In *Second International Conference on Language Resources and Evaluation. Proceedings, Volume I*, Athens, Greece, 31 May–2 June, 2000. 21–26.

Something Borrowed, Something Blue: Rule-Based Combination of POS Taggers

Lars Borin

Department of Linguistics, Uppsala University
Box 527, SE–751 20 Uppsala, SWEDEN
Lars.Borin@ling.uu.se

Abstract

Linguistically annotated text resources are still scarce for many languages and for many text types, mainly because their creation represents a major investment of work and time. For this reason, it is worthwhile to investigate ways of reusing existing resources in novel ways. In this paper, we investigate how off-the-shelf part of speech (POS) taggers can be combined to better cope with text material of a type on which they were not trained, and for which there are no readily available training corpora. We indicate—using freely available taggers for German (although the method we describe is not language-dependent)—how such taggers can be combined by using linguistically motivated rules so that the tagging accuracy of the combination exceeds that of the best of the individual taggers.

1. Introduction

1.1. The problem

Linguistically annotated text resources are still scarce for many languages and for many text types, mainly because their creation represents a major investment of work and time. For this reason, it is worthwhile to investigate ways of reusing existing resources in novel ways. In this paper, we investigate how off-the-shelf part of speech (POS) taggers can be combined to better cope with text material on which they have not been trained, and for which there are no readily available training corpora, i.e. consisting of (automatically or manually) POS tagged and manually checked text material.

The wider context of this research is the ETAP project, a multilingual parallel translation corpus project funded by the Bank of Sweden Tercentenary Foundation. The aim of the project is to create an annotated and aligned multilingual translation corpus which will be used as the basis for the development of methods and tools for the automatic extraction of translation equivalents for applications such as machine translation systems, but it will also be useful as a source of data for contrastive linguistic research of a more traditional cut.

The ETAP corpus consists to a considerable extent of text material of the kind referred to above, either because the text type is special (e.g. truck maintenance manuals), or because of the language (e.g. Polish and Serbian–Bosnian–Croatian), or both. Furthermore, in the various ETAP subcorpora, at most one of the languages represents originally produced native language text, while the others are translations, and it is well-known that translated texts differ in many respects from original texts (cf. Gellerstam, 1996; Johansson, to appear).[1] Not much is known about how tagger performance changes when moving from the training domain domain or text type to other domains or text types, but it is probably safe to assume that, on the average, it will not improve, but rather the opposite (see Padró and Màrquez, 1998).

Since it is not within the brief of the ETAP project to create new training corpora in the traditional way, an extremely work-intensive and time-consuming process, while tagging as much of the text material as possible *is*, we are investigating to what extent it is possible to reuse existing—meaning either developed in our department in some other context, or freely available on the WWW—NLP resources for the task of tagging the languages of the project. As a general rule, we may say that the amount of such resources is growing quite fast at the present moment. On the other hand, their availability is highly dependent on the language, from almost unlimited numbers for English, over a few different POS taggers for German or Swedish,[2] to practically nothing for a language like Polish.[3]

Even in the cases where more than one tagger is available, their performance on our corpus may be fairly uneven, since they represent different tagger technologies, come with lexicons and tagsets of different coverage and size, and have been trained on different types and amounts of text.

However, the differences between existing taggers can actually be used to our advantage, provided that these differences are *complementary* and *systematic*. We want complementarity, because there would be no point in combining part of speech taggers which make the same errors, or where the errors of one tagger is a proper subset of those of the other one, and systematicity (rather than randomness), because if the differences are systematic, we can presumably exploit them to enhance tagging performance.

1.2. A solution: Tagger combination

In the machine learning community, the idea of combining classifiers—e.g., neural networks trained on the same classification task—for enhancing accuracy is an old one, going back at least to the mid-sixties (Tumer and Ghosh, 1999). Several regimes for classifier combination have

[1] More detailed information on the ETAP corpora is available elsewhere (Borin, to appear a; Borin, to appear b).

[2] In addition, the tagged corpora which are used to train POS taggers are still relatively few in number, so that taggers for any language (with the possible exception of English) tend to be trained on the same corpora.

[3] This is the difference between *high-density* and *low-density* languages, according to recent NLP terminology.

been proposed, from simple averaging, over majority voting and more complicated non-linear models, to training a new classifier on the basis of the combination. All these methods have in common that they are *knowledge-poor*, i.e. they require no domain knowledge for their implementation. With other such methods, they share the need for relatively large amounts of training data,[4] and the feature of being supervised methods, i.e. the 'right' answer must be part of the training data.

