Bilaga 1: Project plan
CrossCheck – a grammar checker for second language writers of Swedish

Professor Viggo Kann
Nada, KTH
viggo@nada.kth.se

2001-09-06

1 Introduction

Language technology has a potential to play a major role in the process of learning a language. Until recently, the use of language technology in systems for language learning has been nearly nonexistent. However, this has not been the case with grammar checkers for second language learners learning English (see e.g. [3, 10, 26, 30]). The question if grammar checkers actually improve second language learners' language is still a question of debate [10, 27]. In spite of this, we see the adaptation of grammar checking for Swedish to second language learners as a first step to put language technology in computer assisted language learning environments.

Designing and developing a grammar checker for second language learners sets new demands on the tools for text analysis. Second language learners are a very heterogeneous group of writers, with different language background, competence and performance. Using exactly the same methods and tools as for a native speaker grammar checking is not a fruitful way to proceed. The error types are too many and too unpredictable [20] and therefore we believe in refinement of current methods and development of new ones.

1.1 Research questions

In this project, we will focus on the detection and diagnosis of the errors. Grammar checking second language learner's text is difficult in many ways. First, how should a text that contains a lot of errors be analysed in general? Second, how should a program detect and diagnose errors in a text difficult even for humans to understand? Finally, what kind of linguistic analysis is needed to support unsure writers?

We are aware of that second language learners need a more comprehensive feedback and instruction to interact with a grammar checker. However, issues about feedback and instruction are out of reach for this project and will instead be investigated in close collaboration with the new Nada project *The use of language tools for writers in the context of learning Swedish as a second language*, see section 2.1.5.

In detail, we will focus on the following research questions:

1. How should ill-formed input in the context of ill-formed constructions be analyzed? Native speakers' errors are often isolated in the sentence. The errors are like occasional islands in a sea of correct text, which makes the error detection in many cases predictable. However, non-native speakers' sentences are in many cases incorrect in several locations within the sentence, and on different levels. Which part of the sentence should the program start analyzing in order to detect and diagnose the main error?

2. How should the program recognize and diagnose grammatical errors in non-native speakers' text production? Many sentences are both syntactic and semantic erroneous, and which error types are most important to get rid of? Moreover, how should this be done? Should all error rules be applied or should a more general grammar checking be done as a first step? For what kind of errors are probabilistic methods in grammar checking needed and better than rule based methods?

3. What kind of linguistic analysis is needed in a grammar checker for second language learners? Writing in a second language is often learning a second language, and the writers' self-confidence and language competence is probably not strong enough to judge the different kinds of proposals from the program. What happens when the program is wrong and causes false alarms? What kind of linguistic analysis is needed to get high precision and still detect errors? Is semantic analysis required in order to get a reliable error detection and useful recall?

1.2 Objectives

We have the following objectives for the project.

- Development of a prototype/demonstrator for grammar checking Swedish as a second language. The tool will be built on the GRANSKA platform.

- Method development: We want to further explore and develop new approaches to grammar checking, focussing on unpredictable errors.

- Extension and improvement of the linguistic analysis in Granska which will be of benefit to standard grammar checking, as well as to other language technology applications.

- Development of an error typology as a base for the grammar checking of Swedish as a second language.

- Development of a corpus of second language Swedish (a Swedish *learner corpus*), both for immediate use in constructing the error typology mentioned above and as a general resource for the second language learning research community. Decide upon standardized storage and annotation formats and develop computational tools for the effective use of the corpus (together with the KTH-corpus project), both independently and also contrastively, in comparison with native Swedish reference corpora such as the KTH-corpus and SUC.

- Development of an annotated news corpus for public domain use. Together with the KTH-corpus project (also applying for funding from the Language technology program) tag and correct the large KTH-corpus. We believe that the KTH-corpus can be used as a native reference corpus for many purposes (e.g. training of probabilistic grammar checking and evaluation of the grammar checker), if due care is exercised.

Every part of the project will be carefully evaluated and documented.

2 Research field

Through natural language technology we can find ways of living comfortably with technology. Our knowledge of language can be used to develop computer systems that help us in composing and correcting text, recognizing speech and writing, understanding text well enough to select information, translating between different languages, and generating speech as well as the printed word.

