Automatic Grammar Checking for Second Language Learners – the Use of Prepositions

Jens Eeg-Olofsson, Ola Knutsson
Numerical Analysis and Computer Science
Royal Institute of Technology, Sweden
{knutsson}@nada.kth.se

Abstract

This paper presents a partial extension of a tool for automatic grammar checking of Swedish text. The work was carried out within a research project aiming at designing the grammar checker to meet the needs of second-language writers. The paper discusses the construction and implementation of a new set of matching rules for the grammar checking system, written with the purpose of detecting second language learners' writing errors related to the use of prepositions. A minor evaluation indicated high precision for the performance of this set of rules and an F-score of 40 per cent. The paper also discusses the conclusions drawn during the research process for the future treatment of problems with the use of prepositions.

1 A grammar checker for Swedish text

GRANSKA is an automatic grammar checker developed at the Department of Numerical Analysis and Computer Science (Nada) at the Royal Institute of Technology in Stockholm (Domeij et al., 2000; Knutsson, 2001). Other grammar checkers for Swedish have been developed at the universities in Gothenburg (Andersson et al., 1999) and Uppsala (Sågvall Hein, 1998) and at Lingsoft Oy in Finland (Birn, 2000; Arppe, 2000).

Like Lingsofts grammar checker, GRANSKA is a phenomenon-based system. This means that the program is provided with error detection rules which are designed to match expected writing errors in the input text. This also means that the program does not try to make a full syntactic analysis of the text and that it is not expected to react to a syntactically distorted but very unusual sequence like *Arg gula Nicke på var mannen den hatten med* (*Angry the yellow Nicke with was the man the hat with*) whereas agreement violations at a morpho-syntactic level like *Mannen med det gula hatten var arga på Nicke* (*The man with the yellow hat was angry with Nicke* - the Swedish noun phrase and predicative (dis)agreement markers disappear in the English translation) are detected, providing the user with diagnosis and suggestions for the correction of the errors.

The underlying morpho-syntactic analysis is carried out by means of a statistical trigram tagger (Carlberger and Kann, 1999). The text that is to be checked is first analyzed by the tagger and the tagged text is the input to the rule-based grammar checking process.

Since the fall of 2001 the main target for the research on grammar checking at Nada are writers who learn Swedish as a second language (Knutsson et al., 2003b). The aim of the research project named CrossCheck is to develop a useful tool for this particular group of writers. There is a fairly well-defined set of phenomena known to be particularly difficult to second language learners of Swedish, namely noun phrase agreement, word order, tense and the use of prepositions (Knutsson et al., 2003a; Staerner, 2001; Öhrman, 2000; Pitkänen-Koli, 1990). It is the latter category that we have paid attention to in our search for error patterns in second language learners' texts for the subsequent formalization and eventually the construction of a new set of rules for the system.

2 The corpus

Thanks to a reborn interest for second language acquisition and -learning some of the data collected in the seventies have been subject to OCR scanning and electronic storing. This is the case with the SSM (Svenska som målspråk/Swedish as a target language) corpus, a collection of student essays written by adult learners of Swedish as a second language and representing 10 reasonably different mother tongues (Hammarberg, 1977). We analyzed a subset of the SSM corpus, i. e. the first part of it that was electronically stored, containing 140 student essays (30 000 words) representing three mother tongues: English, Arabic and Japanese. A balanced ten per cent of the material was extracted and saved for the evaluation (se section 5 below). The error analysis of the essays provided information as to which phenomena the program should look for. For information on standard text distribution and frequency of certain constructions or phenomena we used the Stockholm-Umeå Corpus - SUC 2.0 (Ejerhed et al., 1992; Källgren, 1998).

