag ska inte dra i dig, du kommer att tycka det är roligt, sa jag lugnande. Jag kunde inte kommendera min egen ko, och jag ville inte heller. I stället tog jag det lugnt och väntade. Efter ett litet tag blev hon nyfiken och började backa ut ur båset. Nu höll jag henne i halsremmen och så gick vi ut. Det började snöa och Knaggen stannade i dörren. Hon hade aldrig sett snö förut och blev rädd och nervös. Hon försökte vända och sedan brölade hon till de andra korna. Ja, kom då, min lilla ko! Det här är inget konstigt. Då blev hon genast lugnare. Och så började hon nyfiket lukta på den frusna marken. Vi gick runt på gården. Hon vred ängsligt sina stora öronlurar åt alla håll. Men där fanns inget farligt och hon vande sig långsamt vid snön och kylan. Jag tror till och med att hon började tycka om snön, för hon sträckte ut sin stora tunga och fångade några snöflingor. Hon blev nog förvånad när det bara smakade kallt. Jag undrade om hon frös, men hon hade en mjuk och fin päls, så det var nog ingen risk. Jag hämtade en ryktborste och borstade henne så hon blev glänsande i sin svartvita päls. Den lilla hårtofsen som hon hade mellan hornen kammade jag. Vid rumpan hade hon flagor av torkad gödsel, dem fick jag pilla bort. Jag borstade benen så fint, bort med allt gammalt skräp! Sedan putsade jag klövarna så de blev som blanka, svarta skor. Sist tog jag upp min näsduk och torkade henne runt mulen. Kor är alltid våta, man får torka dem hela tiden. Jag skulle riva mig en mycket större duk från ett gammalt lakan sedan, en som räckte till Knaggen, en riktig konäsduk! Knaggen var den finaste kon som någonsin gått i det snöiga Skåne. Jag gick in till farmor. Hon satt och lyssnade på predikan i radio som vanligt. Vad ska du, hick, göra? frågade hon. Jag ska rida till stan på min ko, sa jag. Du vill inte ha någon smörgås? Nej. Jag kom plötsligt på något fiffigt. Men vänta! Jag tar med mig en mugg och lite chokladpulver. Jag stoppade allt i fickan. Jag kommer nog hem till kvällen. Farmor lyssnade aldrig riktigt, hon hörde mest på prästen. Jag ledde fram min ko till mjölkpallen, klättrade upp och satte mig på hennes rygg. Först spratt det till i benen på henne som om hon ville kasta av mig. Tycker du inte vi ska rida in till stan? sa

jag som om det var den naturligaste sak i världen. Då lugnade hon sig. Fast Knaggen var väldigt hård på ryggen. Jag förstår varför hon fått sitt namn. Hennes ryggrad var alldeles knaggig, det var som att sitta på en trästock med en massa hårda knastar som stack ut. Hon tog sig fram på ett lustigt sätt. Först rusade hon några meter. Sedan stannade hon och bligade oroligt omkring sig. Sedan blev det en ny rusning. Varje steg hon tog var en pina för mig. Jag slängdes fram och tillbaka och var rädd att kastas mot de vassa hornen. När hon stannade såg hon mycket ledsen ut. Hon förstod nog också hur dålig hon var som ridko. Jag satte mig på marken framför henne och funderade. Hennes ögon var så stora. När jag tittade i dem såg jag en spegelbild av mig själv. Hon blinkade tungt. Vi ska klara av det! sa jag. Hur skulle jag göra? För visst kan man väl rida på en ko! Jag kom att tänka på mitt höhus. Och på mjukt hö. Och på en tom säck. Nog skulle man väl kunna sitta mjukare på en vass korygg... Sedan kom jag att tänka på en cykel. Och på ett cykelstyre. Och på de röda och gröna gummihandtagen som sitter på ett cykelstyre. Nog skulle man väl kunna... Jag satt en stund och tänkte på en hösäck och på cykelhandtag. Kons öga speglade hela världen. Jag reste mig upp och gick hämtade en säck i ladan och fyllde den med mjukt hö. Sedan gick jag till garaget. Farbror Gustav använde inte sin cykel på vintern. Jag drog loss gummihandtagen. Ja, de var nog lagom stora. Sedan gick jag tillbaka till Knaggen och visade henne vad jag hade. Hon nosade på handtagen och frustade. Hon förstod ingenting. Hösäcken tyckte hon var mycket bättre. Nej, du