(Semi-)Automatic Detection of Errors in PoS-Tagged Corpora

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

Abstract

This paper presents a simple yet in practice very efficient technique serving for 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 bigram *ARTICLE - FINITE VERB*). Further, the paper describes the generalization of the "negative bigrams" into "negative *n*-grams", for any natural *n*, which indeed provides a powerful tool for error detection in a corpus. The implementation is also discussed, as well as evaluation of results of the approach when used for error detection in the NEGRA® corpus of German, and the general implications for the quality of results of statistical taggers. Illustrative examples in the text are taken from German, and hence at least a basic command of this language would be helpful for their understanding - due to the complexity of the necessary accompanying explanation, the examples are neither glossed nor translated. However, the central ideas of the paper should be understandable also without any knowledge of German.

1. Errors in PoS-Tagged Corpora

The importance of correctness (error-freeness) of language resources in general and of tagged corpora in particular cannot probably be overestimated. However, the definition of what constitutes an error in a tagged corpus depends on the intended usage of this corpus.

1.1 If we consider a quite typical case of a Part-of-Speech (PoS) tagged corpus used for training statistical taggers, then an error is defined naturally as any deviation from the regularities which the system is expected to learn; in this particular case this means that the corpus should contain neither errors in assignment PoS-tags nor ungrammatical constructions in the corpus body[1], since if any of the two cases is present in the corpus, then the learning process necessarily:

- gets a confused view of probability distribution of configurations (e.g., trigrams) in a correct text

and/or, even worse (and, alas, much more likely)

- gets positive evidence also about configurations (e.g., trigrams) which should not occur as the output of tagging linguistically correct texts, while simultaneously getting less evidence about correct configurations.

1.2 If we consider PoS-tagged corpora destinated for testing NLP systems, then obviously they should not contain any errors in tagging (since this would be detrimental to the validity of results of the testing) but on the other hand they should contain a certain amount of ungrammatical constructions, in order to test the behaviour of the tested system on a realistic input.

Both these cases share the quiet presupposition that the tagset used is linguistically adequate, i.e. it is sufficient for unequivocal and consistent assignment of tags to the source text[2].

---

[1] In this paper we on purpose do not distinguish between "genuine" ungrammaticality, i.e. one which was present already in the source text, and ungrammaticality which came into being as a result of faulty conversion of the source into the corpus-internal format, e.g., incorrect tokenization, OCR-errors, etc.

[2] This problem might be – in a very simplified form – illustrated on an example of a tagset introducing tags for NOUNs and VERBs only, and then trying to tag the sentence *John walks slowly* - whichever tag is assigned to the word *slowly*, it is obviously an incorrect one. Natural as this requirement might

1.3 As for using annotated corpora for linguistic research, then it seems that even inadequacies of tagset are tolerable provided they are marked off properly - in fact, these spots in the corpus might well be quite an important source of linguistic investigation since, more often than not, they constitute direct pointers to occurrences of linguistically "interesting" (or at least "difficult") constructions in the text.

2. Automatic PoS-Tagging Errors Detection

In the following, we shall concentrate on the first case mentioned above, i.e. on methods and techniques of generating "completely error-free" corpora, or, more precisely, on the possibilities of (semi-)automatic detection (and hence correction) of errors in a PoS-tagged corpus. Due to this, i.e. to the aim of achieving an "error-free" corpus, we shall not distinguish between errors due to incorrect tagging, faulty conversion or ill-formed input, and we shall treat them on a par.

The approach as well as its impact on the correctness of the resulting corpus will be demonstrated on the version 2 of the NEGRA® corpus of German (for the corpus itself see www.coli.uni-sb.de/sfb378/negra-corpus, for description cf. Skut et al. (1997)). However, we believe the solutions developed and presented in this paper are not bound particularly to correcting this corpus or to German, but hold generally.

The error search we use has several phases which differ in the amount of context that has to be taken into consideration during the error detection process. Put plainly, the extent of context mirrors the linguistic complexity of the detection, or, in other words, at the moment when the objective is to search for "complex" errors, the "simple(r)" errors should be already eliminated.

The first, preliminary phase, is thus the search for errors which are detectable absolutely locally, i.e. without any context at all.

2.1 Preliminary Phase: Trivial Errors

When aiming at correction of errors in general, the basic condition which is to be met is that the local assignment of PoS-tags is if not correct then at least (morphologically) plausible. In particular, the first errors to be corrected are those where the assignment of PoS-tags violates morphological (and possibly other local, e.g., phonological) laws of the language. Important point is, however, that only the error ***detection*** is strictly local - for the ***correction***, a vaster context might be (and as a rule is) needed.

From this it follows that the first phase should be the search for "impossible unigrams", i.e. for tags which are assigned in conflict with morphological or lexical information. A simple (but unrealistic) example from English would be the cases that the word *table* were assigned the tag PLURAL-NOUN or the tag PREPOSITION. As realistic examples from NEGRA®, it is possible to put forward tagging the (German !) word *die* as a masculine singular form of an article, tagging *ein* as a definite article, assigning a word starting with a capital letter and standing on a non-first position of a sentence a verbal tag (typically, such word is a verbal noun) or tagging *bis* as a preposition requiring dative case[3].

A particular case of locally recognizable errors is constituted by numerals with round thousands, written in digits, with blank between the last three zeroes (e.g., *12 000*) which in NEGRA® are systematically tokenized/segmented as two cardinal numbers following each other, e.g., *12 000* is segmented as

<position> 12 tag=CARD <end of position>
<position> 000 tag=CARD <end of position>

while it is obviously to be segmented as a single numeral, i.e.