By applying such technologies we have the ability to extend the current limits of our use of language. Language enabled products will become an essential and integral part of everyday life.

Language technology is composed of two parts: a linguistic part and an engineering part. Both are essential for creating good and practically useful systems. In Sweden there has traditionally been an emphasis on linguistic methods in language technology (except in speech recognition and generation), while the current global trend is towards statistical methods and large data sets. An excellent survey of the state of the art in language technology has been compiled by Ron Cole [11].

Language technology systems are often very language specific. For example, usually an English language technology system cannot be translated into Swedish without quite large modifications and extensions, since English is in many ways a much easier language to handle than Swedish. This means that specific methods and algorithms have to be constructed in order to manage the Swedish language.

Spelling error detection and correction are relatively well studied areas, see for example [25]. Grammar checking and proof-reading are less studied areas, but have been studied for some languages, for example by Vosse [28]. Most approaches have been grammar-based, unlike our approach which is rule-based and statistical.

There are three Swedish grammar checking systems today: Grammatifix (Lingsoft), Scarrie (an EU project, in Sweden mainly developed at Uppsala University) and Granska (developed by the language technology group at Nada and mainly funded by the HSFR/Nutek Language technology program 1997–2000). We have exchanged ideas with both the other projects and now have a quite close cooperation with Lingsoft. None of the existing Swedish grammar checking systems have been adapted for Swedish as a second language.

Computer-supported writing has been a research topic at IPLab since the 80's. The focus has been on how to support the process of composing long texts at the computer, and the development of computer-based research tools for observing and analysing writing processes. Recently, the work has included studies of overview problems, writing research tools, collaborative writing, and language tools for writers.

Learning and teaching Swedish as a second language does not constitute a new area of research. In Sweden many studies have been conducted in the area of acquisition and learning of Swedish as a second language [19, 21]. However they have been conducted from linguistic and socio-cultural perspectives; issues regarding the use of language technology for the development and acquisition of Swedish as a second language have been peripheral in comparison.

This is true also outside Sweden. The use of language checking technology for second language writers and learners is in its infancy. Some work has been done for English as a second language, but it is fair to say that much remains do be done. See, e.g., [12, 13] and the references given there.

The development and use of grammar checking and proof reading tools for Swedish have an important place within the writing process of native speakers. The pedagogical potential has often been neglected [27].

Although the project proposed here will focus on the development of a writing tool for second language writers, the same underlying technology could in principle be used also in foreign and second language instruction. The error spotting methods can be the same, but language learners, having different goals from somebody simply composing a text, need a different form of feedback. Even in this case, the error typology will be helpful in determining the feedback, however.

2.1 Previous and current research at Nada

2.1.1 Algorithms for Swedish language tools

The project was funded by HSFR and Nutek in the Language technology program 1993–1996 and was led by Viggo Kann. In this project we mainly studied and constructed tools for Swedish spelling error detection and correction, and for Swedish hyphenation. The work led to two generally appreciated programs: Stava and Avstava, the first which is available on the web as `http://www.nada.kth.se/stava`.

Stava's word recognition is based on rules for compounding words, suffix rules for inflections and a probabilistic hashing algorithm called Bloom filter [2] for storing the dictionaries. The program also uses word frequencies, misspelling rules and letter 4-grams to give ranked corrections to misspelt words [14]. Stava has evolved for several years and is now probably the best existing Swedish spell checking program with respect to speed and coverage of Swedish words.

When developing Stava we found several other applications for our methods, besides spelling error detection and correction. These applications include correction of optically scanned documents, extension of part-of-speech lexicons, tagging of unknown words, stemming and correction of search questions in information retrieval [22].

2.1.2 Integrated language tools for writing and document handling

During the period 1997–2000 Viggo Kann and Kerstin Severinson Eklundh at Nada led the Granska project on Swedish grammar checking and proof reading. The project was funded mainly by HSFR and Nutek, but also by TFR (as part of the *ramanslag* Algorithms and complexity). In this project we cooperated with professor Gunnel Källgren at the department of Linguistics at

Stockholm University, professor Robin Cooper at the department of Linguistics at Gothenburg University, and Margareta Westman at the Swedish language council (Svenska språknämnden).

