Achieving an Almost Correct PoS-Tagged Corpus

Pavel Květoň and Karel Oliva
Austrian Research Institute for Artificial Intelligence (ÖFAI)
Schottengasse 3, A-1010 Wien, Austria
{pavel,karel}@oefai.at

Abstract

After some theoretical discussion on the issue of representativity of a corpus, this paper presents a simple yet very efficient technique serving for (semi-)automatic detection of those positions in a part-of-speech tagged corpus where an error is to be suspected. The approach is based on the idea of learning and later application of "negative bigrams", i.e. on the search for pairs of adjacent tags which constitute an incorrect configuration in a text of a particular language (in English, e.g., the bi-gram *ARTICLE - FINITE VERB*). Further, the paper describes the generalization of the "negative bigrams" into "extended negative bigrams of length *n*", for any natural *n*, which indeed provides a powerful tool for error detection in a corpus. The approach is illustrated throughout on the case of the NEGRA corpus (hence some command of German might be helpful, even though not really necessary). Finally, some general implications for statistical taggers are mentioned.

0. Introduction

One of the pressing problems in current corpus linguistics, if not the most important issue of corpus linguistics at all, is the quality of language corpora. This quality, however, is a many-faceted problem in itself, comprising not only practical matters (such as the assignment of a particular tag to a particular word in a particular corpus), but also issues of more theoretical nature. In order to reflect this division (albeit only roughly and only within a very limited area of part-of-speech tagging of written texts), this paper will have two layers.

In the first layer, we shall touch upon the predominantly theoretical problems of quality of part-of-speech (PoS) tagged corpora, in particular on the issue of representativity of a corpus.

In the second layer, the applications of the theoretical ideas developed in the first layer will be discussed, in particular, we shall

- present some ideas concerning a method for achieving a high-quality (exactly tagged) PoS corpus
- show the practical results which were achieved when this method was applied on the NEGRA corpus of German
- discuss the impact which the method as well as the results achieved have on statistical taggers, in particular on *n*-gram based taggers.

1. The Issues of Representativity of a Corpus

The notion of ***representativity of a PoS-tagged*** corpus of a language can be understood in at least the following two ways:

- representativity wrt. the presence/absence of a phenomenon (broadly defined) of this language,

or

- representativity wrt. the relative frequency of occurrences of a (broadly defined) phenomenon in this language[1].

[1] Obviously, the former case is a trivial instance of the latter one - if a phenomenon occurs in a corpus in the correct relative frequency, then it is present in the corpus if it is a possible phenomenon and it does not occur there (i.e. it occurs with zero frequency, it is absent) if it is an impossible phenomenon.

From this it is easy to conclude that representativity is dependent on what is taken to be a phenomenon (i.e. dependent on what is to be represented - not a surprising conclusion in fact). Thus we might have a corpus representing all the possible agreement patterns in a language, all possible subcategorization patterns (understood as patterns displaying different possibilities of number of complements with a given head and their position around the head), etc. Similarly, we can have a corpus representing all possible kinds of free adjunction. However, we cannot have a corpus representing all possible free adjunction patterns[2], unless we allow for infinite corpora (which is usually not the case).

In order to expose the idea of representativity and its importance in a non-trivial level of detail, let us consider the example of representativity as needed for a learning corpus for a bigram-based statistical tagger[3]. In this case, the phenomena[4] whose presence and relative frequency are at stake are:

- bigrams, i.e. pairs *[First_Tag,Second_Tag]* of tags of words occurring in the corpus adjacently and in this order
- unigrams, i.e. the individual tags.

The ***qualitative representativity*** wrt. to bigrams consists in this case of two complementary parts:

- the representativity wrt. the presence of all possible bigrams of the language in the corpus, which means that if any bigram *[First_Tag,Second_Tag]* is a bigram in a correct sentence of the language, then such a bigram occurs also in the corpus - this might be called positive representativity
- the representativity wrt. the absence of all impossible bigrams of the language in the corpus, which means that if any bigram *[First_Tag,Second_Tag]* is a bigram which cannot occur in a correct (i.e. grammatical) sentence of the language, then such a bigram does not occur in the corpus - this might be called negative representativity

If a corpus is both positively and negatively representative, then indeed it can be said to be a qualitatively representative corpus[5]. In our particular example this means that a bigram occurs in a qualitatively representative (wrt. bigrams) corpus if and only if it is a possible bigram in the language (and from this it already follows that any unigram occurs in such a corpus if and only if it is a possible unigram[6]). From this formulation, it is also clear that the qualitative representativity depends on the notion of grammaticality, that is, on the "language competence", in particular on the ability of distinguishing between a grammatical and an ungrammatical construction (sentence).