POS taggers are classifiers in this sense, and it is natural to see how the methods developed for general machine learning could be applied for this specific machine learning task as well. The experiments reported in the literature (Brill and Wu, 1998; van Halteren et al., 1998; Màrquez et al., 1998) have all adhered to the knowledge-poor, supervised training regime, and to my knowledge, the experiment reported in the present paper represents the first attempt to apply a knowledge-rich method to the problem of combining POS taggers, by formulating linguistically motivated rules for how tagger differences should be utilized in the combination of taggers.

In the remainder of this paper, I describe an experiment that we made on knowledge-rich POS tagger combination with freely available German taggers, report on the results of the experiment, describe ongoing work with other taggers and languages, as well as point to further work which needs to be done in this area.

2. An experiment with knowledge-rich POS tagger combination

2.1. Step 1: Finding taggers

The first step in the tagger comparison procedure was the procurement of taggers to compare. Here, the comparison of German taggers is described, but the method used is quite independent of language.[5] For German, we found three publicly available POS taggers, *Morphy* (Lezius et al., 1998), *QTAG* (Mason, 1997), and *TreeTagger* (Schmid, 1994; Schiller et al., 1995).

2.2. Step 2: Evaluating the taggers

The evaluation of the taggers was carried out according to the following procedure. Two short texts from two ETAP subcorpora were tagged with each of the taggers. Ten sentences were then picked out and the number of correct and incorrect tags in them counted.

Of the three German taggers evaluated, one, QTAG, turned out to yield an accuracy below 90% (on both text types), which we had set as the minimum for inclusion in the experiment. This was probably due to it having been trained on nineteenth century fiction (Oliver Mason, p.c.), while our texts are contemporary non-fiction (as well as translations, which also certainly has a bearing on the matter; cf. above).

In table 1, the performance of the other two German taggers is shown for the two text types, technical manuals from the Scania subcorpus (two texts, 2376 tokens), and political prose from the German translation of the Swedish *Statement of Government Policy* (SGP) of 1988 and 1996 (two texts, 4815 tokens). Accuracy percentages are calculated as: CORRECTLY TAGGED TOKENS/ALL TOKENS. For comparison, the best published accuracy figures are also given for the two taggers.

tagger/tagset	*Best published*	*Scania*	*SGP*
TreeTagger	97.5% (Schmid, 1994)	96.3%	96.2%
Morphy/full "large":	84.7% (Lezius et al., 1998)	90.4%	93.8%
Morphy/ reduced "small":	95.9% (Lezius et al., 1998)	94.7%	95.4%

Table 1: The performance of the two taggers

The 'full' and 'red(uced)' tagsets used with Morphy refer to the way tagging errors were counted; with the 'full' tagset, the whole morphosyntactic description (resulting in a 'tagset' with about 1000 tags) had to be correct, i.e., if any part of it was incorrect—e.g., if the case was given as 'dative' instead of 'nominative' (a fairly common error)—the error count would be increased by 1. In the case of the 'reduced' set, however, a correct part of speech[6] together with an error or errors in gender, case, and number for nominal parts of speech, and person/number for finite verbs, only counted as 0.25 errors.

The results seem to confirm that tagger performance is dependent to some extent on the text type. This means that rules for combining taggers most probably will need to refer to the text type (see below). Other than this, however, we can at the present time only note that the dependence of tagger performance on the text type is a topic that merits further investigation.

2.3. Step 3: Finding tagger differences

Next, a correspondence table was constructed for the tag sets of the taggers, for a tagger comparison program developed in the project (see Bengtsson et al., to appear), and the program was used on the output of the taggers to make pairwise comparisons of the taggers. For Morphy, the reduced tagset was used as the basis for the correspondences, as being more directly comparable to the TreeTagger set than the full Morphy tagset.

The hypotheses on which we based the experiment described here were the following.

[4]Except in the case of majority voting without weights, of course, but this method is most suitable in the case of binary classification, which POS tagging is not.

[5]Apart from such obvious considerations as the availability of computational resources for a particular language, of course. Thus, for English, our search for freely available resources turned up three taggers with altogether 10 different tag sets to choose among, while we have not been able so far to find even a single tagger for Polish.

[6]Here we used, roughly, the part of speech inventory of TreeTagger (48+6 tags), so that, e.g., finite verbs, infinitives, and participles were counted as different parts of speech, even though they have the common major part of speech "VER" in Morphy's tag set. In practice, this makes the reduced set very similar, but not identical, to the "small" Morphy tagset (Lezius et al., 1998).

