Different Ways of Evaluating a Swedish Grammar Checker

Rickard Domeij, Ola Knutsson and Kerstin Severinson Eklundh

Department of Numerical Analysis and Computer Science
Royal Institute of Technology
SE- 100 44 Stockholm, Sweden
{domeij, knutsson, kse}@nada.kth.se

Abstract

Three different ways of evaluating a Swedish grammar checker are presented and discussed in this article. The first evaluation concerns measuring the program's detection capacity on five text genres. The measures (precision and recall) are often used in evaluating grammar checkers. However, in order to test and improve the usability of grammar checking software, they need to be complemented with user-oriented methods. Consequently, the second and the third evaluations presented in the article both involve users. The second evaluation focuses on user reactions to grammar error presentations, especially with regard to false alarms and erroneous error identification. The third and last evaluation focuses on problems in supporting users' cognitive revision processes. It also examines user motives behind choosing to correct or not to correct problems highlighted by the program. Advantages and disadvantages of the different evaluation methods are discussed.

1. Introduction

Tools for checking mechanics, grammar and style in writing are widely used as an integrated part of common word processors. Until recently, advanced tools have been lacking for smaller languages, such as Swedish. However, there are now one commercial grammar checker, Grammatifix (Arppe, 2000), and two research prototypes available, Scarrie (Sågvall-Hein, 1998) and Granska (Domeij et al, 2000).

There are many reasons for further research and development of authoring aids. First, the need for such aid has increased, especially when the computer as a writing tool has reached many new and different user groups, for example high school students and second language learners. Secondly, before adapting the grammar checkers to new user groups, there is a need for more sophisticated methods for evaluating the functionality and usability of the programs and their effects on users' ability and practices of revision in writing.

This paper will focus on evaluations made in relation to the development of the Swedish grammar checker Granska. We argue that the evaluation of grammar and style checking must go further than merely measuring the functionality by measures of precision and recall, and thus seriously address the issue of usability. By giving examples of three different studies made during the development of Granska, the advantages of using a broader approach to evaluation are demonstrated.

2. The evaluated system

Granska is a grammar checker for Swedish developed at the Royal Institute of Technology in Sweden. It is together with other language tools integrated in a writing environment supporting different aspects of the writing process. Granska combines probabilistic and rule-based methods to achieve high efficiency and robustness (see also Carlberger & Kann, 1999). Using special error rules, the system can detect a number of Swedish grammar problems and suggest corrections for them that are presented to the user together with instructional information.

The interface of a grammar checker serves several important functions. On a general level, it gives a picture of the program's capabilities and way of working for the user. More specifically, it communicates with the user about the errors encountered, describing these errors as well as giving suggestions for correcting them.

Importantly, the interface is also where the program communicates with the user's writing process. If properly designed, it provides for a transparent and easy switch between the grammar checking and other processes of text composition. Although it constitutes a part of the general process of revision, there is no predefined place in writing to which grammar checking can be confined. This is because writing is a highly complex, recursive and individual activity (Flower & Hayes, 1981). Accordingly, the interface should provide means for invoking the grammar checker interactively at any time, and for going back to writing without delay or inconvenience. We have considered these aspects of the design of the interface in our work on the Granska system.

Granska is presently being adapted for second language learners of Swedish. The evaluations presented in the article have been made during different stages in the development of Granska. The development is still an ongoing process, involving recurrent evaluation of functionality and usability.

3. Related research

In other research areas such as information retrieval and information extraction, evaluation methods have been seriously developed in relation to forums such as TREC, MUC and, for Europe, CLEF. Notably, the grammar checking area is short of empirical evaluative efforts of this kind, although some efforts have been made (see the Eagles report for an overview of different evaluations and evaluation methods).