ska inte äta höet, sa jag. Inte nu! Jag trädde handtagen på hornen. De passade precis. Hornen blev som ett cykelstyre! Ha, nu behövde jag inte vara rädd för att hornen var för vassa. Jag kunde sitta där och hålla i hornen och styra min ko. Jag lade säcken på hennes rygg. När jag nu hoppade upp, så var det mjukt och skönt. Jag böjde mig framåt och tog tag i handtagen. Jag låg som en speedway-förare på Knaggens rygg. Sedan ska jag skaffa mig en ringklocka, sa jag till henne. Eller en tuta att sätta fast på hornen! Nu styrde jag henne ett par lunkande varv på gården. Ska vi ge oss ut på en riktig runda? ropade jag. Hon sprang i full fart upp mot kyrkan. Hon tyckte faktiskt om att springa! Och jag gungade mjukt fram på min hösäck som på ett moln. En ko kan man inte kommendera. Vill hon inte så bromsar hon och bara tjurar. Men vill hon själv, så går det undan så man knappt kan stoppa henne. Jag böjde hennes huvud framåt och då sprang hon fortare. När vi kom ut på stora vägen, var det som om hon flög fram. Jag kramade styret och fartvinden fick mina ögon att tåras. Åh, vad hon kunde springa, hon som aldrig fått springa förr! Jag frös lite om fingrarna men då stoppade jag bara in händerna i hennes stora varma öron och värmde mig ett ögonblick. Nu red vi ut i världen! Det gick undan och Knaggens svans stod rakt upp som en spik. Vi kom in till stan på bara några minuter. Jag bromsade in henne framför vårt hus. Det var mörkt i fönstren, mamma låg fortfarande kvar på BB och väntade på att min bror skulle födas. Och pappa var på tullen. Där var ingen som kunde se hur fint vi red. Knaggen lunkade ner till stan. Till en början blev hon rädd för alla bilarna, men hon vande sig. Vi svängde in vid BB och jag kikade in i alla fönster men jag kunde inte se mamma någonstans. Folk samlades i fönstren och kikade ut på mig och min ko, men mamma var inte där. I stadsparken stannade vi och vilade en stund. Knaggen letade efter gräs att äta, men gräset låg gömt under snön. Jag gav henne hö från säcken. Ät inte upp allt bara, för då blir det hårt på din rygg! Hon åt eftertänksamt. Det smaskade i hennes mun och det lät som om detta torra hö var mycket saftigt att äta. Jag tog fram muggen och satte mig på huk vid hennes spända juver. Långsamt mjölkade jag en kopp full. Det var kyligt ute och det ångade om den varma mjölken. Så blandade jag i lite chokladpulver. Vips hade jag en kopp varm mjölkchoklad. Den värmde gott. Vi har det bra, du och jag, sa jag till Knaggen. Vi tittade på bilar susade som förbi, spårvagnar som plingade och människor som jäktade fram och tittade förvånat på oss. Vi hade det lugnt på gräset hon och jag. När hon ätit hälften av höet var hon nöjd och det som var kvar räckte till sadel på hemvägen. Vi red hem till farmor igen. Det gick fort nu när Knaggen var mätt och glad. Vi körde om en bil så snön yrde och Knaggen sprang på grusvägen så småstenarna for. Det var som att rida en puma! Jag önskade så att farbror Gustav hade sett oss komma farande i triumf. Men farbror Gustav var nog inne och vilade. Nu gick Knaggen rakt in i stallet. Hon ville mjölkas, det droppade redan ur spenarna. Hon skyndade in i sitt bås. Efter mjölkningen gick farbror Gustav alltid in och drack kaffe. Då gick jag ut i stallet. Min ko skulle sova i höhuset på natten. Jag backade ut henne i stallgången. Hon stannade till vid en kätte där det stod en kalv. De nosade på varandra och kalven råmade ynkligt. Det lät som om man tryckte på ett litet leksaksdjur av plast. Det är ju din kalv! Hon får också