1. that the cases in which the taggers agree are ‘certainly correct’ and not in need of any special treatment,[7]

2. that there would be differences between the two taggers in the errors they made, such that there would be a number of cases where one of the taggers was right and the other wrong (and, crucially, that the overall lower-performing tagger would sometimes be right in these cases), and

3. that these differences would show some systematicity, which could be utilized to improve tagging accuracy by combining the two taggers.

The first hypothesis was not tested in the experiment, but rather taken as axiomatic. The two other hypotheses turned out to be confirmed in the study, however. There were differences between the taggers, and some of the differences seem to be systematic.

In figure 1, we show an (abbreviated) example of the kinds of statistics that the comparison program generates (for a text in the Scania corpus; the output of the comparison program has been translated into English from the original Swedish, and slightly edited for clarity). We see that in roughly 10 percent of the cases, the two taggers disagree on which tag to assign to a word. Earlier we saw (in table 1), that the expected accuracy of the better tagger, TreeTagger, on this text type is in the order of 96%. Hence, some of the disagreements must fall in the remaining 4 percent interval to be useful. In table 2, we give a breakdown into categories and percentages of the 10 percent disagreements. It is expected, of course, that TreeTagger is correct in about 60% of these cases (this follows from its overall accuracy of about 96%). The interesting figure in table 2 is the one in the category labelled “neither correct”—i.e., 5.5% in the Scania case—, because this figure gives an indication of the theoretically attainable maximum accuracy of a combination of the two taggers. Thus, for the Scania texts, this should lie above 99% (since 5.5% of 10% makes about 0.5% of the whole).

Consequently, these results add further support to the idea that it would be worthwhile to explore ways of combining the two taggers. We have seen that others have reached the same conclusion and have experimented with knowledge-poor methods taken from the machine learning work on classifier combination. Here, however, we wish to investigate whether it would be possible to exploit linguistic knowledge to the same end, especially as we do not have training corpora for our text material (a prerequisite for the use of knowledge-poor methods; see above).

```
=============== Results ===============
    1096 tagged units
    Equivalent tags [985 / 89.9 %]
    Non-equivalent tags [111 / 10.1 %]

    ======= Non-equivalent tags: statistics =======
    10 VER 3 SIN corresponding to NN
(verb 3rd singular – common noun)
    7 EIG DAT SIN NEU corresponding to NN
(proper noun dative singular neuter – common noun)
    6 VER 3 PLU corresponding to VVINF
(verb 3rd plural – verb infinitive)
    5 VER PA2 corresponding to NE
(verb perfect participle – proper noun)
    5 ABK corresponding to NE
(abbreviation – proper noun)
    3 VER 3 SIN corresponding to VVPP
(verb 3rd singular – verb perfect participle)
    3 SUB NOM SIN MAS corresponding to NE
(common noun nominative singular masculine
        – proper noun)
    3 SUB NOM PLU FEM corresponding to NE
(common noun nominative plural feminine
        – proper noun)
    3 EIG NOM SIN MAS corresponding to NN
(proper noun nominative singular masculine
        – common noun)
    3 ADJ SOL AKK PLU MAS corresponding to NN
(adjective without article accusative plural masculine
        – common noun)
    2 ZUS corresponding to PAV
(verb supplement – pronominal adverb)
    2 VER PA2 corresponding to ADJD
(verb perfect participle – adverbial/predicative adjective)
    2 VER 3 SIN corresponding to ADJD
(verb 3rd singular – adverbial/predicative adjective)
    2 SUB NOM SIN NEU corresponding to NE
(common noun nominative singular neuter
        – proper noun)
    2 SUB NOM PLU MAS corresponding to NE
(common noun nominative plural masculine
        – proper noun)
    2 SUB DAT SIN NEU corresponding to ADJA
(common noun dative singular neuter
        – attributive adjective)
    2 SUB AKK SIN NEU corresponding to VVFIN
(common noun accusative singular neuter – finite verb)
    2 ART DEF NOM SIN FEM corresponding to PRELS
(article definite nominative singular feminine
        – substitutive relative pronoun)
    2 ADJ IND NOM SIN NEU corresponding to NN
(adjective indefinite nominative singular neuter
        – common noun)
    2 ADJ ADV corresponding to ADJA
(adjective adverbial – attributive adjective)
    [. . . ]
```

Figure 1: Comparison program statistics (tags explained in parentheses)

[7]This assumption has been made elsewhere (Màrquez et al., 1998), although the authors in that case utilize it in a different way than this is done here.

corpus	*Morphy correct*	*TreeTagger correct*	*neither correct*	*total*
RF	101 (35.5%)	176 (62.0%)	7 (2.5%)	284
Scania	86 (36.1%)	139 (58.4%)	13 (5.5%)	238
total	187 (35.8%)	315 (60.4%)	20 (3.8%)	522

Table 2: Tagger differences: Which tagger was right how often?