Earlier studies of grammar and style checking software have involved measuring the program's error detection capacity in terms of precision (i.e. error detection correctness) and recall (i.e. error coverage) (see e.g. Kukich, 1992; Birn, 2000; Richardson & Braden-Harder, 1993). The need of measuring the quality of correction alternatives and instructions has also been recognized (see

e.g. Kohut & Gorman, 1995; TEMAA-report, 1997 pp. 34).

Richardson & Braden-Harder (1993) take different text genres into account and report large differences in error detection rates between for instance texts from professional writers and freshman compositions. They also report that professionals are more forgiving to wrong proposals than students.

Kohut & Gorman (1995) evaluate the effectiveness of several commercial grammar and style packages in the writing of business students. In this study, real errors detected by the program were further classified as correctly identified (incorrect usage accurately classified by the program) or incorrectly identified (incorrect usage misclassified by the program). For the correctly identified errors, the remedial advice was rated by experts as very helpful, helpful or not helpful.

Other studies have investigated the impact of specific software on the quality of produced text (see Kohut & Gorman, 1995 for an overview). The studies have often been conducted in pedagogical settings, comparing improvements in text quality between two groups of students, one group using a grammar checker, the other not. Some studies report positive effects while others report no measurable effects at all. The mixed results may be due to problems in controlling the relevant variables or not using sufficiently sensitive variables.

An advantage with the measurements of recall and precision mentioned above is that they are well defined. On the other hand, the results are hard to interpret. Do users prefer high precision before high recall, or perhaps the other way around? The truth is that we do not know what users prefer before we study them. Therefore, measures of precision and recall can only be a starting point. On top of that, aspects such as user abilities and needs, variability of text genres and user groups, the complexity of error types and error presentations must also be taken into consideration.

Although most of the studies mentioned above in some sense are user-oriented in their approach, none of the studies did study real users during computer-aided revision. To get a deeper understanding of user related issues in grammar checking, we decided to study users in process.

4. Three evaluations

In the following three sections, we will present three different evaluations performed in different stages during the development of the Swedish grammar checker Granska. The first evaluation concerns precision and recall of error rules on five text genres for the Swedish grammar checker Granska. It focuses on the functionality of the system and aims at measuring its error detection capacity for three error types across different genres. This study was made during the error rule implementation phase of the project.

The second and the third evaluations involve users in two different ways. The second evaluation is formative and focuses on user reactions to error presentations, especially with regard to false alarms and erroneous error identification. It relies on observational methods complemented with tape recordings of users thinking aloud. The evaluation was performed during the work with error presentations and correction alternatives.

The third and last evaluation focuses on problems in supporting users' cognitive revision processes. The main research question addressed here is if a grammar and style checker has the capacity to support the user in managing three important steps in the revision process: detection, diagnosis and correction. It also examines user motives behind choosing to correct or not to correct problems highlighted by the program. Revision processes and motives for revising are studied by analyzing think-aloud protocols in depth. This study was carried out early in the design process using an experimental prototype of the grammar checker. The work with coding and analyzing the vast amount of data went on during later phases. The study both served to inform and evaluate design decisions.

After the three evaluations have been presented in closer detail in the following sections, the different methods used will be further discussed.

5. Evaluation 1: A text analysis evaluation

Granska was evaluated on five text genres comprising about 200 000 words (Knutsson, 2001). The detections and diagnoses from Granska on these texts were manually examined. The result indicates differences in the outcome of the grammar checking between text genres. In the following text, recall is defined as 'detected errors/all errors' and precision is defined as 'correct alarms/all alarms'.

Collecting and annotating an evaluation corpus are a demanding task, and one problem is to obtain texts that are under revision. The texts in the material have to varying extent been proofread, which is demonstrated in the evaluation results on the different text genres. The text genres were sport news, international news, public authority text, popular science text and student essays. The evaluation corpus contained 418 syntactic errors.

The largest groups of error types in the evaluation material are the following: disagreement within the noun phrase (17%), split compounds (18%), verb chain errors (21%), missing words (13%) and so called context-sensitive spelling errors (13%). The remaining 18% of the errors belonged to about ten broad error types. Granska tries to cover about 60% of all errors in the material. We are continuously working on expanding the error coverage of Granska, and presently focusing on errors specific for second language learners.