The project resulted in several useful tools:

- a very efficient probabilistic part-of-speech tagger and tag disambiguator [8] The performance of our tagger is currently 98% correctness for known words and 93% correctness for unknown words. This is comparable to the best taggers in the world.
- a new object-oriented rule language for describing grammatical errors using rules consisting of regular expressions, words, part-of-speech tags, help rules and recursive rules [24]. Although this was not an original objective of the rule language it turned out to be very useful in detecting phrases such as NP and PP with good precision [23].
- a word inflector that given a word and a tag will inflect the word corresponding to the tag.
- grammar checking rules and help rules [23]. We have constructed and evaluated rules for all error types handled by other grammar checkers for Swedish and also for the very common Swedish error type: split compounds, which is an intrinsically hard error to find by computers.
- GRANSKA, the complete spelling and grammar checking program [15, 7], available on the web as `http://www.nada.kth.se/theory/projects/granska`.
- three user interfaces for Granska (web, Word, stand alone Windows application).

We have put much effort in optimizing the tagging and rule matching using good algorithms and data structures. The rule set is precompiled into a form that makes the rule matching very fast [7].

2.1.3 Interactive assistants

We have started a TBSS-funded project (ending in May 2002) where we will improve the accuracy of the interactive web assistant Relite, developed by Askalot. The system will answer questions from customers on a company and its products. The manager of the system will have to prepare the system by constructing a database of questions and answers in advance. In the project we will improve the system by using language technology tools such as spelling error detection and correction, lemmatization and grammar checking.

2.1.4 Small projects

In some recent smaller projects we have constructed:

- a Swedish word predictor, saving almost 50% of the keystrokes [9].
- a Swedish key word extractor [9].

- a language identifier, that with almost no errors will find out in which of 40 European languages a document is written. Usually one or two sentences is enough to determine the language.
- a Swedish stemmer improving precision in information retrieval [6].
- an automatic stemming rule constructor (under construction).
- an automatic document clustering algorithm, specialized in clustering Swedish news (under construction).

2.1.5 Planned projects

The use of language tools for writers in the context of learning Swedish as a second language
Tessy Cerratto and Kerstin Severinson Eklundh at Nada have recently applied (from Vetenskapsrådet/Utbildningsvetenskap) for funding of this project.

The project aims to investigate issues that are related to the use of computer support for learning Swedish as a second language. In particular, the project deals with the problem of the use of computer-based language tools for writers in the context of learning a second language.

The goal of the project is to study how learners develop their writing practices in the context of learning Swedish as a second language, and to contribute to improving the design of existing language tools for writing in learning contexts.

The work is focused on learning and human-computer interaction issues, but it is closely related to the CrossCheck project and we will cooperate in several areas.

KTH-corpus – A Swedish tagged news corpus for public domain use
Hercules Dalianis, Viggo Kann, Erik Åström, Johan Carlberger, Martin Hassel at Nada apply for funding of this project from the Language technology program. We will work closely with this project in the work on the KTH-corpus.

NEA – A mobile multi-modal multi-lingual news extraction agent
Hercules Dalianis is project leader for this planned project, and he has applied for funding from SSF. The KTH departments DSV (Henrik Boström) and TMH (Rolf Carlson) are also involved in the project. The result will be a news extraction agent adapted for e.g. telephones, SMS, and PDA.

3 Project plan

The research is strongly interdisciplinary between computer science and computational linguistics. At Nada, KTH there is both a strong computational linguistics group (Kerstin Severinson Eklundh, Hercules Dalianis, Ola Knutsson, Martin Hassel) and an algorithmic group (Viggo Kann, Johnny Bigert, Johan Carlberger, Jonas Sjöbergh, Erik Åström) working in the area.

In the Department of Linguistics at Stockholm University, there is a computational linguistics group with a research tradition in the areas of monolingual and multilingual corpus linguistics (Benny Brodda, Torbjörn Lager, Janne Lindberg, Lars Borin), Swedish grammar checking (Rickard Domeij), and computational text linguistics (Sofia Gustafson-Capková and Jennifer Spenader), as well as a well-known researcher in the linguistic subfield of second language acquisition (Björn Hammarberg). Lars Borin also works in the Department of Linguistics at Uppsala University, where some of the learner corpus material will be collected, and where another renowned SLA researcher (Åke Viberg) has agreed to act in the capacity of consultant in matters of learner corpus collection and evaluation.