3 The empirical analysis

3.1 The grammarians and the preposition

An initial investigation into the very concept of preposition and various descriptions of its function and distribution in Swedish revealed a number of fairly divergent views and conclusions (Eeg-Olofsson, 2002). This pre-study was important to obtain harmony between our own analysis and that of the tagger, but also in order to evaluate the distinction preposition - verb particle inherent in the tagger, who's look-up lexicon provides information from the manually tagged SUC corpus. The pre-study was also important for the conclusions as to what paths to follow in the future revision and extension of the rules.

3.2 The tagger and the preposition

We had the tagger analyze the student essays. Subsequently we examined the prepositions in a randomly extracted subset of the morpho-syntactically tagged text to check the status of the statistical analysis. It was surprisingly consistent. The inconsistences were practically restricted to the distinction preposition - verb particle and they were very few.

3.3 The error typology

When examining the tagged text we found four hundred instances in which the writer had made some kind of mistake in connection with the preposition. Two criteria had to be fulfilled for the mistake to be classed as a prepositional error:

1. There had to be something wrong with the preposition itself (or with the syntactic position normally occupied by the preposition).

2. The error had to be correctable from a human point of view.

Furthermore, mere semantic distortions but also genitive use of the preposition *av* (*of*) were not considered (in the latter case we judged the correction operation too complicated and uncomfortable for the rule construction).

The erroneous instances were eventually categorized as follows:

Word- or phrase form errors

- Form(a): Misspellings
  Example: *fran, fram, from, fråm* for *från* (correct English: *from*)

- Form(b): Ill-formed idiomatic units
  Example: *till en stor grad* for *till stor del, i hög grad* (correct English: *largely*)

- Form(c): Split compounds
  Example: *till baka* for *tillbaka* (correct English: *back*)

Erroneous use of prepositional constructions

- Use(a): Insertion (erroneously invoking preposition)
  Example: *Hon hjälpte till mig* for *Hon hjälpte mig* (English: *She helped to me*)

- Use(b): Deletion (erroneously omitting preposition)
  Example: *Jag väntade tåget* for *Jag väntade på tåget* (English: *I waited the train*)

- Use(c): Substitution (erroneous choice of preposition)
  Example: *De skrattade på mig* for *De skrattade åt mig* (English: *They laughed on me*)

Some of the problems in the word- or phrase form errors are detected by the spell checking module STAVA which is enclosed in the system. Others constitute so called real-word errors and will have to be detected by the context sensitive module (i.e. the grammar checker).

Within the latter category of the typology, the types Use(a) and Use(b) represent more clearly syntactic violations. Subcategory Use(c) is the one that most obviously represents the lexicality and arbitrariness of the use of the preposition. The detection and correction of an incorrectly chosen preposition in most cases requires the matching of an entire (ill-formed) lexicalized phrase.

4 The construction of the rules

We wrote 31 rules for the detection of word- or phrase form errors, and 9, 8 and 5 rules for the detection of errors of categories Use(a), Use(b) and Use(c) respectively.

4.1 The rules, their organization and syntax

The error detecting rules are expressed in a partly object oriented rule language that has been exclusively developed at Nada for the GRANSKA system (Knutsson, 2001). The grammar checking or scrutinizing is executed on the morpho-syntactic level and an additional syntactic analysis, especially of complex noun phrases, is provided if a rule asks for it.

```
exempelregel@kongruensregler
{
X(wordcl=dt),
(JJ/Y)(),
Z(wordcl=nn &
(gender!=X.gender | num!=X.num |
spec!=X.spec) & (gender=Y.gender &
num=Y.num & spec=Y.spec))
-->
mark(all)
corr(X.form(gender:=Z.gender,
     num:=Z.num, spec:=Z.spec))
info ("Om" italics(X.real_text)
       "syftar på" italics(Z.real_text)
        "är det kongruensfel")
action(scrutinizing)
}
```

The rules are designed with a left hand side corresponding to the (erroneous) sequence that is to be matched, in this case a determiner followed by a one or more unproblematic adjectives (the rule invokes a help rule for the adjectival sequence), and a noun representing some kind of violation of gender agreement with the determiner (and not with the adjective), e.g. *mannen med det gula hatten* or *den lilla gula huset på prärien* (*The little yellow house on the prairie* - again, the (dis)agreement is invisible in the translation). The right hand side provides feedback for the user and returns detection (the statement mark), diagnosis (the statement info) and correction (the statement corr). The dot notation is a transparent way of assigning values to the objects. Compulsory in the right hand syntax is the statement action.

