Automatic proofreading for Norwegian: The challenges of lexical and grammatical variation

Koenraad de Smedt, University of Bergen <desmedt@uib.no>
& Victoria Rosén, University of Bergen <victoria@uib.no>

Abstract

In this paper we present some techniques, experiences and results from the SCARRIE project, which has aimed at developing improved proofreading tools for the Scandinavian languages. The focus is on methods which were used for spelling and grammar checking and particularly some novel analyses and treatments dealing with the extensive lexical and grammar variation in Norwegian Bokmål.

The major findings are that (1) since in Bokmål, lexical variants may differ with respect to grammatical features, stylistic replacement at the word level causes a need for grammar checking, and (2) the different systems for gender agreement in Bokmål can be handled in an economical way by a single grammar and lexicon if the features in the lexicon are interpreted dynamically depending on the subnorm or style preferred by the author.

1. Introduction

Among language technology applications, proofreading can be equally challenging as, for instance, machine translation. In a fair number of cases, errors in texts cannot be adequately corrected without understanding the intention of the author in the given context. In practice, however, automatic proofreading systems excel not by their understanding of the text but by their consistency and tirelessness in processing high volumes without becoming 'blind' to relatively simple errors as humans tend to become.

But even with limited expectations, the user may may find a proofreading system unacceptable if the number of false alarms is higher than the number of actual errors spotted, or if many suggestions for correction are inappropriate. It is therefore useful to invest in research aimed at improving the coverage of the system as well as the system's ability to propose corrections that are appropriate in the given context, whether grammatical or stylistic.

The SCARRIE project is a language technology project aimed at building high-quality proofreading tools for the Scandinavian languages (Danish, Swedish and Norwegian). The project was sponsored by the European Commission through the Telematics programme. The project ran from December 1996 through February 1999. The coordinator was WordFinder Software AB (Växjö, Sweden). The other main partners in the project were the HIT-programme at Universitetet i Bergen, Institutionen för lingvistik at Uppsala Universitet, Center for Sprogteknologi (København) and Svenska Dagbladet (Stockholm). Although the project aimed at eventual commercial exploitation, it did involve a great deal of linguistic and computational research.

Proceedings of NODALIDA 1999, pages 206-215

At the end of the project, prototypes and evaluation reports were delivered for these languages. The prototypes correct simple misspellings and mistypings by means of advanced spelling and sound based matching criteria. They also have good coverage in their recognition of new compounds and derivations. Furthermore, they can detect repeated sequences, correct diacritical marks, correct words in the context of idioms and multi-word expressions, correct words based on different styles or norms, and perform limited grammar correction.

We will in the remainder of this paper only report on the Norwegian part of the project. Earlier publications (Rosén & De Smedt 1998, De Smedt & Rosén 1999) have highlighted different aspects of the linguistic and computational methodologies which are at the basis of SCARRIE for Norwegian. In this paper, we concentrate on the problems of proofreading for a language which shows rich variation not only in the lexicon but also in grammar. The specific problems related to grammar correction and style which are discussed below have to our knowledge never before been thoroughly researched with natural language processing methods.

2. Lexical and inflectional variants in Bokmål

Designing a system for automatic proofreading is difficult for any language, but Norwegian Bokmål presents a special challenge. Bokmål allows rich variation in the form of stems as well as inflectional endings. As we will see, this variation has grammatical consequences. First, we observe that many word stems in Bokmål have variants, as shown in for instance (1) and (2).

(1) *melk / mjølk* (milk)

(2) *gress / gras* (grass)

There is also variation in inflection, as exemplified in (3) and (4).

(3) *bok+en / bok+a* (book+DEF)

(4) *arbeid+et / arbeid+a / arbeid+de* (work+ed)

When computing the possible combinations of different stems and endings, we observe that the situation becomes more complex and the number of allowed variants increases, as demonstrated in (5).

(5) *melk+en / melk+a / mjølk+en / mjølk+a* (milk+DEF)

When compounding also enters the picture, word forms can easily have a dozen or more variants. At sentence level it is obvious that even more possible combinations may be found. Consider sentence (6) containing thirteen words; this sentence as a whole has no less than 165,888 possible spellings when all combinations of variants are enumerated.

(6) *De lavtlønte sykehjemsansatte ble helt utmattet og slukket tørsten med den surnete fløtemelken.*
(The low-paid hospital employees became totally exhausted and quenched their thirst with the soured cream milk.)