2.4. Step 4: Finding the systematic differences

Finding the systematic differences between POS taggers implies making a decision as to which variables should be taken into account, i.e. should provide the input parameters for the if–then rules which should be the result of the next step. This amounts to a hypothesis about which factors influence tagging performance, and our initial hypothesis has been that the following parameters are relevant:

- the individual tags themselves;
- disjunctions of tags, denoting linguistically natural categories, e.g., both common nouns and proper nouns are nouns, both verbs and adjectives are verbal words in many languages, etc.;
- the text type, in our case the technical text of the Scania corpus vs. the administrative-political text type of the SGP (see above);

2.5. Step 5: Formulating rules for combining taggers

Using the lists of differences between taggers and the hypothesis about which parameters were likely to influence tagger performance, rules were formulated to choose the output of the inferior tagger (Morphy) over that of the better tagger (TreeTagger) under certain, systematically recurring conditions.

The general format of the rules is:

- *if* Morphy and TreeTagger assign non-equivalent tags to a text word,
  and the following conditions are fulfilled,
  then choose the tag that Morphy assigned,
  else choose the tag that TreeTagger assigned.

The conditions that could be inferred from the test material are shown in table 3.

Thus, we could formulate concrete rules for when to choose Morphy's tag over that of TreeTagger, such as:

text type	TreeTagger tag(s)	Morphy tag(s)	+ cases	– cases
both	–	ABK	33	0
Scania	ADJA, ADJD	SUB *, EIG *	15	1
both	ADJD	VER PA2	7	2
SGP	ADJD	VER *	4	0
SGP	ADV	KON *	13	0
both	NN	ADJ (–ADV)	15	3

Table 3: Inferred rule conditions.

- Regardless of text type, if Morphy says ABK (abbreviation)[8]
- If the text type is SGP, and TreeTagger says ADV, and Morphy says KON *

The conditions are not absolute. In table 3, "+ cases" (or *confirming cases*) indicates the number of times that both a particular condition was fulfilled in the material, and Morphy chose the right tag, while under "– cases" is given the number of times that the condition was fulfilled, but Morphy assigned an incorrect tag.

Given the rules, and the output of the tagger comparison program, it was straightforward to calculate the expected improvements from using the rules, taking into account both the confirming and disconfirming cases. The expected improvements are shown in table 4 (the previous results are repeated there for convenience).

tagging regime	*Scania*	*SGP*
TreeTagger only	96.3%	96.2%
Expected improvement from combination with Morphy (% units)	+1.7	+0.8
Resulting accuracy	98.0%	97.0%

Table 4: Expected tagging improvements using the rules

We see that there is an expected improvement in tagging performance, even a marked improvement in the case of the Scania texts. In this context, we must remember that even an improvement of a single percent unit is much here, considering that the span between the chance baseline and maximum human interjudge agreement is somewhere

[8] TreeTagger does not have a tag for abbreviations, directly corresponding to ABK in Morphy. Instead, abbreviations should be tagged as either common (NN) or proper (NE) nouns, according to the TreeTagger tagging scheme. We decided that Morphy's scheme was better, among other things because of cases like "d. h." = "das heisst", 'that is', tagged by TreeTagger as "ADJA NN", i.e. 'adjective–common noun', and by Morphy as "ABK ABK", i.e. 'abbreviation–abbreviation'. In this case it is doubtful whether the third person indicative singular present tense verb "heisst" even *should* be tagged as a noun. Out of the 33 instances of ABK, 5 corresponded to NE, and 6 to NN. Out of these, only one NN instance was correctly tagged by TreeTagger ("B" in "z. B." = "zum Beispiel", 'for example'.

in the region of 10 percent units (Voutilainen 1999). Earlier, we calculated the maximum theoretically possible improvement in the Scania case to be on the order of 3 percent units ($> 99\% - > 96\%$). Thus the expected improvement from the use of these 6 combination rules is more than half the theoretical maximum, which must be considered a fairly good yield.

