Bilaga 4: **CrossCheck project status report**

based on input provided by
the CROSSCHECK *project group*
and compiled by Lars Borin
Nada, KTH • Computational Linguistics, SU

2002-10-15

1 Introduction

This is a report on the progress made in the first year of the CROSSCHECK project, a collaboration between NADA/KTH and Linguistics/SU on the development of a grammar checker for second language writers of Swedish, and the collection of a corpus of written learner Swedish.

The CROSSCHECK grammar checker is based on the GRANSKA platform; hence, further development of GRANSKA forms a natural part of the project. We have implemented a server version of GRANSKA. GRANSKA is run continuously on the server, and therefore the scrutinizing tasks can be performed immediately without any tiresome initializing. Grammar checking of a web page now takes less than a second compared to half a minute before. We use PHP for the communication between the web page and the server.

The rest of this report is structured as follows. In the next section, we report on our progress for the items (work packages) specified in the project work plan submitted with the original project application in October 2001. These are repeated here for convenience:

- Collect the SSM [Björn Hammarberg's *Svenska som målspråk* material] part of SSL [Swedish as a Second Language] corpus, and part of the university student material
- Define the annotation system and develop tools for manipulating and annotating the corpora in collaboration with the KTH-corpus group
- Start annotating and complete a preliminary version of the annotated corpora
- Develop the error typology
- Study clause chunking and clause type recognition
- Start working on probabilistic grammar checker and phrase and clause reduction
- Initial user study: second language learners using GRANSKA as is

In separate sections, we report on scientific and other collaborations, on conference etc. presentations, and on popularization. Finally, there is a list of publications and manuscripts produced in the project.

2 Progress in relation to work plan items

2.1 Collect the SSM part of SSL corpus, and part of the university student material

The SSM corpus is a corpus of hand-written essays from adult learners of Swedish written between 1973 and 1975. The corpus comprises around 100,000 words and it consists of entrance exams and final exams for the Swedish course, descriptions of picture sequences, political debate or descriptions of a film, among other things. The mother tongues represented in the SSM corpus are English, Spanish, Greek, Polish, Persian, Finnish, Hungarian, Turkish, Japanese and Arabic.

The conversion of the corpus to computer-readable format comprises scanning the typed versions of the texts, and with the help of the hand-written originals ("originals" often being photo copies), a computer-readable version of the texts has been produced. This work is mostly completed, although a number of issues in connection with the encoding of certain features of the original documents still remain to be resolved, such as, e.g., how to handle cases of bad photo-copy, resolving unclear handwriting, and punctuation (mostly full stops) concealed by the supporting lines below the text.

The first processing of the SSM texts is well on its way to completion, more or less according to the work plan. As already mentioned, the details of the encoding still remain to be finalized, and the texts will need an additional proofreading. Finally, it would also be useful to provide scanned pictures of the hand- written originals for a user of the corpus to verify uncertainties as well as to study phenomena that could not be represented on a computer as for instance certain typographic variants of characters, text layout and so on.

The university student material is being continuously collected as part of the *Svenska 2* courses in the Dept. of Scandinavian Languages, Uppsala University. We have also acquired a subcorpus consisting of 1839 email messages in Swedish exchanged over a period of two years (April 2000 – May 2002) among a group of 15 foreign students in Uppsala.

Finally, Inger Lindberg in the Institute of Swedish as a Second Language, Dept. of Swedish, Göteborg University, has agreed to add a material consisting of tests of Swedish as a Second Language to the corpus. This (hand-written) material is being computerized at Göteborg University.

2.2 Define the annotation system and develop tools for manipulating and annotating the corpora in collaboration with the KTH-corpus group • Start annotating and complete a preliminary version of the annotated corpora

A corpus format to facilitate linguistic searching and automatic evaluation of language technology applications has been proposed. The format separates the annotation of a corpus from the actual text. This makes it possible to have many parallel annotations of the same data, correct errors while leaving the underlying corpus unchanged and to distribute corrections and annotations separately from the original corpus data. A small example corpus has been annotated using this format. See `http://www.nada.kth.se/~johnny/corpus/format.html`.