Not all combinations of variants are equally acceptable in all contexts, because variation is not free, but bound to more or less established subnorms within Bokmål. In other words, for almost all words that have variants, it is the case that the choice between them is not neutral, but depends on the author's style. Although the situation is vastly complex, we have in SCARRIE for Norwegian distinguished between three basic styles: radical, conservative and neutral. The stem *melk*, for instance, is conservative or neutral, whereas *mjølk* is radical; the ending *+en* is conservative or neutral, while *+a* is radical or neutral. Example (6) has only neutral variants; entirely conservative or radical variants of this sentence, as well as a great number of inconsistent combinations, can easily be constructed. As a final remark on basic styles, we mention that SCARRIE for Norwegian also handles a *school book norm (læreboknormalen)* in Bokmål, but this is another, quite complicated story which we will not go into here.

The fact that lexical items are associated with a norm or style value has a number of consequences. First, the user of a proofreading system should be able to state a preferred style. The system should be sensitive to that style so that whenever it makes a suggestion for a correction of a spelling error, it proposes a form that fits with the author's style. Second, we can observe that some forms are rarely or never used because they are infelicitous combinations of different styles, such as *mjølken* in (5), which combines a radical stem with a non-radical ending. Even though such forms may be allowed in Bokmål, they will need to be replaced under all major styles (conservative, neutral and radical) if consistency is to be achieved. Third, variants may have different grammatical features; this final complication is an important theme of this paper.

3. Lexicon

SCARRIE uses full-form lexicons which contain all inflectional forms of words except genitives (which are very regular). In order to restrict the system's suggestions for correction to those word forms that occur in the author's chosen style, it would be possible to construct separate lexicons for each subnorm. However, since there is considerable overlap between subnorms, this would be a wasteful and inflexible solution. Moreover, separate lexicons would not allow straightforward correction of word forms belonging to other styles than the author's stated preference. Therefore, one integrated lexicon was constructed with replacements depending on style. Table 1 presents a simple example, consisting of the lexical entries belonging to the lemma *bok* (book).

Table 1. Lemma for *bok* (without frequency information)

word form	*style code*	*compound codes*	*replacement*	*grammar code*
bok	N	N,sg,indef		N_f_sg_indef
boka	C2	N,sg	boken	N_f_sg_def
boken	C3	N,sg	boka	N_fm_sg_def
bøkene	N	N,pl		N_f_pl_def
bøker	N	N,pl,indef		N_f_pl_indef

The entries for the indefinite singular *bok* (book), plural definite *bøkene* (the books) and plural indefinite *bøker* (books) all have a style code N which means they are normal forms and do not need to be replaced under any styles. The entry for the singular definite *boka* (the book) specifies that under style code C2 (conservative), it should be replaced by *boken*. Conversely, the entry for *boken* specifies that under style code C3 (radical), it should be replaced by *boka*. In other subnorms, both word forms are acceptable and therefore never replaced. For forms with more variants, the coding in the lexicon can be quite complex; for more examples from the lexicon, we refer the reader to Rosén & De Smedt (forthcoming).

We focus now on grammar checking, which obviously relies on grammatical information associated with lexical entries. The last column in Table 1 contains grammar codes that are used by a parser which can for instance detect lack of agreement in the NP, as in (7).

(7) * *Den lille bøkene* (the little+SG+DEF books+PL+DEF)

Before discussing the grammar codes in the lexicon in more detail, the grammar correction mechanism itself will first be sketched.

4. Grammar correction in SCARRIE

Various approaches to grammar correction have been tried out for the various languages covered in the SCARRIE project. The system for Norwegian is based on the CORRie platform, which has a built-in LR parser based on augmented context free grammar (Vosse 1992, 1994). Grammar rules for Norwegian were written for use with this parser. The following kinds of grammatical errors can be automatically corrected by the Norwegian SCARRIE grammar:

1. Lack of gender, number and/or definiteness agreement between (a) determiner, adjective phrase and noun in NP, (b) subject or object and nominal or adjectival complement in S, and (c) noun and postposed possessive in NP.

2. Errors involving (a) the wrong sequence of verb forms in VPs and (b) finite vs. non-finite verb forms.

3. Errors involving case forms for object pronouns in topicalized position and for corresponding subject pronouns in inverted position.

Although native speakers of Norwegian would clearly recognize these kinds of errors, they are not uncommon as results of mistypings and editing routines and are occasionally overlooked by human proofreaders. An example of lack of gender agreement is (8), corrected as (9).