4.2 Rules for word- and phrase form errors

The left hand part of these rules typically contains a disjunctive list of (expected) ill-formed varieties of the target token in context.

```
Example: Confusion set rule

Input:
Jag är from Atlanta (I am from Atlanta)

Output:
detection: Jag är [from] Atlanta
diagnosis: Du menar antagligen [från]
correction: Jag är [från] Atlanta
```

```
spell3@formpprules
{
X1(wordcl!=dt),
X2(text="fran" | text="fron" |
   text="from" | text="fråm" | text="fram"),
X3(text!="som" & wordcl!=pp)
-->
mark(X2)
corr(X2.replace("från"))
info("Du menar antagligen" italics("från"))
action(scrutinizing)
}
```

The rule also detects the grammatical sequence *Fransiskus var from och vis* (*Franciscus was pious and wise*) producing a false alarm with the correction *Fransiskus var från och vis* (*Franciscus was from and wise*), but the word *from* is a low frequency word and the probability for this adjective to be detected is very low indeed. The string *fram* on the other hand, was eventually subtracted from the confusion set since it produced too many false alarms, being homo-graphic with the high frequency adverb *fram* (*forward*, *ahead*).

4.3 Rules for the erroneous use of prepositional constructions

Type Use(b), Deletion, proved to be one of the most frequent error types in the corpus samples. Of course these errors were less easily identified as they could not be marked or tagged in the text - the challenge was to find something that was not there!

Jag måste byta ett annat tåg. Jag väntade tåget. När jag väntade tåget såg jag en liten flicka. (*I had to change another train. I waited the train. When I waited the train, I saw a little girl.*)

We will show a slightly simplified version of the *wait-rule*, detecting *Jag väntade tåget* as well as *När jag väntade tåget*.

```
input:
Jag väntade tåget

output:
detection: [väntade tåget]
diagnosis: Du har nog glömt
           preposition före [tåget]
correction: Jag väntade [på] tåget
            Jag väntade [i] tåget
            Jag väntade [med] tåget

delpp5@lexpprules
{
X1(lemma="vänta"),
(NPall/X2)(text!="barn" &
           text!="tillökning"),
X3(wordcl!=ie & wordcl!=dt &
wordcl!=pn & wordcl!=pm & wordcl!=pp)
-->
mark(X1 X2)
corr(X2.insert("med"))
corr(X2.insert("i"))
corr(X2.insert("på"))
info("Du har nog glömt preposition
      före" italics(X2.real_text))
action(scrutinizing)
}
```

The rule invokes a help rule (X2) matching a complex noun phrase, e.g. *Jag väntade mannen med den gula hatten* (*I waited the man with the yellow hat*), and it blocks the matching of the lexicalized expression *vänta barn/tillökning* (the English equivalent being *expecting*). The variable X3 blocks the matching not only of any preposition immediately following the verb but also of a whole series of well formed sequences possibly even with some preposition involved in the construction, such as *Det väntade jag mig inte av dig* (*That, I did not expect from you*).

The user has to choose from three (ordered) alternative suggestions for the correction. After all there is a not very small possibility that the writer actually intended to write *Jag väntade i hallen* (*I waited in the hall*) or *Jag väntade med läxorna till dagen därpå* (*I postponed doing my homework until the next day*). The given list of alternative corrections is similar to the functionality in most spell-checkers.