följa med. Kalven ville genast dricka mjölk från sin mamma. Jag lät kalven suga ett litet tag, buffa och stöta i juvret, sedan var den nöjd och följde med ut i snön. Det mörknade och vi gick till det lilla höhuset. Knaggen bör- jade genast äta på väggarna. Det är ett ordentligt hus, sa jag till henne. Du kan äta hur mycket du vill! Höet räcker ända fram till midsommar och då behöver man ju inget hus. Då kan du ligga ute i sommargräset. När jag sett till att de hade allt de behövde, gick jag in till farmor. Var har du varit? frågade hon. Åh, vad du luktar ko, hick. Jag har ju varit i stan med min ko, sa jag. Hon burrade mig i håret. Du hittar på så mycket! Hon gjorde stora smörgåsar till mig och så lyssnade hon vidare på radion. Det var någon sorts kvällsgudstjänst.Sedan gick vi och lade oss som vanligt. Jag låg och tittade i det mörka taket och kunde inte somna. Efter ett tag hörde jag att farmor började snarka. Hum hum, sa jag. Alldeles tyst. HUM HUM, ropade jag en gång till. Nej, hon hade somnat. Ingen brydde sig om mig. Min vita katt Esmeralda låg och sov i fotändan i kväll men ville inte vakna. Jag drog ut fötterna och katten åkte hit och dit, men sov vidare ändå. Jag steg upp. Farmor låg i sin säng som ett träd som fallit tungt. Hon skulle aldrig gå att väcka. Hennes löständer låg och flinade i glaset. Jag kände mig så isande ensam igen och frös om fötterna. Då går jag väl ut till min lilla ko, sa jag för mig själv. Jag drog loss täcket. Katten halkade ner i ett hörn, men öppnade inte ens ögonen. Jag virade in mig i täcket och gick med små steg ut i farstun. Farmors trätofflor stod vid dörren och jag stack fötterna i dem. De var stora och iskalla. Jag tog ficklampan, den som farmor använde om hon behövde gå på dasset mitt i natten. Så hasade jag över gårdsplanen i mitt vita täcke, som ett stort spöke i snön. Stjärnorna syntes på himmeln och det var gnistrande kallt. Jag öppnade dörren av höbalar. Kovärmen slog emot mig. Jag lyste in med lampan. Knaggen och hennes kalv låg bredvid varandra och sov. De tittade upp på mig men blev inte rädda. Kalven reste sig upp och vinglade fram till mig och började suga på min tumme. Jag gick in till dem och stängde med höbalarna efter mig. Det var verkligen varmt därinne! Där skulle man kunna bo hur kall vintern än blev. Knaggen såg fundersamt på mig och tuggade och tuggade. Jag lägger mig bredvid dig, sa jag tyst till henne. Jag låg vid hennes mage och kalven låg på andra sidan och prasslade i höet. Jag släckte lampan. Hum hum, sa jag. Kalven svarade med sitt leksaksljud. Mitt emellan oss låg Knaggen och idisslade. Jag klappade henne på den lilla tofsen mellan hornen. Sedan kände jag på öronen. De var kalla och våta. Båda öronen. Och kalvens också! Vad gör ni? Suger ni öron? Jag tände ficklampan igen. De slickade varandra då och då. Framför allt öronen tyckte de om att slicka! Jag lade mig med huvudet mot kons varma mage. Det bullrade inne i den, det bubblade och pyste inne i magen som i en stor fabrik. Och det var så varmt, som om det fanns en långsamt brinnande brasa därinne. Jag dåsade till. Men varje gång Knaggen klöktes och fick upp en ny tugga att idissla, ryckte jag till och slog upp ögonen. Jag kände mig inte ett dugg ensam på hela natten. När jag vaknade på morgonen, stod Knaggen redan upp. Kalven drack mjölk från henne. Det var så varmt i vårt höhus att jag hade sparkat av mig täcket i sömnen. Jag låg i höet i min ljusblå pyjamas. Jag knuffade ut den ena balen som var dörr. Därute lyste solen redan. Frisk och kylig luft strömmade in genom hålet.Knaggen stod och åt på höhusets vägg. Jag tog på mig farmors trätofflor och gick ut och rullade ihop ett snöklot som jag gav till Knaggen. Hon var törstig och hon åt genast av snön. Så tog jag min mugg och satte mig bredvid kalven. Jag mjölkade i de spenar som var lediga. Du äter vit snö och jag kan mjölka vit mjölk ur dig. Idag ska vi rida ut i världen! Först ska vi galoppera förbi farbror Gustav, så han får se att man kan bli bästa vän med en ko. En ko är inte bara en mun och en rumpa! Nu var muggen full. Jag blandade i chokladpulver och rörde om med fingret. Jag drack en kopp varm mjölkchoklad. Vi har det väl bra, sa jag till kalven. Men kalven kunde inte svara för den hade munnen full av mjölk.