Work has started on the construction of a tree bank, necessary for evaluation of existing shallow parsers, but also for the gathering of phrase structure statistics.

The existing GRANSKA tagger has been substantially improved. We have developed various utilities (tools) for building lexicons, for comparing the output of different taggers, for extracting a corpus from the electronic version of *Nationalencyklopedin*, and for interactively classifying errors in text.

A preliminary report describing POS tagging of Swedish is being prepared. Several taggers were tested on Swedish text, the best achieving 96.0% tagging accuracy. Combining taggers by voting achieved 96.5% accuracy and combining them by training a new classifier on their results achieved 96.6% accuracy.

2.3 Develop the error typology

An error typology in the domain of prepositional phrases is the subject of ongoing Computational Linguistics Master's thesis work in the CROSSCHECK project. The thesis contains a thorough analysis of L2 errors in prepositions and prepositional phrases extracted from a subset (30 000 words) of the SSM Corpus. The thesis work suggests a template for new error rules, specifically designed for L2 learners, to be implemented in GRANSKA. Furthermore, the work implies an evaluation of the performance of a complete set of rules for the detection and correction of prepositional errors in L2 learner text, as well as a discussion concerning the contributions, shortcomings and further requirements of language technology within the context of second language acquisition.

2.4 Study clause chunking and clause type recognition • Start working on probabilistic grammar checker and phrase and clause reduction

We have continued the work on the statistical methods for context-sensitive spelling error detection.

A robust probabilistic method for the detection of so-called context sensitive spelling errors has been developed. Context sensitive errors, also called "real word errors", are frequent in second language writer's texts, for instance when the noun "kalas" is replaced with the verb "kallas". The program, PROBGRANSKA, identifies less-frequent grammatical constructions and attempts to transform them into more-frequent constructions while retaining similar syntactic structure. If the construction is not susceptible to transformations, it is likely to contain an error.

We have investigated PROBGRANSKA in two experiments. In the first experiment only information derived from a part-of-speech tagged corpus was used. This experiment showed a good error detection capacity but also a high rate of false alarms. ...also a high rate of false alarms. This is often due to phrase and clause boundaries which are likely to produce rare grammatical constructions.

In the second experiment, we have combined the first method with robust phrase and clause recognition to avoid many of the false alarms in the first experiment. A comparative evaluation of the experiments showed that the introduction of linguistic knowledge dramatically increases the precision of the error detection method.

We illustrate the boundary problem with an example:

Den lille mannen som är gammal är sjuk.

The part *är gammal är* appears suspicious if observed in isolation. We identify the NP *Den lille mannen som är gammal* by using the GRANSKA framework as a phrase

recognizer. We transform the NP and get

Mannen är sjuk.

which is a much more common grammatical construct.

2.5 Initial user study: second language learners using GRANSKA as is

A pilot study of three writers has been carried out during six months. The pilot study had two aims; the first aim was to explore and evaluate a naturalistic way to collect data from writers' revisions during free text production. The second aim was to study how GRANSKA should be adapted to second language learners.

In order to adapt GRANSKA to the context of second language learning, we wanted to know about the detection capacity of the program when used by second language writers and more importantly, to study the role of feedback from the program to the users/writers/learners. In particular, we were interested in determining if the users follow the advice from the program. We were also interested in how the users would react to false or misleading responses from Granska.

The prerequisite for the users in the study was the following: "Use GRANSKA whenever you want and when you think it will help you". The control of the data collection was thus left to the users. According to the instructions the user should save the original text scrutinized with GRANSKA and also the final version, written after the revision aided by GRANSKA. Collecting the two versions gave us the opportunity to follow the users' actions.

We also instructed the users to judge the program's detections and feedback using grades. By using a grading procedure we aimed at determining if grades could track some of the revision process and at identifying if feedback was problematic for the user.

The analysis of the two versions of the writer's text seems to be appropriate to collect data about the users' decisions during revision of their texts. Complemented with the users' judgement some of the revision process can be traced. The judgements from the users gave fine-grained information of the different steps using GRANSKA (detection, diagnosis and correction). The judgments also seem to give rise to other comments from the user, which often are richer than just the grades.