4.4 The unpredictable writer

It would of course be impossible to predict and implement all the possible deviations from the expected input. A phenomenon based grammar checker is after all not likely to detect all the errors concerning prepositions in the following sentence:

After sex månader comde Jag till båcka att förtset min studing

An English translation of the intended sentence would be *After six months I came back to continue my studies*. The sequence contains one uncomplicated word form error (*After*), one split compound (*till båcka*) and one deletion (*till båcka [ ] att*). The problem here is that the ill-formedness in the context (*comde*, *båcka*, *förtset*) was not predicted by the rule designer and the sequence will not match the rules as the following sequence would:

After sex månader kom jag till baka att fortsätta mina studier

Deviant context can also cause the matching of a rule which was designed for a problem other than the one at hand. The diagnosis turns out to be erroneous and by consequence the correction becomes misleading:

```
input:
Vi [träffade i] Minneapolis
(We saw in Minneapolis)

output:
detection: [träffade i]
diagnosis: Du ska nog inte ha preposition
           efter verbet [träffade]
correction: Vi [träffade] Minneapolis
            (We saw Minneapolis)
```

The rule aims at constructions with *träffa* used intransitively, as in *Jag träffade med min chef* (*I saw with my boss*). But in this case the writer certainly meant to use the reciprocal form of the verb, i.e. *Vi träffades i Minneapolis* (*We saw each other in Minneapolis*). The rule assigns the noun of the actually unproblematic prepositional complement as the syntactic object of the sentence, producing a semantic clash.

User studies conducted at Nada have showed that the users are uncomfortable with the diagnosis provided and that they want to know more more about the problem at hand (Knutsson et al., 2003b). A more exhaustive diagnosis probably would improve the chances for the user to find out the grammar checker provided an erroneous analysis, and maybe subsequently find out herself (or with human assistance) what the problem really was.

5 The evaluation

False alarms can have a strongly negative effect on the writing process, particularly for writers with low self esteem (Domeij et al., 2002; Domeij, 2003). Therefore we have paid particular attention to sharpen the precision of the program during the rule construction, inevitably lowering the recall rate.

The minor evaluation, which was executed on the new set of prepositional rules exclusively, actually resulted in a precision rate of one hundred per cent, but then we must keep in mind that any detected prepositional problem counted as a successful performance of the program, even in cases when the diagnosis was wrong and the correction misleading.

The recall was 25 per cent. More concretely - in the evaluation text of the size of 2 800 words, we found 40 prepositional problems manually that fulfilled the criteria presented above in section 3.4. The program, supplied with our new set of rules, detected 11 of these 40 errors and no other error.

6 The Janus-faced preposition

Mastering the use of prepositions in a second language is generally considered notoriously difficult, due to lexicalization. The choice of preposition is often highly arbitrary and unpredictable.

The traditional and not only phrase structural notion of the preposition as governing a succeeding element thus forming a prepositional complement, very often blurs the semantic interpretation of a sentence. As a matter of fact, the preceding verb or predicate seems to play a very important role in the use of prepositions, and the preposition is frequently more closely attached to the preceding predicate than it is to the succeeding nominal. Consider the following sentence:

August skrev ett brev (August wrote a letter)

This is a trivial well-formed proposition around a transitive verb. Now let us add a preposition forming a prepositional phrase in the usual way:

August [VP skrev [PP på [NP ett brev]]]

This is a syntactically perfectly well-formed proposition (around an intransitive verb) but the straightforward syntactic interpretation makes it semantically odd. In other words, what the speaker wanted to express was not that August wrote on a letter (meaning that he wrote on a piece of paper that happened to be a previously written letter) but that he was actually performing the activity of writing a letter:

August [VP skrev på [NP ett brev]]

If we tie the preposition to the verb like this, the proposition is formed around a slightly more complex and still transitive verbal unit and what's more, the verb obtains progressive aspect!