In the last five years we have built a tool-box of language technology tools around GRANSKA (for tokenizing, tagging, finding phrases, keyword extraction, clustering, language identification etc.) together with our powerful linguistic rule language and some resources such as dictionaries, corpora, etc. All this machinery is just waiting to be used in new problem areas. Therefore the research will naturally build on the GRANSKA platform.

We have since many years collaborated with Svenska språknämnden (Swedish Language Council). In this current project, we will support a master project by Ola Karlsson (working at Svenska språknämnden), on *Using Granska as a support tool for second language learning exercises.*

The project work can be divided into the following parts:

3.1 Constructing a corpus of second language Swedish

The second language corpus, or *learner corpus*, is a relative newcomer to the field of corpus linguistics. It is a corpus of the linguistic production of second or foreign language learners. Like other corpora, learner corpora can comprise written language, spoken language, or both. Also like other corpora, English is by far the best represented language, with at least two large learner corpora, the International Corpus of Learner English – ICLE [18], and the Uppsala Student English Corpus – USE [1]. They have proved to be invaluable sources of empirical data on learner language, useful for both basic research and pedagogical purposes. For learner Swedish, there is some spoken language corpus material, namely the ASU corpus at Stockholm University [19] and Åke Viberg's primary school material at Uppsala University. The ASU corpus, which is freely available for research purposes, has a small component of written material as well, about 50,000 words.

There is a definite need for a large, balanced Swedish learner corpus as a general, publicly available resource. Thus, we see the work on the collection of such a corpus in this project as furthering two aims:

1. the shorter-term aim of supplying relevant text material in sufficient quantities for building the error rules in GRANSKA (see section 3.3);

2. the longer-term aim of building the 'core' of a balanced, extensible Swedish learner corpus, together with the computational tools needed to explore it and relate it to a comparable corpus of native Swedish. The main dimensions of coverage aimed for are *learner background* (native language, educational level, etc.) and *proficiency in Swedish* (beginner, advanced, nativelike fluency, etc.).

The collection of a Swedish learner corpus is more difficult and resource-consuming than the collection of native Swedish material. At the moment, we are aware of three sources for texts written by second language learners of Swedish:

1. the written part of the ASU corpus, which is available in machine-readable form, but not necessarily in the needed format;

2. the SSM corpus, collected by Björn Hammarberg in the 1970's, comprising about 100,000 words of short essays written by adult second language learners of Swedish, representing 10 different native languages. This is a well-balanced corpus which should be included, even though it exists only on paper, in hand-written and typewritten versions. Hence, the typed version must be scanned, OCR-processed, and checked against the hand-written original;

3. essays written by students of Swedish as a second language at educational institutions as a regular part of their courses, and collected by the teachers in computer-readable form (e.g., as MS Word documents). This is how the USE learner English corpus has been collected, resulting in a million word corpus in a bit over a year [1]. We have a preliminary agreement to collect such material in courses of Swedish as a second language offered in the Department of Scandinavian Languages at Uppsala University (Berit Söderman), starting this semester (Fall 2001), and the Department of language and communication at KTH (Cecilia Weissenborn).

   During the first half year of the project we will endeavor to form additional such agreements, e.g. for the corresponding courses at Stockholm University. The learners at these three institutions, being students preparing themselves to enter a Swedish university, correspond fairly well to the intended target group for the shorter-term use of the learner corpus, namely highly educated advanced learners of Swedish as a second language. For the longer-term goal, we must also collect material from other kinds of institutions, in order to get a wider variety of material. This will be initiated in the second year of the project. However, we have had initial contacts with the teacher in charge of the teaching of second language Swedish at the Celciusskolan *gymnasieskola* in Uppsala (Hillevi Torell), who expressed a great interest in contributing material to this project as well as in trying out a prototype grammar checker in class.