<position> 12 000 tag=CARD <end of position>

2.2 Medium Phase: Impossible Bigrams

The errors described in the previous section were cases of incorrect morphological analysis (e.g., *die* tagged as masculine singular), errors in lexical analysis (the case of preposition *bis* tagged as requiring a dative case) and diverse errors in lemmatization, conversion and segmentation, and if discussed alone, they had better be classified as such. In fact, calling these kind of errors

seem, it is in fact not met fully satisfactorily in any tagset we know of; for more, cf. Květoň and Oliva (in prep.).

[3] The corrections to be performed are not presented, since they might differ from case to case, in dependence on the particular context.

"impossible unigrams" (as above) makes little sense apart from serving as a motivation for error detection based on search for "impossible *n*-grams", i.e. *n*-tuples ($n \in N$) of tags which, if occuring as tags of adjacent words in a text of a particular language, constitute a violation of (syntactic) rules of this language.

The starting point for application of this idea is the search for "impossible bigrams". These 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
- 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[4].

For 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 an error-free and representative (wrt. bigrams) corpus of sentences of a certain language at our disposal. By saying ***error-free*** and ***representative***, we have in mind that, for the case of bigrams:

- any sentence in the set of sentences constituting the corpus is a grammatical sentence of the language in question (error-freeness wrt. source)
- any bigram can occur in a grammatical sentence of the language if and only if it occurs at least once in the corpus (i.e. if any bigram is a possible bigram in the language, it occurs in the corpus (representativity), if any bigram is an "impossible bigram", it does not occur in the corpus (error-freeness wrt. tagging)).

[4] 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.

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 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 both to error-freeness and to 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 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 correctness
- manual pruning of the **CB** and **IB** sets for achieving representativity.

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, prun-

ing 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 (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.

2.3 Advanced Phase: Variable-length *n*-grams

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 erroneous since the rules of German orthography require that some kind of clause separator (comma, dash, coordinating conjunction) occur inbetween two finite verbs[5].

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[6] - 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 "imposible 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[7].
- 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

[5] 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 subcategorizing for a subject (and possible only marginally with passive forms, e.g., in *Gestern wurde getanzt*, or – obviously – with verbs which do not subcategorize for a subject, such as *frieren, grauen* in *Mich friert, Mir graut vor Statistik*).

[6] 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).

[7] Unlike English, (standard) German has no preposition stranding and similar phenomena - we disregard the colloquial examples like *Da weiss ich nix von*.

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 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".

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 configurations 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 an error-free and representative (in the above sense) 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,Between,Second]* occurring in the corpus, and collect all the possible tags *Between* 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 learn corpus prevents any 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 signalled as a "suspect spot". Obviously, this idea is again based on the assumption of error-freeness and representativity of the learning corpus, so that for training on a realistic corpus the correctness of the resulting "impossible *n*-grams" has to be hand-checked. This, however, is well-worth the effort, since the resulting "impossible *n*-grams" are an extremely efficient tool for error detection.

The implementation of the idea is a straightforward extension of the above approach to "impossible bigrams".

2.4 Extensions

The above approach does not guarantee, however, that all "impossible *n*-grams" are considered. In particular, any "impossible trigram" *[First,Second,Third]* cannot be detected as such (i.e. as impossible) if the *[First,Second]*,

[Second,Third] and *[First,Third]* are all possible bigrams (i.e. they all belong to the set **CB**). Such an "impossible trigram" in German is, e.g., *[nominative-noun,main_verb,nominative-noun]* - this trigram is impossible[8] since no German verb apart from *sein/werden* (which, as said above, are not tagged as main verbs in NEGRA®) can occur in a context where a nominative noun stands both to its right and to its left, however, all the respective bigrams occur quite commonly (e.g., *Johann schläft, Jetzt schläft Johann, König Johann schläft*). Here, an obvious generalization of the approach from "impossible bigrams" to "impossible trigrams" (and "impossible tetragrams", etc.) is possible, however, we did not perform this in full due to the amount of possible trigrams as well as to the data sparseness problem which, taken together, would make the manual work on checking the results unfeasible in practice. We rather applied only about 20 "impossible trigrams" and 6 "impossible tetragrams" stemming from "linguistic invention" (such as the trigram discussed above).

3. Evaluation of the Results

By means of the error-detection techniques described above, we were able to correct 2.661 errors in the NEGRA® corpus. These errors were of all sorts mentioned in Sect. 1, however the prevailing part was that of incorrect tagging (only less than 8% were genuine source errors, 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[9].

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[10], brought up a relative error improvement of 9,97%. This improvement documents the old and hardly surprizing truth that - apart from the size - also the correctness of the training data is absolutely essential for the results of a statistical tagger.

---

[8] Exempted are quotations and other metalinguistic contexts, such as *Der Fluss heisst Donau, Peter übersetzte Faust - eine Tragödie ins Englische als Fist - one tragedy*, which, however, are as a rule lexically specific and hence can be coped with as such.

[9] Which is a much nicer number than 355.096, and thus an additional motivation for correcting corpora ☺

[10] We did not perform training on the "corrected" corpus and testing on the "old" one, because it is not clear how the results of such an experiment should be evaluated: in particular, in such a case it is to be expected that it often happens that the tags assigned by the tagger and the ones in the "pyrite standard" (since it cannot be really called "golden", then) differ due to an error in the "standard" - and hence the measuring of the accuracy of the results of the tagger are problematic at best within such an architecture.

4. 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 allows for straightforward checking whether an error really occurrs - 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.