The overall recall for all errors in the five genres is 52% and the precision is 53%. The results from the most frequent error types are presented in table 1.

Error type	Sport news	International news	Public authority	Popular science	Student essays	All texts
Verb chain errors	100/91	100/71	75/86	100/78	100/76	97/83
Split compounds	100/11	-/0	71/42	60/27	40/67	46/39
Disagreement within NPs	88/38	100/11	100/25	100/37	74/72	83/44

Table 1. Recall/precision percentages on five text genres for three frequent error types in the material.

There is a big difference between the results from the different text genres. Granska achieves the best results on verb chain errors (e.g. *Han har spela fiol*/*He has play violin*). Verb chain errors got a recall ranging from 75% in public authority texts to 100% in sport news. This may indicate that these errors are easier to find and correct than for instance split compounds (e.g. *Jag samlar bok märken*/*I'm collecting book marks*).

The results on split compounds need further explanations. Split compounds are very difficult to detect without generating false alarms, and therefore there needs to be quite a few errors in the texts in order to achieve a precision over 50%. Student texts contain more errors than the other texts, which results in a precision of 67% and a recall of 40%. Looking at the same error type in public authority texts gives a precision of 42% and a recall of 71%. Moreover, in international news, Granska only generated false alarms and no detections, which can be explained by the fact that there were no split compounds occurring at all in international news text.

Comparing the results with other evaluations is difficult because of factors such as different languages, text types, the complexity of error types, error frequencies in the texts and more. However, some comparisons might be interesting despite all difficulties. The Critique system for English has also been evaluated (Richardson & Braden-Harder, 1993) on different text genres with lower accuracy on texts from professional writing (about 40%) and much higher on freshman composition (72%). The results from the evaluation of Critique are in line with Granska's results on different text genres. For Swedish, an evaluation made by Birn (2000) has been conducted on newspaper texts, and reports a recall of 35% and a precision of 70%. The system evaluated was the Swedish grammar checker in Microsoft Word. The precision is higher than Granska's overall results, while recall is lower, which may suggest different design choices made during the program development in the intricate trade-off between recall and precision. One notable difference is that Word's grammar checker does not address the complex error type split compounds, which Granska does with some loss of precision as a result.

6. Evaluation 2: A formative study of two grammar checkers

During the development of Granska a formative evaluation was carried out. The evaluation consisted of a small user study involving Granska and a commercial grammar checker (Knutsson, 2001). Five users participated in the study. The users were all experienced writers and had all, to some extent, used grammar checking tools before.

Direct observation was used complemented with tape recordings of users thinking aloud. The tape recordings were used as background information in the study, which focuses on the observations. The user's task was to use the two grammar checkers for checking a text containing errors possible for at least one of the programs to detect. When an alarm from the grammar checker occurred, the users could either accept or reject the alarm. They could also correct the errors themselves if they found it suitable.

The study focused on users' responses to false alarms, wrong diagnoses and multiple suggestions from the programs. These three problems are important to study during the development process of a grammar checker. They all address the problem of the trade-off between recall and precision.

If false alarms really are a problem for the users, we have to increase precision, which also means decreased recall, because of the inverse relation between the two measures. If users found multiple diagnosis and suggestions problematic we have to implement a decision mechanism that presents only one diagnosis and suggestion, with the risk of presenting one erroneous diagnosis and suggestion instead of two or more possible error interpretations. In other words, should the user or the program select among alternative interpretations?

