

# Comma checking in Danish

Daniel Hardt

Copenhagen Business School & Villanova University

## 1. Introduction

This paper describes research in using the Brill tagger (Brill 94,95) to learn to identify incorrect commas in Danish. Trained on a part-of-speech tagged corpus of 600,000 words, the system identifies incorrect commas with a precision of 91% and a recall of 77%. The system was developed by randomly inserting commas in a text, which were tagged as incorrect, while the original commas were tagged as correct. Then the tagger was trained to recognize the contexts in which incorrect commas occur. In what follows, we first describe the corpora and tag sets used in this research, and give background on the Brill Tagger. We then describe the methodology for learning to identify comma errors, and then we examine some of the principles that the system learned to identify comma errors. Finally, test results are presented, and we discuss plans for future research. The method used here is quite general, and could be applied fairly directly to a wide range of grammar checking problems, in Danish or other languages.

## 2. Background

- Corpora and tag sets

This research uses two Danish corpora: the Parole corpus (Parole 1998) and the Bergenholtz corpus (Bergenholtz 1988). The Brill tagger was trained on the manually tagged Parole corpus to recognize Danish part of speech tags. The Danish Parole tag set consists of 151 distinct Tags, containing information such as syntactic category, number, gender, case, tense and so on. As described below, we have used a reduced version of the Danish Parole tag set for the current project.

- Brill tagger

The Brill tagger learns by first tagging raw text with an Initial State Tagger, which tags words with their most frequent tag. The resulting file is termed Dummy, and is compared to a file called Truth, which has been manually tagged, and is thus assumed to be completely correct.<sup>1</sup> The system *Contextual-Rule-Learn* searches for transformations that can be used to make Dummy more closely resemble Truth. The system searches among transformations that instantiate the following templates:

Change tag *a* to tag *b* when:

1. The preceding (following) word is tagged *z*.
2. The word two before (after) is tagged *z*.
3. One of the two preceding (following) words is tagged *z*.
4. One of the three preceding (following) words is tagged *z*.
5. The preceding word is tagged *z* and the following word is tagged *w*.
6. The preceding (following) word is tagged *z* and the word two before (after) is tagged *w*.

Learning proceeds iteratively as follows: *Contextual-Rule-Learn* tries every instantiation of the transformations templates, and finds the transformation that results in the greatest error reduction. (See Fig. 1.) This transformation is output to the Context Rules list, and the transformation is applied to Dummy. The process continues until no transformation results in an improvement above a preset

---

<sup>1</sup> We ignore the lexical rules, which are learned in a separate phase. These are not relevant to the present study. See Brill 94, 95 for details.

threshold. The tagger can then be run with rules that determine part of speech tagging for Danish, based on the Danish Parole Corpus. We term this the Base Tagger.