(8) * *Et morsomt gutt ler.* (A(neuter) funny(neuter) boy(masculine) laughs.)

(9) *En morsom gutt ler.*

Grammar correction of Norwegian in SCARRIE is based on the detection and correction of mismatches of grammatical features. *Error weights* attached to phrase structure rules make it possible not only to find such feature mismatches, but also to suggest corrections for them. Each feature on the right hand side of a phrase structure rule may have an error weight associated with it, the default being 1. A weight higher than 1 indicates that the feature 'carries more weight'. An example of such a rule is (10).

(10)
```
NP(Gender Number Definiteness NCase)
-> Det(Gender Number Definiteness:3 [dem quant])
   AP(Gender _ Number Definiteness)
   N(Gender:5 Number Definiteness NCase)
```

Trying to correct a feature mismatch by changing the gender of the noun will now produce a total weight of 5, whereas changing the gender of both the determiner and the noun gives a total of 2. The system chooses the analysis with the lowest error weight, and looks up the word forms *et* and *morsomt* in the lexicon. It will find other word forms in the same lemmas with the feature masculine, and can therefore suggest the correction in (9).

The features in the grammar rules refer to features associated with word forms in the lexicon. However, this coding in the lexicon (cf. the last column in table 1) is not straightforward. The reasons for this will become apparent after a discussion of systematic gender variation in Bokmål.

5. Gender systems

Besides the considerable variation in stems and endings, Bokmål has several systems for gender agreement. We can distinguish between three major gender systems. The most obvious lexical characteristic is that feminine singular nouns sometimes behave like masculine ones, both with respect to endings and agreement. This variation is schematically shown in table 2.

Table 2. Main gender systems in Bokmål

3 gender system	*2.5 gender system*	*2 gender system*
ei lita bok	*ei lita bok	*ei lita bok
*en liten bok	en liten bok	en liten bok
boka mi	boka mi	*boka mi
*boken min	boken min	boken min

The first two rows deal with the indefinite form. Here we see that the indefinite form *bok* occurs in all styles. However, it agrees with feminine determiners and adjectives in one system, while it agrees with masculine determiners and adjectives in the other systems.

The bottom two rows show the definite variants *boka* and *boken,* which are both acceptable in the 2.5-gender system. In the 2-gender system, *boka* is not acceptable, while *boken* is unacceptable in the 3-gender system. We have outlined above how lexical entries with replacements can deal with this variation depending on specified styles. In addition, however, we have to take care of agreement, just like we have to for the indefinite form.

The main question is, how can we achieve this variation of the treatment of gender, which not only seems to require different allowable word forms under different styles, but also different grammatical features for the same entry under different styles? One might think it was necessary to use multiple lexicons, multiple grammars, or both. We will show how in fact a more practical and economical solution was devised, consisting of a flexible interaction between a single lexicon and a single grammar.

This solution requires that lexical entries are coded appropriately to reflect the described variation. Unfortunately, the consequences of this variation were never taken care of by lexicographers before the need for a proper natural language processing treatment manifested itself. In Bokmålsordboken and in NorKompLeks, which the Norwegian SCARRIE lexicon is based on, all feminine words are coded as both *m* and *f.* Unfortunately, this does not differentiate between those nouns that are obligatorily *f* in a 3-gender system (e.g. *bok, jente*), and those that may be either *m* or *f* in such a system (e.g. *art, krokodille, nytte,* etc.). This coding does not allow for correct agreement in a 3-gender system.

6. Grammatical codes for gender

We wil now turn our attention to the way in which the codes in the lexicon (cf. the last column in table 1) are related to the features used in the grammar rules. We have opted to create new codes and add them to the SCARRIE lexicon of fully inflected word forms. Here we differentiate between the two classes mentioned before: only words like *krokodille* are treated as *m* or *f,* which means they have the full inflectional pattern of both genders. All other feminine nouns are treated as only *f* in the lexicon, except for

the form with masculine inflection (e.g. *boken*), which receives a special code *fm*, as shown in the last column of table 1.

However, the codes in the lexicon are not to be taken at face value; they are interpreted by subnorm-dependent *translation tables* that convert them to the feature structures required for grammatical analysis. For example, it could be specified that a code as in (11) is to be translated to the grammatical expression (12) which matches expressions in rules such as (10).