In order for it to be useful as a general resource, the corpus should be stored and annotated using a standardized format (e.g. XCES). We will need to make decisions as to the annotation of errors, however. These are matters which should be decided upon together with the end users of the corpus, e.g. SLA researchers, and also be in communication with international language resource standardization initiatives (e.g. EAGLES/ISLE).

Computational tools for manipulating (e.g., searching) and annotating (e.g., part-of-speech tagging) the learner corpus will be developed in collaboration with the group working on the KTH-corpus.

The issue of comparing the learner corpus with a representative native speaker corpus will also be addressed, as will the development of tools for making contrastive learner language – native speaker language investigations on the basis of such a ‘corpus pair’ (or comparable corpus). One good candidate for a suitable native speaker corpus is certainly the Stockholm Umeå Corpus [16], being a balanced corpus. Unfortunately, its size (1 million words) may be insufficient in many cases. However, we believe that the KTH-corpus can be used as a native reference corpus for many purposes, if due care is exercised. In developing a methodology and tools for working with these comparable corpora, we will build on earlier and ongoing work where we have investigated so-called 'translationese' [5] and learner language (English) [4] using parallel and comparable corpora.

3.2 Constructing an annotated news corpus for public domain use

There is a great need for making public a large tagged written corpus in Swedish for the development and evaluation of various human language technology tools specifically for Swedish. Together with the KTH-corpus project (also applying for funding from the Language technology program) we will tag and correct the hopefully 100 million word KTH-corpus.

The main technical work and implementation of the corpus tools will be done in the KTH-corpus project. In this project we will do the linguistic work and then use the result to improve Granska. The idea is to tag the corpus automatically, make it public, and then encourage language technological researchers and companies to use it and report errors using a system that almost automatically will include the corrections in the public and therefore constantly evolving corpus.

It is especially the probabilistic grammar checker that needs a much larger tagged corpus than the one million word SUC corpus. Then the second language corpus can be utilized for extraction of grammar error categories. The parts detected as ungrammatical by the probabilistic grammar checker will be analyzed so that salient features in tags, words, phrases and clauses can be collected. Our aim is to generalize these features so that existing methods and grammar rules can be adapted accordingly.

We plan to detect and correct the spelling errors and grammar errors that we will find in the text so that the corpus also will become a spelling and grammar error corpus, see section 3.4.

3.3 Adapting Granska to second language learners

The group of second language learners is very heterogeneous and the error types will differ a lot between the users. Building rules for all specific error types is an everlasting job; hence, we will explore new approaches to grammar checking. The first goal is therefore to build error detection rules for the most frequent error types in the SSL corpus. These rules together with Granska's original rules will be a major part of the development of a grammar checker for second language learning. The second goal is to develop methods for detection of less frequent error types and unpredictable errors that occur in text produced by second language learners and by native learners text (probabilistic and heuristics methods). One major challenge here is how rule based grammar checking and probabilistic grammar checking should be integrated; exploiting the best use of each method and the combination of the two methods.

3.3.1 Probabilistic grammar checking

Modern grammar checkers, including Granska, are bad at certain grammatical errors, such as finding missing words and misspelled words yielding a semantic error, for example *för* (for) is easily misspelt as *frö* (seed). In a recent article in Svenska Dagbladet the head of the Swedish Language Council reviewed Granska and had the main objection that it could not find this this type of errors.

Our plan is to detect improbable language constructs using trigram frequencies. Due to the limited size of the corpora used, many acceptable trigrams have never been encountered. To mitigate the bad effects thereof, we have used the corpora to build a representative matrix, giving us a probability of tag t being replaced with tag r for all tags t, r. The representatives are then used to improve the trigram frequency checks significantly. Furthermore, we identify and use representatives for NPs and PPs to eliminate the difficult trigrams originating from phrase boundaries.

Sentences containing improbable language constructs could also be sent to a separate set of rules that perhaps can give the user a clue on why the sentence is wrong.

We will also investigate other possibilities of finding and correcting errors without having constructed error rules in advance.

3.3.2 Refinement and development of the linguistic analysis

Second language learners place new demands on the general linguistics analyses in Granska. Granska's general text analysis capabilities will be extended with the following linguistic tools:

- recognition of syntactic functions
- clause type recognition
- extended phrase recognition
- phrase and clause reduction.