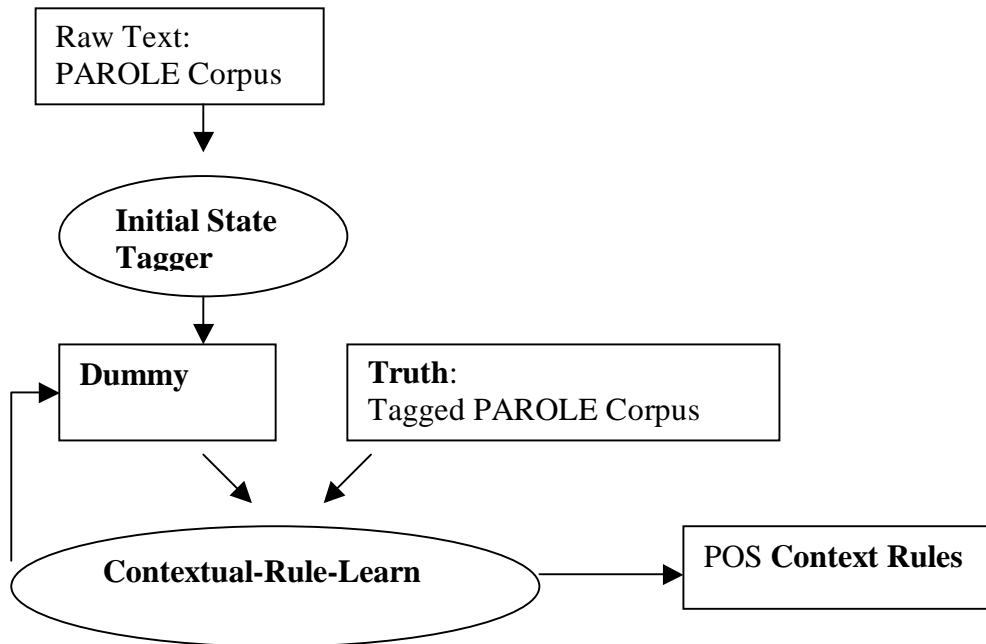


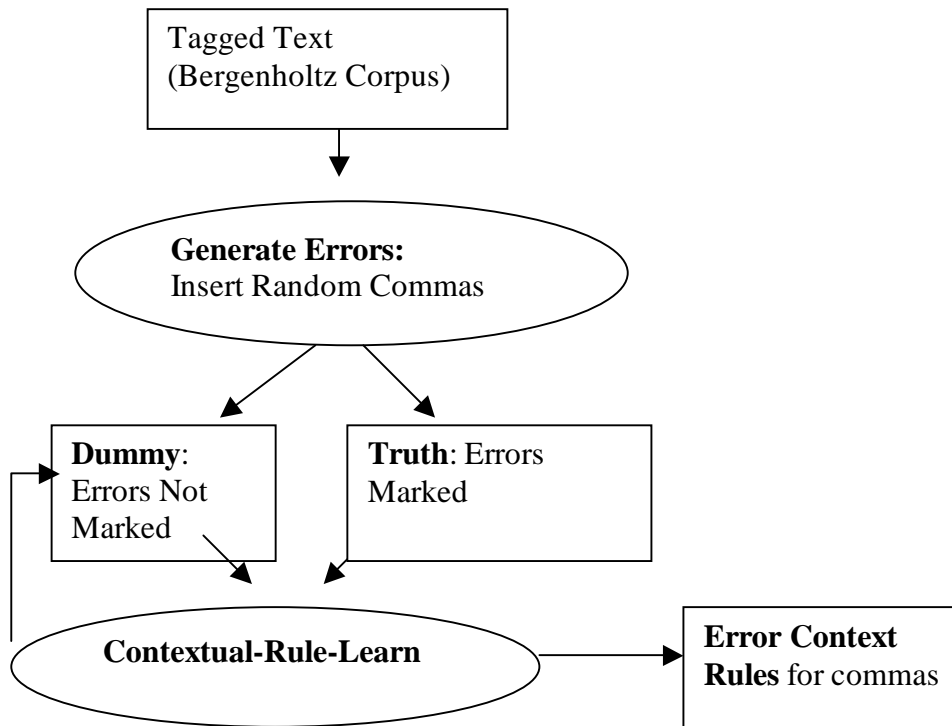
Fig 1. Training the Base Tagger

### 3. Training the comma checker

We produced the training file by tagging 600,000 words of text from the Bergenholtz corpus, using the Base Tagger. We converted the tags to the Reduced Parole Tag Set. This was done to facilitate the learning of generalizations such as “no comma between a preposition and a noun”. In the original tag set, there are 23 tags for common nouns, because of differences in number, gender, etc. In the reduced tag set, there are just two: N (common noun), and N\_GEN (genitive noun). Other categories have similarly reduced numbers of tags.

To use this tagged file as the training corpus for developing a comma checking system, we make the simplifying assumption that all existing commas are correct, and that no additional commas would be correct. Thus all existing commas in the training corpus are given a new tag, *GC* (good comma). Next, two copies of the training corpus are created, *Truth* and *Dummy*. In each of these, commas are inserted at random positions (in the same positions in each file). The inserted commas are labeled *BC* in *Truth*, and *GC* in *Dummy* file. Thus the only differences between the two files are that the randomly inserted commas are tagged with *BC* in *Truth* and *GC* in *Dummy*.

Then, *Contextual-Rule-Learn* is run on these two files. The result is an ordered list of *Error Context Rules* for commas. (See Fig. 2.)