The requirement of ***quantitative representativity*** of a corpus wrt. bigrams can then be explained as the requirement that the frequency of any bigram and any unigram occurring in the corpus be "in exactly the right proportion" - i.e. in the proportion "as in the language performance" - to the frequency of occurrence of all other bigrams or unigrams, respectively. However, even when its basic idea is quite intuitive and natural, it is not entirely clear whether quantitative representativity can be formalized really rigorously. The main problem

---

[2] Here for the first time in this paper the tension between a finite language competence and an infinite language performance emerges. We shall devote more space to this point below, albeit only in passing.

[3] The case of a trigram-based tagger, more usual in practice, would be almost identical, but more lengthy. For the conciseness of argument, we limit the discussion to bigram-based taggers only, without loss of generality of argument toward *n*-gram taggers for any natural *n*.

[4] In an indeed broadly understood sense of the word phenomenon - of course a bigram is nothing that would be called a linguistic phenomenon (or at least not called a phenomenon in linguistics).

[5] The definitions of positive and negative representativity are obviously easily transferable to cases with other definitions of a phenomenon. Following this, the definition of qualitative representativity holds of course generally, not only in the particular case of corpus representative wrt. bigrams.

[6] This assertion holds only on condition that each sentence of the language is of length two (measured in words) or longer. Similarly, a corpus qualitatively representative wrt. trigrams is qualitatively representative wrt. bigrams and wrt. unigrams only on condition that each sentence is of length three at least, etc.

here is that what is at stake is measuring the occurrence of a bigram (or unigram, for that matter) within the full "language performance", understood as set of utterances of a language. This set, however, is infinite if considered theoretically (i.e. as set of all possible utterances in the language) and finite but practically unattainable if considered practically as a set of utterances in a language realized within a certain time span (also, due to immanent language change, it is questionable whether such a concept of utterances over a time span is a true performance of a single language).

2. Towards Achieving a Qualitative Representativity wrt. Bigrams

Notwithstanding the theoretical problems, the frequencies are used in practice (e.g., for the purpose of training statistical taggers), and hence it is useful to have a look what they (the practical ones) really mean: in our example, it is the relative frequencies of the bigrams (and unigrams) in a particular (learning or otherwise referential) corpus. However, nothing can be said as to whether this corpus is quantitatively representative or not in the original, intuitive sense[7].

For this reason, it seems more appropriate to strive (at least at the beginning) for achieving a qualitatively representative corpus, rather than for a quantitatively representative one[8].

The starting point for implementing this idea will be the search for a corpus which is qualitatively negatively representative wrt. bigrams. In particular, we shall consider a method of finding and removing "impossible (i.e. ungrammatical) bigrams" in a corpus. In practice, the "impossible bigrams" as a rule occur in a realistic large-scale PoS-tagged corpus for the following reasons:

- in a hand tagged corpus, an "impossible bigram" results from (and unmistakeably signals) either an ill-formed text in the corpus body (including wrong conversion) or a human error in tagging[9]
- in a corpus tagged by a statistical tagger, an "impossible bigram" may result also from an ill-formed source text, as above, and further either from incorrect tagging of the training data (i.e. the error was seen as a "correct configuration (bigram)" in the training data, and was hence learned by the tagger) or from the process of so-called "smoothing", i.e. of assignment of non-zero probabilities also to configurations (bigrams, in the case discussed) which were not seen in the learning phase[10].

For the sake of easiness of explanation of learning the process of detecting errors in PoS-tagging, let us make a provisional and in practice unrealistic assumption (which we shall correct immediately) that we have a qualitatively representative (wrt. bigrams) corpus of sentences of a certain language at our disposal[11]. Given such a (hypothetical) corpus, all the bigrams in the corpus are to be collected to a set **CB** (correct bigrams), and then the complement of **CB** to the set of all possible bigrams is to be computed; let this set be called **IB** (incorrect bigrams). The idea is now that if any element of **IB** occurs in a PoS-tagged

---

[7] This is not to be understood as criticism of the selection and mix of corpus sources. However, it is to be understood in the sense that such a selection is always a matter of more or less subjective decision, and the rigid formalization of the concept of quantitative representativity in the above sense is very difficult at best.