3.3.3 Semantics for grammar checking – how to use semantic analysis of words in grammar checking

In many cases, semantic analysis of some kind is necessary to achieve high precision in grammar checking. One example for Swedish is the semantic agreement in the predicative, like *Gröt är gott*, where *Gröt* is non-neuter and the semantically agreeing adjective *gott* is neuter. What kind of semantic analysers do we want and need for grammar checking? To start with, we see three applications for "light" semantics; improvement of grammar checking rules, the probabilistic grammar checker and also as a tool for the learners, both reading and writing, to determiner the sense of a word.

We will start this work by using information from the Lexin lexicon. Lexin is made for second language learners and has at least six language pairs. Lexin provides information about the senses of words and some of this information could be used as is, but also as a starting seed for machine learning of word senses.

We will explore and consider the use of a very large corpus (KTH-Corpus) for bootstrapping methods to build semantic lexicons to be incorporated in Granska [17, 29]). Eventually we will build a word sense tagger) if we found it suitable for our needs. We also want to integrate named entity tagging (from the KTH-Corpus project) into Granska as a first step towards semantic analysis.

3.3.4 New tools for writing

Adapting a grammar checker for Swedish as a second language will result in a new tool for writers with Swedish a second language. However, we also see many possibilities to create new tools for writers in a second language. Among all ideas, we want to start consider and explore two specific applications: Writing memories and linguistic search/editing. A writing memory can work like a translation memory, but instead of giving translations, it will give the user partly matching phrases, clauses and sentences from a very large corpus (KTH-corpus). These matchings will give input to the learners' language generation. The linguistic information in Granska makes it possible to introduce new linguistic functions which are of interest to writers as well as language learners. These include linguistic search, i.e. searching for linguistic units rather than strings of characters. For example, a writer may need to locate all verbs in a text in order to consider the tense choice, and possibly change a verb to present instead of past tense. The latter is an example of linguistic editing functions, which use the linguistic structure of the text to provide powerful tools for revision.

3.4 Evaluation and user studies

In order to automatically assess the rule based and probabilistic grammar checking, we need text annotated for grammar errors. Using this, the software displays statistics such as rule usage, coverage and precision.

Therefore we will develop tools for annotating grammar errors in the KTH-corpus semi-automatically using XML.

In a planned explorative user study we will study second language learners using Granska, as is. The learners will use Granska for a longer period of time and for naturalistic writing tasks. We will use user diaries and interviews as well as direct observations to collect data. The results from the study will lead to directions for further development of the Granska prototype and future work. The texts produced will also be studied from a linguistic point of view, investigating error types and error frequencies. This must be done in order to extend the grammar checking recall of Granska.

A second study in which second language learners using the adapted Granska will be conducted during autumn 2002. This study we will based on the first study, but extended.

In the planned Nada project on second language learners (see section 2.1.5) there will be more user studies complementing the studies of this project.

4 Preliminary results

In a master's project [20] Öhrman investigated how Granska worked on second language learner texts from the ASU corpus [19]. Öhrman reports that Granska finds about 32 % of the errors in ASU with a precision of 85 %. This evaluation is limited but very promising. In addition, Öhrman developed an error typology based on the errors in the ASU corpus. The typology will be a suitable point of departure for the development of an error typology from the SSL-corpus (developed in the current project).

In another master's project, Staerner (forthcoming) studied how Granska can be used in second language learning. One of the main questions addressed by the study was how a grammar checker should be modified and adapted to second language learning. The master's thesis reports findings from interviews conducted with six second language teachers. The interviewed teachers were positive of using computer supported "free writing", based on grammar checking. The teachers all agreed that computer support for language learning is already and will continue to be an important part of education.

5 Relevance and spreading of information

Simple systems for natural language processing are used widely in the society, for example in word processors. Most of the development is done in big programming companies in USA, and they are not thinking about internationalization or localization to Swedish. Therefore it is important for the usefulness and even survival of the Swedish language that Sweden develop Swedish processing tools, and for this reason research in Swedish language technology is essential.