**Fig 2. Learning Comma Context Rules**

Thus what the system learns is contexts in which a comma's tag should be changed from GC to BC, and in this way marked as an error. The list of such contexts is produced by the learner as an ordered list of rules, specifying when the comma tag should be changed. It is important to note that these rules are ordered, so that a decision specified by a rule early on the list will sometimes be reversed by a rule later on the list.

In all, 166 *Error Context Rules* for commas were produced. The first 12 rules are shown below:

1. GC -> BC if one of the three following tags is End-of-sentence
2. GC -> BC if one of the two previous tags is Beginning-of-sentence
3. GC -> BC if the next tag is Preposition
4. GC -> BC if one of the two following tags is Verb(Infinitive)
5. GC -> BC if the previous tag is Conjunction
6. BC -> GC if the previous tag is Interjection
7. GC -> BC if one of the two previous tags is Subordinating Conjunction
8. GC -> BC if the previous tag is Preposition and the following tag is N
9. GC -> BC if the previous tag is Pronoun and the following tag is N
10. GC -> BC if the previous tag is Verb(past) and the following tag is Pronoun(personal)
11. BC -> GC if one of the next two tags is Subordinating Conjunction
12. GC -> BC if the previous word is *er* (is)

The first two rules state that a comma is marked bad ("BC") if it is within 3 words of the end of a sentence, or within 2 words of the beginning of the sentence. These rules were learned because there were comparatively few correct commas in these environments in the Truth file, and a large number of

incorrect commas in these environments. However, the system soon learns that these rules are overly general. For example, the sixth rule states that a comma is correct if preceded by an interjection. This occurs typically near the beginning or end of a sentence, as in the following example from the training corpus:

Naa/INTERJ ./GC I/PRON\_PERS sidder/V\_PRESENT stadig/RGU og/CC hygger/V\_PRESENT  
Well , you sit still and enjoy  
jer/PRON\_PERS ./XP  
yourselves.

Rule 7 doesn't permit commas between prepositions and nouns, and Rule 8 doesn't permit commas near the beginning of a subordinate clause. This is related to the fact that a comma typically introduces a subordinate clause in Danish. This fact is partially captured in Rule 11, which permits commas just before subordinating conjunctions. Rule 9 disallows commas between a Pronoun and Noun. In the Parole corpus, there is no category for Determiner, and words like *the* and *a* are tagged as pronouns.

#### 4. The resulting system

We build a system that corrects commas in raw text, based on the rules learned above. Text is first tagged by the Base Tagger, and then commas are all tagged **GC**. Next the Comma Corrector is executed – this is the tagger with the Comma Error Rules. In the output, any incorrect commas are tagged with **BC**. (See Figure 3.)

Here is a sample run of the system, with different comma positions in the (constructed) sentence *Det er godt, at du kom* (It is good, that you came):

##### Input

Det er godt, at du kom.  
Det er godt at, du kom.  
Det er godt at du, kom.  
Det, er godt at du kom.  
Det er, godt at du kom.

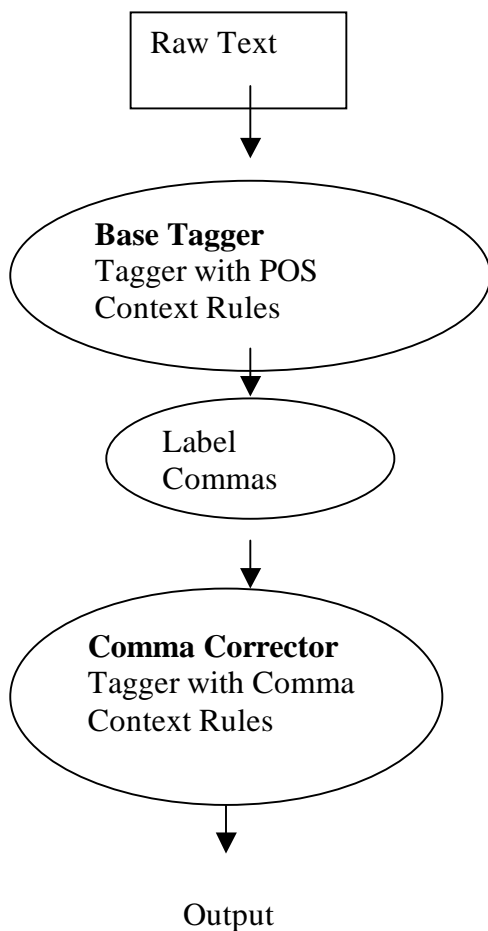
##### Output

Det er godt , at du kom .  
Det er godt at ./BC du kom .  
Det er godt at du ./BC kom .  
Det ./BC er godt at du kom .  
Det er ./BC godt at du kom .