False alarms did not entail important problems for the users during the study. According to the comparison between the users' final text versions and the output from GRANSKA, the program did not seem to fool the users. So far, the method used seems to be adequate both in pointing out the directions for an extended study, and also in pointing on which parts of the language tool that must be further developed and improved. As a complement to the study of the writers, interviews with teachers have been conducted. There is an overlap between the errors the teachers said were important and the errors the language tools searches for.

In the near future, diagnose provided by GRANSKA should be rewritten and evaluated together with users. We also plan to involve teachers in the process. Some of the diagnoses given by GRANSKA seem to annoy or/and disorientate the users. We hypothesize that adequate diagnosis and correction proposal would make the users to get useful help from both functions provided by the program. Therefore we think that it is important to concentrate on the study of GRANSKA's capacity for generating adequate

correction proposals.

3 Collaborations

The CROSSCHECK project collaborates (or has collaborated) with the following research projects and individuals:

- *The use of language tools for writers in the context of learning Swedish as a second language*. This project currently runs at IPLab-NADA and is funded by The Swedish Research Council-Vetenskapsrådet. It focuses on human computer interaction issues. In particular, it concentrates on pedagogical aspects of the design of writing software for second language learning.

  The URL of the project is `http://www.nada.kth.se/~knutsson/call.html`.

- *Squirrel – Corpus based language technology for computer-assisted learning of Nordic languages*. This project was funded by the Nordic Council of Ministers April 2001 – March 2002. The main outcome of the project has been a prototype web search engine for locating suitable reading material for second language learners of Nordic languages. Points in common with CROSSCHECK are mainly in the field of determination of general characteristics of texts/reading matter which impact on their use (but also their production) by language learners.

- *PADLR – Personalized Access to Distributed Learning Repositories*. This project is funded by the Knut and Alice Wallenberg Foundation as part of the *Wallenberg Global Learning Network* initiative, October 2001 – September 2003. Issues in common with the CROSSCHECK include standard formats for storage, annotation, and exchange of linguistic resources.

- *Forskningstillgänglighet för ASU-korpusen*. This project is funded by Magnus Bergvalls Stiftelse, collaboration with Björn Hammarberg, Dept. of Linguistics, Stockholm University. January – December 2002. The aim of the project is to provide an easy-to-use interface/research tool for second language acquisition researchers, primarily to the ASU corpus of learner Swedish, but the tool is meant to be general. Here, a close collaboration with CROSSCHECK has been seen as vitally important, mainly for reasons of format compatibility between the corpora in the two projects.

- *IT-based collaborative learning in grammar*. This is a pedagogical project, where one aim is to use a Swedish treebank as a source of 'generative' grammar exercises for linguistics students. Here, too, it is important that storage, annotation, and exchange formats are maximally compatible among the resources developed in this project and CROSSCHECK.

4 Presentations at conferences, etc.

Bigert, Johnny and Ola Knutsson: Robust error detection: a hybrid approach combining unsupervised error detection and linguistic knowledge. Presented at *2nd Workshop on Robust Methods in Analysis of Natural Language – ROMAND'02*, July 2002, Frascati, Italy.

Bigert, Johnny, Ola Knutsson, Viggo Kann and Jonas Sjöbergh: Annotated clauses and flat phrase structures for Swedish. Accepted for presentation at *Swedish Treebank Symposium*, November 2002, Växjö, Sweden.

Borin, Lars: Where will the standards for intelligent computer-assisted language learning come from? Presented at *LREC 2002 pre-conference workshop on international standards of terminology and language resources management*, May 2002, Las Palmas, Spain.

Borin, Lars: What have you done for me lately? The fickle alignment of NLP and CALL. Presented at *EuroCALL 2002 pre-conference workshop on NLP in CALL*, August 2002, Jyväskylä, Finland.

Borin, Lars and Klas Prütz: By their fruits ye shall know them: interference in a learner language corpus. Presented at *EuroCALL 2002*, August 2002, Jyväskylä, Finland.