Now, if we when uttering the sentence put the prosodic stress on the preposition, we tighten the preposition even closer to the verb and we obtain yet a new meaning of the sentence, namely that August signed a letter:

August [VP skrev PÅ [NP ett brev]]

For most grammarians, the latter operation on the prepositional unit actually turns its identity into another grammatical category, that of the particles bound to a group of verbs. This is also the expected analysis of the GRANSKA tagger. But what we wish to show with the example is that the citizenship so to speak, of the preposition, could perfectly well be viewed upon as a scalar phenomenon, where the preposition in the function of verb particle is at one extreme of the scale and that prepositional units like in the prepositional phrases *till exempel* (*for example*) and *ibland* (*sometimes*) are to be found at the other extreme (the latter indeed having obviously been fused into an adverbial entity).

Looking at the preposition from the predicate or the nominal surrounding it, more interesting features are exposed. The nominals to the right tend to be much less connected to a particular preposition than the predicates to the left. The noun phrase *tiden* (*(the) time*) for instance, concords with a remarkable set of prepositions - (*i, ur, under, över, på, med, för, av, till*) *tiden*.

The predicate is distributionally more closely attached to maybe only one particular preposition (e.g. *laugh at*, *allergic to*). In other words, the prepositional constructions seem to be more lexicalized on the left hand side of the scale (in spite of instances such as *till exempel*). The left-bound prepositions also seem to be the most difficult ones for the second language learner:

Jag har tänkt mycket med skrivningen (*I have thought a lot with the test*)

När de kom in i kassa tittar de på henne och skrattar på henne (*When they came into the till they look at her and laugh on her*)

Kvinnor vill inte konkurrera mot män för att bli chefer (*Women do not want to compete against men to become managers*)

Han borjade vissla till hans hund (*He started whistling to his dog*) (correct in English)

Even more illustrative are the deletions, where the obligatory preposition is left out:

Vi hade ingen tid för att prata varandra (*We had no time to talk each other*)

Jag väntade tåget (*I waited the train*)

Då kallade Lasse en taxi (*Then Lasse called a taxi*) (correct in English)

Jag kände att skriva några rader (*I felt writing a couple of lines*)

Han visslar hans hund (*He whistles his dog*)

The predicate very often is not a verb but an adjective like *allergisk (mot)* (*allergic (to)*), a participle or an adverb, although syntactically or functionally a predicative:

Det är svårt till mig att prata om politik här i Sverige (*It is hard to me to talk about politics here in Sweden*)

Jag vet ingenting om jag ska bli intressare för svensk språk (*I know nothing if I will be interester for Swede language*)

Han var framme till Göteborg klockan 14,30 (*He arrived to Gothenburg at 14:30*)

Hon hade ont med pengar (*She was short with cash*)

Sometimes there is really nothing wrong with the prepositional item itself but rather with the verb that has been chosen to go with it. The sentence below does not fulfill the criteria for a prepositional error in our analysis:

Måste en läkare prata om för en patient om denne skulle snart dö, eller skulle det vara hemligt? (*Does a doctor have to talk at a patient if they was soon to die, or would it be kept secret?*)

Sometimes it is very hard to judge if an error should be analyzed as an instance of substitution or as an instance of ill-formed idiomatic unit. And it could very well be that this distinction is indeed superfluous.

Dålig utbildning ligger på grunden för de flesta (*Bad education is on the root of most of them*)

...och politiksystemet låter det för att sitta bättre i makten (*The politics system let it to sit better in power*) (errors transfer poorly to English)

7 Consequences for future rule construction

So what has come out of this work? Well we have come to the rather well-justified but not very original conclusion that the use or misuse of prepositions should be treated as lexical phenomena within the system and that future rule construction should be conducted with frequent ill-formed lexicalized sequences as target for the matching rules, in a fashion similar to the detection of ill-formed idiomatic units. An interesting complement might be statistical grammar checking in the lines of Bigert and Knutsson (2002). The lexical approach requires a further study of prepositional constructions, their frequencies and distributions in standard text and in second language learners' writings.