The amount of people learning Swedish as a second language has increased and changed over the last years. Today, more than one million people or one-eight of the Swedish population, are either not born in Sweden or are children of immigrants. Although English represents a bridge between Swedish people and foreigners, it does not always open doors to the Swedish culture and the Swedish society. To master Swedish as a second language is therefore a key to the integration of foreigners to the Swedish society.

We want the results of the research to be used in practice. Therefore we look for partners both in industry and the public sector. The work that we will do in this project should be immediately useful to anyone with Swedish as a second language. This suggests that it might be large interest among ordinary people to read about the systems and to use them.

Stava and Granska have been presented in KTH-nytt, Teknik & Vetenskap and the radio program Vetandets värld. Viggo Kann has given more than 30 talks about Stava and Granska and the ideas behind them for KTH students, high school students and teachers. Since many people use the Stava and Granska proof-reading services on the web, several of them also follow the links and read the web pages about the project.

We plan to spread popular information in similar ways in this project.

We will write about Granska and the project in Swedish popular science journals like Datateknik, Aktuellt forskning och utveckling, Forskning och framsteg and Språkvård.

The language technology group at Nada is teaching an elective last-year course in language technology as part of the Computer science technology program at KTH. The project results will of course be taught in the course.

6 Organization and personnel

Project leader will be Viggo Kann. The leader of the SU part of the project will be Lars Borin. The project groups at Nada and SU will work closely together and also close to the overlapping Nada projects *The use of language tools for writers in the context of learning Swedish as a second language* and *KTH-corpus – A Swedish tagged news corpus for public domain use* mentioned in section 2.1.5.

We will write short progress reports after one and two years of the project, and a final report after the third and last year of the project.

- Prof. Viggo Kann, Nada, 20 % funded by Nada. Project leader, supervisor of Johnny Bigert and Jonas Sjöbergh and assistant supervisor of Ola Knutsson.
- Prof. Kerstin Severinson Eklundh, Nada, 10 % funded by Nada. Supervisor of Ola Knutsson and assistant supervisor of Rickard Domeij.
- Fil. Lic. Ola Knutsson, Nada, 50 %, funded by this project from May 2002. Ph.D. student. Computational linguist specializing in grammar checking.
- Civ. ing. Johnny Bigert, Nada, 60 %, funded by Nada. Ph.D. student. Computer scientist specializing in statistical methods in language technology, especially in grammar checking.
- Civ. ing. Jonas Sjöbergh, Nada, 100 %, funded by this project. Ph.D. student. Computer scientist.
- Lecturer Lars Borin, Department of linguistics SU, 20 %, funded by SU until 2002 and funded by this project 2003–2004.
- project assistent, Department of linguistics SU, 50 % 2001–2002 and 40 % 2003–2004, funded by this project.
- Fil. Mag. Rickard Domeij, Department of linguistics SU, 40 % 2003–2004, funded by this project. Ph.D. student. Computational linguist specializing in grammar checking.
- several diploma works, both by computer linguist students at SU and computer science students at KTH, not funded by this project.

References

[1] M. Westergren Axelsson. USE – the Uppsala Student English corpus: an instrument for needs analysis. *ICAME Journal*, 24:155–157, 2000.

[2] B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM*, 13(7):422–426, 1970.

[3] P. Bolt. An evaluation of grammar-checking programs as self-help learning aids for learners of English as a foreign language. *Computer Assisted Learning*, 5(1–2):49–91, 1992.

[4] L. Borin and K. Prütz. By their fruits ye shall know them: L1 transfer in a learner language corpus. In preparation, 2001.

[5] L. Borin and K. Prütz. Through a glass darkly. Part-of-speech distribution in original and translated text. In Walter Daelemans, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel, editors, *Computational Linguistics in the Netherlands 2000*, pages 30–44. Rodopi, Amsterdam, 2001.

[6] J. Carlberger, H. Dalianis, M. Hassel, and O. Knutsson. Improving precision in information retrieval for Swedish using stemming. In *13th Nordic Conference on Computational Linguistics*, 2001.

[7] J. Carlberger, R. Domeij, V. Kann, and O. Knutsson. A Swedish grammar checker. Submitted to Comp. Linguistics, 2001.