One rather common example of multiple diagnoses and suggestions are split compounds versus disagreement within NPs. Consider for example the sentence *Jag vill ha många vy kort* (eng. *I want many post cards*). It could be interpreted as a split compound *vy kort* (*post card*) or as a number disagreement between *många* (*many*) and *vy* (*post*). In the study, the commercial grammar checker did not present multiple diagnoses but Granska did in form of a list of alternatives presented to the user. At this stage in the development of Granska, we were seeking a metric that could rank and possibly avoid alternative interpretations of an error. Before implementing such a metric, we wanted to know how users reacted to multiple interpretations.

Results suggest that several conflicting diagnoses and proposals seem to be a limited problem for the users if one of the proposals is correct. It only took the users' a minimal amount of extra time to select the correct alternative among several. This gave us valuable information for the further development of Granska. Since there seemed to be limited need for implementing a metric for choosing only one diagnosis and suggestion, our further efforts in the development process were

concentrated on improving the program with regard to false alarms and missed.

Moreover, the results showed that some users seem to need only the detection from a grammar checker, and are able to make the correction in the text by themselves. Surprisingly often, they corrected the text according to the programs' proposals, but instead of inserting them by pressing the buttons in the interface, they typed the correction directly into the text.

False alarms from the programs seem to be of variable difficulty for the users. Easily judged false alarms from the spell checker did not cause users to change the text, but false alarms on more complicated error types sometimes fooled users to change and follow the advice from the two grammar checkers.

7. Evaluation 3: A study of cognitive revision processes in computer-aided editing

In the third evaluation, we wanted to take a closer look at the cognitive processes behind the observed revision behavior. The study is mainly qualitative and focuses on how well human revision processes are supported by writers' aids from a cognitive perspective. Think-aloud methodology is used to track revision processes (such as detection, diagnosis and correction) during computer aided editing. An analysis of the think-aloud protocols reveals how well a grammar checker manages to support these processes; when and why the tool succeeds or fails to support the writer in revising highlighted problems in the text.

The research is influenced by the work of Hayes et al. (1987) in which a detailed psychological model of the revision process is presented and used in studying revision. The revision process is described as being composed of the following three subprocesses: task definition, evaluation and strategy selection. Three stages in the process are pinpointed as problematic, especially for inexperienced writers, i.e. detecting, diagnosing and revising problems in text. In Hill et al (1991) the same theoretical framework and methodology is used to study on-line editing.

The aim of the present study was to examine the usefulness and effect of writers' aids more closely in the light of this framework. It was a further development of a previous study using a similar design but without think-aloud methodology (Domeij, 1998).

In the present study, 11 university students with considerable experience in writing were asked to revise a letter, first using pen and paper, then using computer aids. The letter was originally a negative response from the authorities to a young girl who had asked for permission to marry before the age of sixteen. For the study, the letter had been prepared to contain 37 problems in mechanics, grammar and style, all of which could be analyzed by the computer tool.

Think-aloud methodology was used to track the revision process both during manual and computer-aided editing. The design made it possible to compare the number of changes that subjects made to planted problems with and without computer aid. Most importantly, it made it possible to find explanations to the observed revision behavior by analyzing the think-aloud protocols. Thus, the study combined quantitative and qualitative methods.

The quantitative results showed that, on average, subjects changed 85% of all problems when using the grammar checker, compared to 60% without it. Subjects refrained from changing 15% of all problems although urged to attend to them by the grammar checker. Why did subjects sometimes change further problems when using the grammar checker, and sometimes not? Some interesting answers were found by analyzing the think-aloud protocols.

Subjects made further changes when using the grammar checker because it aided them in a) detecting problems they had missed in the manual revision, b) defining and diagnosing problems that they had problems diagnosing manually, c) correcting problems that they had failed to find corrections for manually, and d) detect, diagnose and correct problems which they did not know before. Negative effects were also observed, as when subjects were fooled to change because of a false alarm. The results also suggest that changes can be less extensive and more surface-oriented when using the grammar checker.

There were two reasons why subjects did sometimes not change when using the grammar checker: a) the reviser wanted to change but failed because of insufficient instructional support from the grammar checker, or because of other kinds of interactional problems such as pressing the wrong button, b) the reviser chose not to change because he or she did not find the response correct or useful in the present situation. The second situation was by far the most commonly observed.