References

R. Andersson, R. Cooper, and S. Sofkova Hashemi. 1999. Finite state grammar for finding grammatical errors in swedish text: a system for finding ungrammatical noun phrases in swedish text. Technical report, Department of Linguistics, Göteborg University.

A. Arppe. 2000. Developing a grammar checker for Swedish. In T. Nordgård, editor, *Nodalida-99 Proceedings from the 12th Nordiske datalingvistikkdager*, pages 13–27. Department of Linguistics, University of Trondheim.

J. Bigert and O. Knutsson. 2002. Robust error detection: A hybrid approach combining unsupervised error detection and linguistic knowledge. In *Proc. 2nd Workshop Robust Methods in Analysis of Natural language Data (ROMAND'02), Frascati, Italy*, pages 10–19.

J. Birn. 2000. Detecting grammar errors with lingsoft's Swedish grammar checker. In T. Nordgård, editor, *Nodalida '99 Proceedings from the 12th Nordiske datalingvistikkdager*, pages 28–40. Department of Linguistics, University of Trondheim.

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

R. Domeij, O. Knutsson, J. Carlberger, and V. Kann. 2000. Granska – an efficient hybrid system for Swedish grammar checking. In T. Nordgård, editor, *Proc. of 12th Nordic Conference on Computational Linguistics (Nodalida-01)*, pages 28–40. Department of Linguistics, University of Trondheim.

R. Domeij, O. Knutsson, and K. Severinson-Eklundh. 2002. Different ways of evaluating a Swedish grammar checker. In *Proc. 3rd Int. Conf. Language Resources and Evaluation (LREC 2002), Las Palmas, Spain*.

R. Domeij. 2003. *Datorstödd språkgranskning under skrivprocessen*. Ph.D. thesis, Stockhom University, Sweden.

J. Eeg-Olofsson. 2002. Prepositioner och automatisk textgranskning för andraspråksinlärare. Master's thesis, Department of Linguistics, Stockholm University.

E. Ejerhed, G. Källgren, O. Wennstedt, and M. Åström. 1992. *The Linguistic Annotation System of the Stockholm-Umeå Project*. Department of Linguistics, University of Umeå, Sweden.

B. Hammarberg. 1977. *Svenskan i ljuset av invandrares språkfel*. Nysvenska studier 57, Lund.

O. Knutsson, T. Cerratto Pargman, and K. Severinson Eklundh. 2003a. Computer support for second language learners' free text production - initial studies. *The European Journal of Open and Distance Learning (EURODL)*.

O. Knutsson, T. Cerratto Pargman, and K. Severinson Eklundh. 2003b. Transforming grammar checking technology into a learning environment for second language writing. In *Proc. HLT/NAACL 2003 workshop: Building Educational Applications Using NLP*, pages 38–45, Edmonton, Canada.

O. Knutsson. 2001. Automatisk språkgranskning av svensk text (in Swedish). Licentiate Thesis. Royal Institute of Technology, Stockholm, Sweden.

G. Källgren. 1998. Documentation of the Stockholm-Umeå Corpus. Technical report, Department of Linguistics, Stockholm University.

T. Pitkänen-Koli. 1990. Fel i svenska uppsatser gjorda av finska grundskole- och gymnasieelever samt universitetsstuderande. In *Andra symposiet om svenska som andraspråk, Skriptor, Stockholm*.

A. Sågvall Hein. 1998. A chart-based framework for grammar checking. initial studies. In *Proc. of 11th Nordic Conference in Computational Linguistic*, pages 68–80, Copenhagen, Denmark.

A. Staerner. 2001. Datorstödd språkgranskning som ett verktyg för andraspråksinlärning. Master's thesis, Department of Linguistics, Uppsala University.

L. Öhrman. 2000. Datorstödd språkgranskning och andraspråksinlärare. Master's thesis, Department of Linguistics, Stockholm University.