[8] In addition, even if there occurred no problems with the quantitative representativity, obtaining a qualitatively representative corpus is a special (and easier) case of obtaining a quantitatively representative one, cf. also footnote 1 above.

[9] Exempted are metalinguistic contexts (e.g., quotations) and collocations displaying otherwise ungrammatical patterns - these will be discussed shortly below. This proviso holds throughout the paper.

[10] This "smoothing" is necessary since - put very simply - otherwise configurations (bigrams) which were not seen during the learning phase cannot be processed if they occur in the text to be tagged.

[11] Note that we do not presuppose that the corpus is error-free - it might well contain tagging errors (and possibly other errors, e.g. ungrammaticalities in the input), only that none of them must cause an "impossible bigram" to occur in the corpus.

corpus whose correctness is to be checked, then the two adjacent corpus positions where this happened must contain an error (which then can be corrected).

When implementing this approach to error detection, it is first of all necessary to realize that learning the "impossible bigrams" is extremely sensible to the qualitative representativity of the learning corpus:

- the presence of an erroneous bigram in the set of **CB** causes that the respective error cannot be detected in the corpus whose correctness is to be checked (even a single occurrence of a bigram in the learning corpus means correctness of the bigram),
- the absence of a correct bigram from the **CB** set causes this bigram to occur in **IB**, and hence any of its occurrences in the checked corpus to be marked as a possible error (absence of a bigram in the learning corpus means incorrectness of the bigram).

However, the available corpora are neither error-free nor qualitatively representative. Therefore, in practice these deficiencies have to be compensated for by appropriate means. When applying the approach to NEGRA®, we employed

- bootstrapping for achieving positive qualitative representativity (to an extent relevant for bigrams, i.e. error-freeness on the level which does not give rise to "impossible bigrams")
- manual pruning of the **CB** and **IB** sets for achieving at negative qualitative representativity[12].

We started by very careful hand-cleaning errors in a very small sub-corpus of about 80 sentences (about 1.200 words). From this small corpus, we generated the **CB** set, and pruned it manually, using linguistic knowledge (as well as linguistic imagination) about German syntax. Based on the **CB** set achieved, we generated the corresponding **IB** set and pruned it manually again. The resulting **IB** set was then used for automatic detection of "suspect spots" in the sample of next 500 sentences from the corpus, and for hand-elimination of errors in this sample where appropriate (obviously, not all **IB** violations were genuine errors !). Thus we arrived at a cleaned sample of 580 sentences, which we used just in the same way for generating **CB** set, pruning it, generating **IB** set and pruning this set, arriving at an **IB** set which we used for detection of errors in the whole body of the corpus (about 20.500 sentences, 350.000 positions).

The procedure was then re-applied to the whole corpus. For this purpose, we divided the corpus into four parts of approximately 5.000 sentences each. Then, proceeding in four rounds, first the **IB** set was generated (already without manual checking) out of 15.000 sentences and then the **IB** set was applied to the rest of the corpus (on the respective 5.000-sentence partition). The corrections based on the results improved the corpus to such an extent that we made the final round, this time dividing the corpus into 20 partitions with approximately 1.000 sentences each and then reapplying the whole process 20 times.

3. Advanced Phase: Extending the Bigrams

The "impossible bigrams" are a powerful tool for checking the correctness of a corpus, however, a tool which works on a very local scale only, since it is able to detect solely errors which are detectable as deviations from the set of possible pairs of adjacently standing tags. Thus, obviously, quite a number of errors remain undetected by such a strategy. As an example of such an as yet "undetectable" error in German we might take the configuration where two words tagged as finite verbs are separated from each other by a string consisting of nouns, adjectives, articles and prepositions only. In particular, such a configuration is

[12] The fact that the lack of representativity of a realistic corpus has to be compensated for manually, that is, by employing the human language competence which cannot be replaced by any other means, is absolutely central to the method described !

erroneous since the rules of German orthography require that some kind of clause separator (comma, dash, coordinating conjunction) occur inbetween two finite verbs[13].