When subjects choose not to change, it was most often in response to problems in style, where some could be seen to disagree heatedly to the advice from the computer. For example, when one of the writers got the suggestion from the program to consider changing "ingå äktenskap" (eng. "enter into marriage") to "gifta sig" (eng. marry) in order to avoid an excessively bureaucratic style, he responded: "No, I don't agree to that because this is kind of a legal text!"

Interestingly, though, the influence of the tool on the number of changes made in style varied greatly between different subjects. While some writers made almost no changes in style, even though they were urged to attend to them by the computer tool, other writers changed many problems in style such as "enter into marriage" both with and without computer support.

Data from the think-aloud protocols suggest that these differences are related to how different writers define the task of revising. Those who made many changes in style were observed to be more reader-oriented than those who refrained from changing. Clearly, writers showed conflicting views about which style is appropriate in a letter from the authorities: a traditional style characterized by high formality and intransparancy, or a less formal reader-oriented style characterized by clarity. This inhomogeneous nature of style even within genres, make style checking problematic.

8. Discussion and future work

It is our hope that the three evaluative studies presented have convincingly shown the advantages of studying users and combining different qualitative and quantitative methods in the evaluation of authoring aids. While the first study contributed to evaluating the

functionality of the error detection capacity, the two other evaluations informed us of how users reacted to different detected problems and their presentations.

The results of the two later studies are interesting mainly in two respects: 1) they use process-tracking methods that shed light on the cognitive processes involved in computer-aided revision, and 2) they pinpoint interactional problems that must be addressed and attended to in designing more useful grammar checking systems. Thus they enabled us to make important design choices based on user data, as what rules to include in the program, what error presentations and instructions to improve on, or how to present different correction alternatives to the user.

The third study was indeed time consuming in its detailed analysis of think-aloud data, but it also produced interesting and general results concerning problems in supporting different cognitive processes in revision. For example, the problem of supporting different users' task definitions involving style decisions was seen to be very complex because of writers' conflicting views of which style to use within the genre. Although style checking is an interesting problem, it needs further research along this line before it can be effectively supported by computers.

We will pursue the cognitive perspective further in the near future as we are adapting the program to writers with Swedish as a second language. Similar studies as those presented here will be performed on users from this group.

Undoubtedly, there are methodological problems associated with using think-aloud data. There is, for example, reason to be careful when generalizing observations made using think-aloud methodology because of the unnatural situation forced upon writers as they are made to speak out their thoughts during the act or writing. However, think-aloud methodology still remains the most effective way of generating data of the thinking processes involved in revision (cf. Hayes & Flower 1983).

When carrying out an evaluation of a grammar checking system, it is very difficult from a methodological perspective to recreate the conditions of an individual writer, using the system as the need arises. Therefore our evaluations have instead been carried out in a partly simulated mode, where writers get a draft text to analyze and correct. This means that some challenging issues of evaluation have not yet been dealt with.

Revision is not necessarily a one-man show. We must not let the cognitive perspective make us forget the socially embedded nature of writing, as the before-mentioned problems in supporting style checking remind us of. In practice, revision, and more generally writing, is performed in a specific social situation, e.g. in a newspaper office or a second language class. Most often, it also involves negotiation and cooperation between people who may contribute to the task in different ways, as for example in newspaper editing and peer reviewing where someone writes a text and others take part in reviewing and revising it.

When designing a grammar checker as an integrated tool in a system for supporting writing, the context and the cooperative practices of revision should be taken into consideration. Evidently, the editor at the newspaper and his colleagues need other support and aid for their work processes, than the second-language student and his peers in their class at high school. Therefore, in future evaluations we are also considering using ethnographical methods of studying work practices in realistic settings.