3. Discussion

Like I said, the aim of the ETAP project, in which the work reported here was done, is to create an annotated and aligned multilingual parallel corpus.

In our work, we have slowly moved towards a particular vision of what kinds of tools are needed in order to attain this aim. They should allow any number of independent knowledge sources to work together, each contributing a piece of the whole solution. This in itself is not a new idea; many NLP programs use more than one knowledge source to do what they were designed to do. More often than not, however, their combination is 'hard-wired' into the program. What we are experimenting with is a setup where new sources of knowledge can be 'plugged in' to the system without extensive rewiring of the whole application.

For the time being, this 'blackboard model' view of tagging and alignment is more of a conceptual tool than a real one, it is still only a way of looking at the problem of tagging and aligning a multilingual parallel corpus. Even so, this way of looking at things has yielded some interesting fresh ideas and insights.

Thus, the idea presented in this article of combining POS taggers by means of linguistically motivated rules is one that came naturally from seing tagging as a cooperation between independent knowledge sources. As a continuation of the work presented here we are in the process of collecting and evaluating other taggers in the same way, for English, French, Spanish, and Swedish, where we consequently soon hope to be able to present analogous results to the ones shown here (Bengtsson et al., to appear).

Tagger combination is not completely straightforward, however, and there are still many unresolved issues in this connection. One such issue is certainly tagset mapping, which was fairly unproblematic in the case described here, presumably because the two tagsets had been designed with the same kind of applications in mind, but which may be much more problematic in the general case (cf. Teufel, 1995). Tagset mapping is consequently an issue which we will need to pay more attention to, if we find that the results reported here hold for other languages and new text material.

Another idea which has flown naturally out of the blackboard model view referred to above is the following. POS tagging and word alignment should not be seen primarily from the point of view of their process aspect, i.e., as two completely separate processes. Rather, we could look at them from the point of view of the kinds of knowledge involved in the processes. Then we see that some of the knowledge is the same in both cases. Furthermore, if we investigate the interdependencies between the different kinds of knowledge involved, we might discover that 'derived' knowledge in one of the processes could be used as 'input' knowledge in the other process. Thus, for those languages in the ETAP corpus for which we have so far been unable to find taggers (Polish and Serbian–Bosnian–Croatian; cf. above), we have investigated the possibility—reported elsewhere (Borin, to appear c)—of using word alignment to transfer POS tags from one language to another in a multilingual translation corpus. The concrete experiment was made on another language pair, namely Swedish–German, and the fact that these two languages are closely related presumably has had a great impact on the results. The jury is still out on the exact extent of this impact, however, as the typology of part of speech systems is very much a current research topic in linguistics (Anward et al., 1996). Thus it still remains to be seen whether more remotely related languages (such as Swedish and Polish) or totally unrelated languages (such as Swedish and Finnish) will benefit from such a transfer. Here, the possibility offers itself of using word alignment with more than one other language in parallel, in combination with some principle for choosing among conflicting POS assignments, which in turn fits in nicely with the work that we have reported on in this article. Some indications that it could be feasible to use several languages in this way come from experiments that we have performed in word alignment, where we have used one or more other languages as 'pivots', or 'detours', and in this way succeeded in raising word alignment recall without lowering the precision (Borin, to appear d; Borin, to appear e).

The results reported in the preceding section are yet to be confirmed by experiments wherethe rule-governed tagger combination is used on fresh text material from the same corpus. This work had yet not been done at the time when this article was written, but we hope soon to be able to present the results of such experiments.

4. Concluding remarks

There are many directions in which the research reported here could be continued. In particular, we can discern at least the following strands of inquiry, which all are worth pursuing, individually or in various combinations:

- trying to clarify the roles of tagger technology, text type, training corpus size, tag set size, etc., i.e. all the variables that presumably play a role in determining tagger performance, in order to make more informed decisions as to if and how POS taggers are to be combined in order to enhance their performance;

- investigating other context factors which may play a role in formulating the rules for tagger combination; here we have in mind at least lexical information, i.e. the identity of the word itself (or its ending or beginning part) may be a good indicator of when to choose one tag over the other, but there may also be other relevant factors (e.g., sentence length, neighboring tags, etc.) which could come into play here;

- exploring machine learning methods as a way to automatize the rule formulation step in this procedure. The question is which method(s) to investigate, but a natural first candidate would be transformation-based learning (TBL), as we have some experience of working with that method both in the project (e.g. Prütz, to appear) and in our department (e.g. Lager, 1999).