In order to be able to detect also such kind of errors, the above "impossible bigrams" have to be extended substantially. Searching for the generalization needed, it is first of all necessary to get a linguistic view on the "impossible bigrams", in other words, to get a deeper insight into the impossibility for a certain pair of PoS-tags to occur immediately following each other in any linguistically correct and correctly tagged sentence. The point is that this indeed does not happen by chance, that any "impossible bigram" comes into being as a violation of a certain - predominantly syntactic[14] - rule(s) of the language. Viewed in more detail, these violations might be of the following nature:

- Violation of constituency. The occurrence of an "impossible bigram" in the text signals that - if the tagging were correct - there is a basic constituency relation violated (resulting in the occurrence of the "impossible bigram"); as an example of such configuration, we might consider the bigram PREPOSITION - FINITE VERB (possible German example string: *...für-PREP reiche-VFIN...*). From this it follows that either there is indeed an error in the source text (in our example, probably a missing word, e.g., *Der Sprecher der UNO-Hilfsorganisation teilte mit, für (Arme) reiche diese Hilfe nicht.*) or there was a tagging error detected (in the example, e.g., an error as in the sentence *... für reiche Leute ist solche Hilfe nicht nötig...*). The source of the error is in both cases violation of the linguistic rule postulating that, in German, a preposition must always be followed by a corresponding noun (NP) or at least by an adjectival remnant of this NP[15].
- Violation of feature cooccurrence rules (such as agreement, subcategorization etc.). The point here is that there exist configurations such that if two wordforms (words with certain morphological features) occur next to each other, they necessarily stand in such a configuration, and because of this also in a certain grammatical relation. This relation, in turn, poses further requirements on the (morphological) features of the two wordforms, and if these requirements are not met, the tags of the two wordforms result in an "impossible bigram". Let us take an example again, this time with tags expressing also morphological characteristics: if the words *... Staaten schickt ...* are tagged as *Staaten-NOUN-MASC-PL-NOM* and *schickt-MAINVERB-PRES-ACT-SG*, then the respective tags *NOUN-MASC-PL-NOM* and *MAINVERB-PRES-ACT-SG* (in this order) create an "impossible bigram". The reason for this bigram being impossible is that if a noun in nominative case occurs in a German clause headed by a finite main verb different from *sein/werden* (which, however, are not tagged as main verbs in the STTS tagset (Schiller et al., 1999) used in NEGRA), then either this noun must be the verb's subject, which in turn requires that the noun and the verb agree in number, or that the noun is a part of coordinated subject, in which case the verb must be in plural. The configuration from the example meets neither of these conditions, and hence it generates an "impossible bigram".

---

[13] At stake are true regular finite forms, exempted are words occurring in fixed collocations which do not function as heads of clauses. As an example of such usage of a finite verb form, one might take the collocation *wie folgt*, e.g., in the sentence *Diese Übersicht sieht wie folgt aus: ...* Mind that in this sentence, the verb *folgt* has no subject, which is impossible with any active finite verb form of a German verb subcategorising for a subject (and possible only marginally with passive forms, e.g., in *Gestern wurde getanzt*, or – obviously – with verbs which do not subcategorise for a subject, such as *frieren, grauen* in *Mich friert, Mir graut vor Statistik*).

[14] Examples of other such violations are rare and are related mainly to phonological rules. In English, relevant cases would be the word pairs *an table, a apple*, provided the tagset were so fine-grained to express such a distinction, better examples are to be found in other languages, e.g. the case of the Czech ambiguous word *se*, cf. (Oliva, to appear).

[15] unlike English, (standard) German has no preposition stranding and the like - we disregard the (heavily) colloquial examples like *Da weiss ich nix von*.

The central observation lies then in the fact that the property of being an impossible configuration can often be retained also after the components of the "impossible bigram" get separated by material occurring inbetween them. Thus, for example, in both our examples the property of being an impossible configuration is conserved if an adverb is placed inbetween, creating thus an "impossible trigram". In particular, in the first example, the configuration *PREP ADV VFIN* cannot be a valid trigram, exactly for the same reasons as *PREP VFIN* was not a valid bigram: *ADV* is not a valid NP remnant. In the second case, the configuration *NOUN-MASC-PL-NOM ADV MAINVERB-PRES-ACT-SG* is not a valid trigram either, since obviously the presence (or absence) of an adverb in the sentence does not change the subject-verb relation in the sentence. In fact, due to recursivity of language, also two, three and in fact any number of adverbs would not make the configuration grammatical and hence would not disturb the error detection potential of the "extended impossible bigrams" from the examples.