In developing and evaluating authoring aids there is need for multidisciplinary approaches using several complementary research methods (see Monk & Gilbert 1995 and Smagorinsky 1994 for an overview of theoretical perspectives and research methods used within human computer interaction and writing research respectively). No single evaluation method gives an exhaustive answer to all important research questions. In this paper we have presented three different ways of evaluating a Swedish grammar checker. In doing that we hope to have contributed somewhat to a broader understanding of the problems involved in evaluating authoring aids.

9. References

Arppe, A., 2000. Developing a grammar checker for Swedish. In *Proc. 12th Nordic Conference in Computational Linguistics, Nodalida-99*. Trondheim, pp. 13-27.

Birn, J., 2000. Detecting grammar errors with Lingsoft's Swedish grammar checker. In *Proc. 12th Nordic Conference in Computational Linguistics, Nodalida-99*. Trondheim, pp. 28–40.

Carlberger, J. & Kann, V., 1999. Implementing an Efficient Part-of-Speech Tagger. In: *Software - Practice and Experience*, 29 (9), pp. 815-832.

Domeij, R., 1998. Detecting, diagnosing and correcting low-level problems when editing with and without computer aids. In *TEXT Technology*, vol 8, no. 1, pp. 14-25. Wright State University, Celina, USA.

Domeij, R., Knutsson, O., Carlberger, J. & Kann, V., 2000. Granska – an efficient hybrid system for Swedish grammar checking. I *Proc. 12th Nordic Conference in Computational Linguistics, Nodalida-99*. Trondheim, pp. 49–56.

EAGLES Evaluation of Natural Language Processing Systems report 1996. http://www.ilc.pi.cnr.it/EAGLES96/browse.html. Last updated 090696.

Flower, L. S., & Hayes, J. R., 1981. A cognitive process theory of writing. *College Composition and Communication*, 32, pp. 365-387.

Hayes, J. R., Flower, L., Schriver, K., Stratman, J. & Carey, L., 1987. Cognitive processes in revision. In: S. Rosenberg (Ed.), *Advances in applied psycholinguistics: Vol. 2*. pp. 176–240. New York: Cambridge University Press.

Hayes, J. R. & Flower, L., 1983. Uncovering Cognitive Processes in Writing: An Introduction to Protocol Analysis. In Mosenthal, Tamer & Walmsley (Eds.), *Research on Writing: Principles and Methods*. New York: Longman.

Hill, C. A., Wallace, D. L., and Haas, C., 1991. Revising on-line: Computer technologies and the revising process. Computers and Composition, 9(1), pp. 83-109.

Knutsson, O., 2001. Automatisk språkgranskning av svensk text. Licentiate thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm.

Kohut, G. & Gorman, K., 1995. The effectiveness of leading grammar/style software in analysing business students' writing. *JTBC,* pp. 341-361. July 1995.

Kukich, K., 1992. Techniques for automatically correcting words in text. *ACM Computing surveys*, Vol. 24, No. 4, pp. 377–439.

Monk, A.F. & Gilbert, N., 1995 *Perspectives on HCI: Diverse Approaches*. London: Academic Press.

Richardson, S & Braden-Harder, L., 1993. The Experience of Developing a Large-Scale Natural Language Processing System: Critique. In Jensen, K. Heidorn, G. E. Richardson, S. D. (eds.), *Natural Language Processing: The PLNLP Approach*, pp. 77-89.

Smagorinsky, P. (Ed.)., 1994. *Speaking about Writing: Reflections on Research Methodology*. Thousand Oaks, California: Sage Publications.

Sågvall Hein, A., 1998. A chart-based framework for grammar checking. Initial Studies. I *Proc. 11th Nordic Conference in Computational Linguistics, Nodalida-98*, Copenhagen, pp.68–80.

TEMAA -A Testbed Study of Evaluation Methodologies: Authoring Aids. Final report. 1997. http://cst.dk/projects/temaa/D16/d16exp.html. Last updated: October 1997.