Domeij, Rickard, Ola Knutsson and Kerstin Severinson Eklundh: Different ways of evaluating a Swedish grammar Checker. Presented at *Third International Conference on Language Resources and Evaluation – LREC 2002*, May 2002, Las Palmas, Spain.

Knutsson, Ola, Teresa Cerratto Pargman and Kerstin Severinson Eklundh: Computer support for second language learners' free text production – initial studies. Presented at *5th International Workshop on Interactive Computer Aided Learning*, September 2002, Villach, Austria.

5 Popularization, etc.

On April 15 2002, we organized a mini conference (temadag) on computer supported grammar checking for second language learners of Swedish. Six talks were given by the researchers in our group and two invited speakers, professor Inger Lindberg from Göteborg University and Olle Josephson, Swedish Language Council. The talks were followed by an open discussion and a possibility to test the current version of the GRANSKA grammar checker in a computer room. Our impression is that the conference was generally appreciated by the more than 50 participants.

The complete program of the conference can be found at `http://www.nada.kth.se/theory/projects/xcheck/temadag.html`.

We have written about the project in the September issue of the journal *Teknik och vetenskap* 2002, and Ola Knutson wrote about our grammar checker in *Språkvård*, the journal of the Swedish Language Council, issue 1 2002. The project was featured in *Nyhetsbrevet Datateknik* ["Grammatik: Svenska som andraspråk svårt för datalingvister" by Torun Bager, issue 6/2002, pp. 2–3].

GRANSKA is regularly used in lectures and practical computer sessions in courses in natural language processing and computational linguistics at KTH and SU, as an example of a concrete computational linguistics application.

6 Project publications, reports, and manuscripts

[1] Alexander Baltatzis. Språkgranskning med reguljära uttryck. Master's thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, 2002.

[2] Johnny Bigert. POS tag distance metrics and unsupervised error detection. Submitted MS, 2002.

[3] Johnny Bigert and Ola Knutsson. Phrase structures in unsupervised error detection. Submitted MS, 2002.

[4] Johnny Bigert and Ola Knutsson. Robust error detection: a hybrid approach combining unsupervised error detection and linguistic knowledge. In *Proceedings of ROMAND'02*, pages 10–19, Frascati, Italy, 2002.

[5] Johnny Bigert, Ola Knutsson, Viggo Kann, and Jonas Sjöbergh. Annotated clauses and flat phrase structures for Swedish. Submitted to Swedish Treebank Symposium, Växjö University, Sweden, 2002.

[6] Lars Borin. What have you done for me lately? the fickle alignment of NLP and CALL. MS; Presentation at EuroCALL 2002 NLP in CALL Workshop, 2002.

[7] Tessy Cerratto and Lars Borin. Overview of the research area (Swedish as a second language). Technical report, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, 2002. NADA technical report TRITA-NA-P0206.

[8] Rickard Domeij, Ola Knutsson, and Kerstin Severinson Eklundh. Different ways of evaluating a Swedish grammar checker. In *Proceedings of LREC 2002*, pages 262–267, Las Palmas, Spain, 2002. ELRA.

[9] Jens Eeg-Olofsson. Feltaxonomi för automatisk språkgranskning av svensk text. Department of Linguistics, Stockholm University, 2002. Bachelor's thesis in Computational Linguistics.

[10] Jens Eeg-Olofsson. Prepositions and automatic proof reading for second language learners. Master's thesis, Computational Linguistics, Department of Linguistics, Stockholm University, Stockholm, 2002.

[11] Magnus Johansson. Hjälpmedel för regelkonstruktion – verktyg för att underlätta skapande av regler till granska. Master's thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, 2002.

[12] Ola Knutsson, Teresa Cerratto Pargman, and Kerstin Severinson Eklundh. Computer support for second language learners' free text production – initial studies. In *Proceedings of ICL 2002*, Villach, Austria, to appear.

[13] Jonas Sjöbergh. Combination of POS taggers for improved accuracy. Report in preparation, 2002.