These linguistic considerations have a straightforward practical impact. Provided a qualitatively representative corpus is available, it is possible to construct the **IB** set. Then, for each bigram *[First,Second]* from this set, it is possible to collect all trigrams of the form *[First,Middle,Second]* occurring in the corpus, and collect all the possible tags *Middle* in the set *Possible_Inner_Tags*. Furthermore, given the impossible bigram *[First,Second]* and the respective set *Possible_Inner_Tags*, the learning corpus is to be searched for all tetragrams *[First,Middle_1,Middle_2,Second]*. In case one of the tags *Middle_1, Middle_2* occurs already in the set *Possible_Inner_Tags*, no action is to be taken, but in case the set *Possible_Inner_Tags* contains neither of *Middle_1, Middle_2*, both the tags *Middle_1* and *Middle_2* are to be added into the set *Possible_Inner_Tags*. The same action is then to be repeated for pentagrams, hexagrams, etc., until the maximal length of sentence in the learning corpus prevents further prolongation of the *n*-grams and the process terminates.

If now the set *Impossible_Inner_Tags* is constructed as the complement of *Possible_Inner_Tags* relatively to the whole tagset, then any *n*-gram consisting of the tag *First*, of any number of tags from the set *Impossible_Inner_Tags* and finally from the tag *Second* is very likely to be an *n*-gram impossible in the language and hence if it occurs in the corpus whose correctness is to be checked, it is to be flagged as a "suspect spot". Obviously, this idea is again based on the assumption of representativity of the learning corpus, so that for training on a realistic corpus the correctness of the resulting "extended bigrams of length *n*" has to be hand-checked. This, however, is well-worth the effort, since the resulting "extended bigrams of length *n*" are an extremely efficient tool for error detection.

In the current stage of research, we have not yet performed the extension from (negative) representativity wrt. bigram to (negative) representativity wrt. trigrams, which might possibly help to discover more errors in the tagging of the NEGRA corpus. In particular, there exist "impossible trigrams" *[First,Second,Third]* which cannot be detected as such (i.e. as impossible) by the using bigrams as described above if the *[First,Second]*, *[Second,Third]* and *[First,Third]* are all possible bigrams[16], that is, the occurrence of such an error (impossible, erroneous trigram) cannot be detected by applying the search for impossible bigrams. However, we plan this to be performed in the near future (for sure before TSD 2002 convenes). Mind the fact that even if the set of all trigrams is much larger than the set of all bigrams, a very substantial subset of this set need not be searched through manually once the previous results concerning impossible bigrams are available, since:

- all those candidates *[First,Second,Third]* for impossible trigram which contain an impossible bigram *[First,Second]* or *[Second,Third]*can be discarded automatically from the search space (these are impossible as bigrams, hence certainly also impossible as trigrams)

[16] A possible example, both for Czech and for German, is the trigram *[preposition,preposition,relative_pronoun]*.

- all those candidates *[First,Second,Third]* for impossible trigram which have been discovered as possible extended bigrams are to be eliminated automatically from the search space, too, since they are already known to be possible trigrams.

4. Evaluation of the Results

By means of the error-detection technique described above, we were able to correct 2.661 errors in the NEGRA corpus. The prevailing part of these errors was indeed that of incorrect tagging (only less than 8% were genuine ungrammaticalities in the source, about 26% were errors in segmentation). The whole resulted in changes on 3.774 lines of the corpus; the rectification of errors in segmentation resulted in reducing the number of corpus positions by over 700, from 355.096 to 354.354.

After finishing the corrections, we experimented with training and testing the TnT tagger (Brants, 2000) on the "old" and on the "corrected" version of NEGRA. We used the same testing as described by Brants, i.e. dividing each of the corpus into ten contiguous parts of equal size, each part having parallel starting and end position in each of the versions, and then running the system ten times, each time training on nine parts and testing on the tenth part, and finally computing the mean of the quality results. In doing so, we arrived at the following results:

- if both the training and the testing was performed on the "old" NEGRA, the tags assigned by the TnT tagger differed from the hand-assigned tags within the test sections on (together) 11.138 positions (out of the total of 355.096), which yields the error rate of 3,14%
- if both the training and the testing was performed on the "correct" NEGRA, the tags assigned by the TnT tagger differed from the hand-assigned tags of the test sections on (together) 10.889 positions (out of the total of 354.354), which yields the error rate of 3,07%
- in the most interesting final experiment, the training was performed on the "old" and the testing on the "correct" NEGRA; in the result, the tags assigned by TnT differed from the hand-assigned tags in the test sections on (together) 12.075 positions (out of the total of 354.354), yielding the error rate of 3,41%.