(11) `N_f_sg_indef`

(12) `N(f sg indef nocase)`

The effects of the different gender systems are achieved by using not just one translation table, but different translation tables dependent on the author's chosen style. An overview of the subnorm-dependent translations for the relevant entries of the lemma *bok* is shown in the table 3 (with the feature *nocase* omitted for simplicity).

Table 3. Style dependent translations of grammatical codes

word form	***code in lexicon***	***3 gender system***	***2.5 gender system***	***2 gender system***
bok	N_f_sg_indef	N(f sg indef)	N(m sg indef)	N(m sg indef)
boka	N_f_sg_def	N(f sg def)	N(f sg def)	N(m sg def) *
boken	N_fm_sg_def	N(m sg def) *	N(m sg def)	N(m sg def)

When we use the translation table for the 3 gender system, the code for *bok* in the lexicon gives rise to the value *f* for the gender feature. Using grammar rules like (2), this enforces agreement with a feminine determiner, as it should in this system. In a 2.5 or 2 gender system, the code gives rise to the value *m*. This enforces agreement with a masculine determiner.

Next, consider the entries for *boka* and *boken*. The forms marked with an asterisk are not acceptable in the given systems and will be replaced, as was discussed in an earlier section. The remaining forms are coded such that *boka* agrees with the feminine and *boken* with the masculine determiner.

7. Interaction between agreement checking and replacement

The two mechanisms described above, style dependent replacement in the lexicon and style dependent agreement checking in the grammar, each deal with specific aspects of the described variation. Still, it is not sufficient to specify these mechanisms separately. Rather, these mechanisms must interact in order to correct entire phrases such that not only the resulting word forms are allowed under the given subnorm, but also appropriate agreement results.

Consider the correction of the phrase *boken min* in radical Bokmål, for instance. The phrase is grammatically correct, but the inappropriate use of the word form *boken*

triggers correction. However, simply substituting *boka* for *boken* would result in an agreement error where there there previously was none: **boka min.*

Therefore, *after* a word form has been substituted, the sentence must be checked grammatically. Since substituting one word form for another may result in changes in grammatical features, the *new* features are used in the syntactic analysis. In the example given, this may cause detection, and subsequent correction of the lack of agreement. In this way, substitution of *boka* for *boken* triggers also the substitution of *mi* for *min,* resulting in the final correction to *boka mi.*

8. Parsing and grammatical correction

The usefulness of the approach taken will be shown with the help of a few examples of how sentence (13) is corrected in different styles.

(13) *Heimeleksen din er ferdig.* (Your homework is finished)

This example contains the word *heimeleksen,* which has a radical stem and a conservative ending. It will be corrected in different ways depending on style. A correction in style 2 (conservative Bokmål), as it appears in the output from SCARRIE, is given in (14).

(14) #1#Heimeleksen din er ferdig.
-- 1.Hjemmeleksen

In this correction, the radical form *heimeleksen* is replaced by *hjemmeleksen.* There is no grammatical error in this case. In style 3 (radical Bokmål), however, the same sentence is corrected differently. The word form *heimeleksen* must be replaced with *heimeleksa*, as shown in (15).

(15) #1#Heimeleksen #2#din er ferdig.
-- 1.Heimeleksa 2.di

This correction implies replacing a masculine form by a feminine form. Although the original sentence was grammatically fine, the replacement *heimeleksa* has a gender feature that now is in conflict with that of the determiner. Rules such as (10) detect such mismatches and the correction of *din* to *di* ensues.

A final parsing example (16) is meant to show how insufficient coverage in the grammar, together with massive lexical and structural ambiguity may lead to problems in grammar checking.

(16) *Resultatet er det vi har kalt for fiksering i problemløsning.*
(The result is what we have called fixation in problem solving)

Sentence (16), which is error free, nevertheless receives the suggestions for correction shown in (17).

(17) Resultatet er det vi #3#har #4#kalt for fiksering i problemløsning.
-- 3.har?4.kalte

Parsing this sentence results in no less than 28 trees, none of them error free. The

reading which the parser chooses for correction is one in which *kalt for* is analyzed as the NP *kalt fôr* (called lining). With a better coverage of the grammar, the parser should have chosen an error free analysis.

9. Results and discussion

The overall results of testing SCARRIE for Norwegian were very favorable compared to existing systems, as was reported in more detail in Rosén & De Smedt (forthcoming). Without giving further details on other test results, we mention that grammar checking was tested on a test suite containing 20 different NP agreement errors (of several types, including types discussed above), 12 VP errors and 32 style/subnorm errors. All except 2 style errors received perfect corrections.