[8] J. Carlberger and V. Kann. Implementing an efficient part-of-speech tagger. *Software–Practice and Experience*, 29(9):815–832, 1999.

[9] J. Carlberger and V. Kann. Some applications of a statistical tagger for Swedish. In *Proc. 5th International Conference on Quantitative Linguistics*, 2000. 51–52.

[10] J. F. Chen. Computer generated error feedback and writing process: A link. *TESL-EJ Teaching English as a second Foreign Language*, 2(3), 1997.

[11] R. A. Cole. Survey of the state of the art in human language technology. `http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html`, 1996.

[12] E. Cornu, N. Kübler, F. Bodmer, F. Grosjean, L. Grosjean, N. Léwy, C. Tschichold, and C. Tschumi. Prototype of a second language writing tool for French speakers writing in English. *Natural Language Engineering*, 2:211–228, 1996.

[13] E. Dagneaux, S. Denness, and S. Granger. Computer-aided error analysis. *System*, 26:163–174, 1998.

[14] R. Domeij, J. Hollman, and V. Kann. Detection of spelling errors in Swedish not using a word list en clair. In *Proc. 2nd International Conference on Quantitative Linguistics*, pages 71–76, 1994.

[15] R. Domeij, O. Knutsson, J. Carlberger, and V. Kann. Granska – an efficient hybrid system for Swedish grammar checking. In *Proc. 12th Nordic Conference on Computational Linguistics*, 1999.

[16] E. Ejerhed and G. Källgren. *Stockholm Umeå Corpus version 1.0, SUC 1.0.* Department of Linguistics, Umeå University, 1997.

[17] W. Gale, K. Church, and D. Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439, 1992.

[18] S. Granger, editor. *Learner English on Computer.* Longman, London, 1998.

[19] B. Hammarberg. *Manual of the ASU Corpus, a Longitudinal Text Corpus of Adult Learner Swedish with a Corresponding part from native Swedes. Version 1999–11–30.* Department of Linguistics, Stockholm University, 1999.

[20] L. Öhrman. Datorstödd språkgranskning och andraspråksinlärare. Technical report, Institutionen för lingvistik, Stockholms Universitet, 2000. D-uppsats i datorlingvistik.

[21] K. Hyltenstam. Svenska som andraspråk – universitetsämne utan forskningsorganisation. In A-B. Andersson, I. Enström, R. Källström, and K. Nauclér, editors, *Svenska som andraspråk och andra språk. Festskrift till Gunnar Tingbjörn. I.* Inst. för svenska språket, Göteborgs universitet, 1997.

[22] V. Kann, R. Domeij, J. Hollman, and M. Tillenius. Implementation aspects and applications of a spelling correction algorithm. In L. Uhlirova, G. Wimmer, G. Altmann, and R. Koehler, editors, *Text as a Linguistic Paradigm: Levels, Constituents, Constructs. Festschrift in honour of Ludek Hrebicek*, volume 60 of *Quantitative Linguistics*, pages 108–123. WVT, Trier, Germany, 2001. Available on WWW from `http://www.nada.kth.se/theory/projects/swedish.html`.

[23] O. Knutsson. *Automatisk språkgranskning av svensk text.* PhD thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, 2001. NADA report TRITA-NA-0105, Licentiat thesis.

[24] O. Knutsson, J. Carlberger, and V. Kann. An object-oriented rule language for high-level text processing. In *13th Nordic Conference on Computational Linguistics*, 2001.

[25] K. Kukich. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439, 1992.

[26] J. C. Park, M. Palmer, and G. Washburn. An English grammar checker as a writing aid for students of English as a second language. In *Proc Conf. on Applied Natural Language Process.*, 1997.

[27] A. Vernon. Computerized grammar checkers 2000: capabilities, limitations, and pedagogical possibilities. *Computers and Composition*, 17:329–349, 2000.

[28] T. Vosse. *Grammar-based spelling error correction in Dutch.* Neslia Paniculata, Enschede, 1994.

[29] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. 33rd Ann. Meeting of the ACL*, pages 189–196, 1995.

[30] M. Yazdani. An artificial intelligence approach to second language learning. *J. Artificial Intelligence in Education*, 1:85–90, 1990.