5. Acknowledgements

The research reported here was carried out within the ETAP project, supported by the Bank of Sweden Tercentenary Foundation as part of the research programme *Translation and Interpreting—a Meeting between Languages and Cultures*. See http://www.translation.su.se/. Henrik Oxhammar wrote the tag comparison program described in this article.

6. References

Anward, Jan, Edith Moravcsik, and Leon Stassen, 1996. Parts of speech: A challenge for typology. *Linguistic Typology*, 1(2).

Bengtsson, Camilla, Lars Borin, and Henrik Oxhammar, to appear. Comparing and combining part of speech taggers for multilingual parallel corpora. ETAP research report etap-rr-03, Dept. of Linguistics, Uppsala University.

Borin, Lars, to appear a. The ETAP project — a presentation and status report. ETAP research report etap-rr-01, Dept. of Linguistics, Uppsala University.

Borin, Lars, to appear b. ... and never the twain shall meet. In Lars Borin (ed.), *Parallel Corpora, Parallel Worlds*. Dept. of Linguistics, Uppsala University.

Borin, Lars, to appear c. Alignment and tagging. In Lars Borin (ed.), *Parallel Corpora, Parallel Worlds*. Dept. of Linguistics, Uppsala University.

Borin, Lars, to appear d. Pivot alignment. In *Proceedings of the 12th Nordic Conference on Computational Linguistics (Nodalida99)*.

Borin, Lars, to appear e. You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In *Proceedings of COLING-2000*.

Brill, Eric and Jun Wu, 1998. Classifier combination for improved lexical disambiguation. In *COLING-ACL'98. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Proceedings of the Conference, Vol. I–II*. Montreal: Université de Montréal.

Gellerstam, Martin, 1996. Translations as a source for cross-linguistic studies. In Karin Aijmer, Bengt Altenberg, and Mats Johansson (eds.), *Languages in Contrast. Papers from a Symposium on Text-Based Cross-Linguistic Studies, Lund 4–5 March 1994*. Lund: Lund University Press, pages 53–62.

van Halteren, Hans, Jakub Zavrel, and Walter Daelemans, 1998. Improving data driven wordclass tagging by system combination. On the WWW: cmp-lg/9807013.

Johansson, Stig, to appear. Towards a multilingual corpus for contrastive analysis and translation studies. In Lars Borin (ed.), *Parallel Corpora, Parallel Worlds*. Dept. of Linguistics, Uppsala University.

Lager, Torbjörn, 1999. The μ-TBL system: Logic programming tools for transformation-based learning. In *Proceedings of the Third International Workshop on Computational Natural Language Learning (CoNLL'99)*. Bergen: University of Bergen.

Lezius, Wolfgang, Reinhard Rapp, and Manfred Wettler, 1998. A freely available morphological analyzer, disambiguator, and context sensitive lemmatizer for German. In *COLING-ACL'98. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Proceedings of the Conference, Vol. I–II*. Montreal: Université de Montréal.

Màrquez, Lluís, Lluís Padró, and Horacio Rodríguez, 1998. Improving tagging accuracy by using voting taggers. On the WWW: cs.CL/9809112.

Mason, Oliver, 1997. Qtag–a portable probabilistic tagger. by Corpus Research, University of Birmingham.

Padró, Lluís and Lluís Màrquez, 1998. On the evaluation and comparison of taggers: the effect of noise in testing corpora. In *COLING-ACL'98. 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Proceedings of the Conference, Vol. I–II*. Montreal: Université de Montréal.

Prütz, Klas, to appear. Part-of-speech tagging for Swedish. In Lars Borin (ed.), *Parallel Corpora, Parallel Worlds*. Dept. of Linguistics, Uppsala University.

Schiller, Anne, Simone Teufel, Christine Stöckert, and Christine Thielen, 1995. Vorläufige Guidelines für das Tagging deutscher Textcorpora mit STTS. draft. Technical report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung / Universität Tübingen, Seminar für Sprachwissenschaft.

Schmid, Helmut, 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Teufel, Simone, 1995. A support tool for tagset mapping. In *Proceedings of SIGDAT 1995. Workshop in connection with EACL 95*. Dublin: Association for Computational Linguistics.

Tumer, Kagan and Joydeep Ghosh, 1999. Linear and order statistics combiners for pattern classification. In Amanda Sharkey (ed.), *Combining Artificial Neural Networks*. Berlin: Springer-Verlag, pages 127–162.