However, the system's grammar checking exhibits considerable discrepancy between lab performance, which has shown great potential, and tests on realistic texts, which show poor reliability. The reasons why grammar checking performs poorly on authentic texts are the following:

1. The coverage of the grammar is too limited. The projected time for working on grammar checking was only 4 person months, while the actual time spent on it was less than 3 person months. Any project aiming at developing a truly wide coverage grammar from scratch should be measured in person years rather than months.

2. Lexical and syntactic ambiguity cause a large number of analyses of correct sentences. For sentences with errors, the number of possible analyses becomes even larger. It is very difficult for an automatic system to choose the 'proper' incorrect analysis for correction. We believe that this is a problem not only for our own approach, but for any grammar checking which is insensitive to meaning. We think it will also affect shallow parsing systems. Such systems will, if they are scanning for NPs, always run the risk of wrongly analyzing the kind of pseudo-phrase shown in example (16).

3. The grammar formalism used by the parser is limited, for instance in its treatment of long-distance dependencies. It is difficult to attain wide coverage without at the same time allowing unwanted rule interactions which result in spurious analyses.

10. Summary and conclusion

From a language technology perspective, we analyzed the problems that variation in Bokmål poses for proofreading and found new solutions that dealt with the problems in a systematic and linguistically motivated way. Some parts of the solutions implied adaptations of the underlying CORRie engine which was used for all languages involved in the project, while other parts were achieved by a creative and efficient design of the lexical and grammatical data for Norwegian.

In this paper, we concentrated on correction of NP agreement in Norwegian, for various reasons. First, an error corpus for Norwegian (Rosén & De Smedt 1998) revealed that a number of these errors indeed occurs in writing. Second, the CORRie parser which was

used has good feature-based mechanisms for handling agreement, which is at the core of our treatment of NPs. Finally, agreement is non-trivial in Bokmål due to the interesting variations and therefore its computational processing poses challenging research questions.

We have described two mechanisms which together handle the variation at the lexical and grammatical levels. One mechanism makes use of lexical replacement depending on style. The other mechanism is agreement checking using a robust LR parser and grammar. We have shown that in Bokmål, both mechanisms are necessary: lexical replacement in Bokmål is dependent on subsequent agreement checking, because variant word forms do not necessarily have the same grammatical features.

Of particular importance is the interaction of the grammatical and lexical levels for handling linguistic variation. By using translation tables dependent on style, we obtain a flexible interface between the lexicon and the grammar. In fact, multiple lexicons or multiple grammars are simulated in this way, which is a powerful feature.

Some remarks are to be made on the limitations of the system. First, grammar checking in SCARRIE for Norwegian slows the system down by a factor of ten compared to running a spelling check without using the parser. Second, even though the current grammar checking performs very well on construed examples, it is not reliable on authentic texts. Due to massive lexical and structural ambiguity, sometimes errors are not detected, or, even worse, they are corrected to something unintended. Therefore, realistic grammar checking is legitimately the subject of more in-depth research.

11. References

De Smedt, Koenraad & Rosén, Victoria 1999. Datamaskinell skrivestøtte. In: Birgitta Lindgren (ed.) *Språk i Norden 1999* (pp. 20-32). Oslo: Novus.

Landrø, Marit Ingebjørg & Wangensteen, Boye 1993. *Bokmålsordboka* (2nd ed.). Oslo: Universitetsforlaget.

Rosén, Victoria & De Smedt, Koenraad 1998. SCARRIE: Automatisk korrekturlesning for skandinaviske språk. In: Faarlund, J.T., Mæhlum, B. & Nordgård, T. (eds.) *Mons 7: Utvalde artiklar frå det 7. Møtet Om Norsk Språk i Trondheim 1997* (pp. 197-210). Oslo: Novus.

Rosén, Victoria & De Smedt, Koenraad, forthcoming. *Er korrekturlesningsevnen di god? Resultater fra SCARRIE. *Proceedings of MONS 8, Tromsø, Nov. 18-20, 1999.*

SCARRIE, Norwegian homepage: http://fasting.hf.uib.no/scarrie/

Vosse, Theo 1992. Detecting and correcting morpho-syntactic errors in real texts. In: *Proceedings of the Third Conference on Applied Natural Language Processing, Trento* (pp. 111-118). Association for Computational Linguistics.

Vosse, Theo 1994. *The word connection.* Enschede: Neslia Paniculata.