Of the five different comma positions, only the first is correct in Danish.<sup>2</sup> The system correctly labels all the other alternatives as incorrect (BC).

---

<sup>2</sup> This is in fact not entirely clear, since there are at least two distinct systems for placing commas in Danish, and this position may not be considered correct in one of the two systems. While it is difficult to get confident judgments, all my Danish informants agree that this example has only one possible comma position, which is the one accepted by the comma correction system.



**Fig. 3 Comma Correction System**

## 5. Empirical results

The system was tested with a file of distinct text from the Bergenholtz corpus, containing 14,044 words. The file contains 869 commas. 389 additional commas were introduced in random positions, as errors. The system marked 327 commas as errors, of which 299 actually were errors. This gives a precision of 91.4% and a recall of 76.9%.

Here is a list of the first 10 examples where the system incorrectly marked a comma as an error:

1. Hulgaard/EGEN ,/BC Århus/EGEN
2. mener/VPRES ,/BC vi/PRONPERS
3. mener/VPRES ,/BC han/PRONPERS
4. menneskemassen/VPRES ,/BC der/UNIK
5. 17-13/NUM ,/BC Norris-Paulsen/N
6. morderiske/VPRES ,/BC psykopatiske/VINF
7. Sørensen/EGEN ,/BC Århus/EGEN
8. nabokommunen/N ,/BC på/SP
9. systemet/N ,/BC kan/VPRES
10. de/PRONDEMO aktive/ADJ ,/BC servicefunktionerne/N

In items 1 and 7 a line break was incorrectly placed immediately before the text in question. Items 4 and 6 involve mis-tagging:

*menneskemassen* (“mass of people”) and *morderiske* (“murderous”) are both nouns, mis-tagged as verbs. Item 10 is an interesting case *De aktive, servicefunktionerne* (the active, service workers). The comma is marked as incorrect because of the following rule:

GC -> BC if the previous tag is ADJ and the next tag is N

This is normally correct; commas don’t tend to appear between an ADJ and a N. Here, however, “the active” is a complete NP, on par with, e.g., “the rich”, and “service workers” is a separate NP.

Total Number Commas	Incorrect Commas	Total System Corrections	Valid System Corrections	Precision	Recall
1258	389	327	299	91.4% (299 / 327)	76.9% (299 / 389)

**Table 1. Results**

## 6. Discussion and further work

The system was developed using the transformation-based learning system of the Brill tagger. This learning system is limited in various ways: for example, only three words or tags before or after a position are examined. It is likely that certain patterns involving commas could be learned if that locality restriction were loosened. Furthermore, the learning system of the Brill tagger maximizes overall success rate, using a greedy strategy. We believe precision is a more relevant measure in grammar checking problems. Thus it would be interesting to modify the learner so that it optimizes precision or some related measure, and we suspect that greedy learning may be problematic in this case. We are contemplating various experiments related to these issues. It is also possible that the precision and recall of the system would be substantially increased with a larger training corpus. Work is proceeding on this. Finally, we plan to apply similar techniques to a wide variety of grammar problems, both in Danish and other languages.

## References

- Bergenholtz, H 1988 Et korpus med dansk almensprog. *Hermes*.
- Brill, E 1994 Some Advances in rule-based part of speech tagging. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA.
- Brill, E 1995 Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. *Computational Linguistics* 21(4).
- Golding, Andrew R and Schabes, Yves 1996 Combining trigram-based and feature-based methods for context-sensitive spelling correction. In *Proceedings of the 34th Annual meeting of the Association for Computational Linguistics*.
- Jacobsen, Henrik Glaberg and Jørgensen, Peter Stray. 1991. *Politikens Håndbog i Nudansk*. Politikens Forlag.
- Parole. 1998. [http://coco.ihu.ku.dk/~parole/par\\_eng.htm](http://coco.ihu.ku.dk/~parole/par_eng.htm)