These results show that there was only a negligible (and, according to the χ^2 test, statistically insignificant) difference between the results in the cases when the tagger was both trained and tested on "old" corpus and both trained and tested on the "corrected" corpus. However, the difference in the error rate when the tagger was once trained on the "old" and once on the "corrected" version, and then in both cases tested on the "corrected" version[17], brought up a relative error improvement of 9,97%. This improvement documents the hardly surprising truth that it is not only the size but also the correctness of the training data that is absolutely essential for the results of a statistical tagger.

5. Conclusions

The main contribution of this paper lies in the presentation of a method for detecting errors in part-of-speech tagged corpus which is both quite powerful (as to coverage of errors) and easy to apply, and hence it offers a relatively low-cost means for achieving high-quality PoS-tagged corpora. The main advantage is that the approach described is based on the combination of focussed search for errors of a particular, specific type with bootstrapping of the search, which makes it possible to detect errors even in a very large corpus where manual checking would not be feasible (at least in practice), since it requires passing through the whole of the text and paying attention to all kinds of possible violations - while the approach described concentrates on violations of particular phenomena on particular spots. Hence, it

[17] For obvious reasons, we did not even consider training on the "corrected" corpus and testing on the "old" one.

allows for straightforward checking whether an error really occurs - and if so, for a direct correction.

As a side-effect, it should be also mentioned that the method allows not for detecting errors only, but also for detecting inconsistencies in hand-tagging (i.e. differences in application of a given tagging scheme by different human annotators and/or in different time), and even inconsistencies in the tagging guidelines. A particular issue is further the area of detecting and tagging idioms and collocations, in the particular case when these take a form which makes them deviate from the rules of standard syntax (i.e. they are detected as "suspect spots" by the method). For details on all these points, including the particular problems encountered in NEGRA, cf. Květoň and Oliva (in prep.).

Acknowledgement

This work has been sponsored by the *Fonds zur Förderung der wissenschaftlichen Forschung (FWF), Grant No. P12920*. The *Austrian Research Institute for Artificial Intelligence (ÖFAI)* is supported by the *Austrian Federal Ministry of Education, Science and Culture*.

References

Brants T.: (2000) *TnT – A Statistical part-of-speech tagger*, in: Proceedings of the 6th Applied Natural Language Processing Conference, Seattle

Hirakawa H., K. Ono and Y. Yoshimura (2000) *Automatic refinement of a PoS tagger using a reliable parser and plain text corpora*, in: Proceedings of the 18th Coling conference, Saarbrücken

Květoň P. and K. Oliva (in prep.) *Correcting the NEGRA Corpus: Methods, Results, Implications*, ÖFAI Technical Report

Müller F.H. and T. Ule (2001) *Satzklammer annotieren und tags korrigieren: Ein mehrstufiges top-down-bottom-up System zur flachen, robusten Annotierung von Sätzen im Deutschen*, in: Proceedings der GLDV-Frühjahrstagung 2001, Gießen

NEGRA®: www.coli.uni-sb.de/sfb378/negra-corpus

Oliva K. (2001) *The possibilities of automatic detection/correction of errors in tagged corpora: a pilot study on a German corpus*, in: 4th International conference "Text, Speech and Dialogue" TSD 2001, Lecture Notes in Artificial Intelligence 2166, Springer, Berlin 2001

Oliva K. (to appear) *Linguistics-based tagging of Czech: disambiguation of 'se' as a test case*, in: Proceedings of 4th European Conference on Formal Description of Slavic Languages held in Potsdam from 28th till 30th November 2001

Petkevič V. (2001) *Grammatical agreement and automatic morphological disambiguation of inflectional languages*, in: 4th International conference "Text, Speech and Dialogue" TSD 2001, Lecture Notes in Artificial Intelligence 2166, Springer, Berlin 2001

Schiller A., S. Teufel, C. Stöckert and C. Thielen (1999) *Guidelines für das Tagging deutscher Textcorpora*, University of Stuttgart / University of Tübingen

Skut W., B. Krenn, T. Brants & H. Uszkoreit (1997) *An annotation scheme for free word order languages*, in: Proceedings of the 3rd Applied Natural Language Processing Conference, Washington D.C.