

# Inter-Annotator Agreement for a German Newspaper Corpus

Thorsten Brants

Saarland University, Computational Linguistics  
D-66041 Saarbrücken, Germany  
thorsten@coli.uni-sb.de

In *Second International Conference on Language Resources and Evaluation LREC-2000*  
May 31 – June 2, 2000, Athens, Greece

## Abstract

This paper presents the results of an investigation on inter-annotator agreement for the NEGRA corpus, consisting of German newspaper texts. The corpus is syntactically annotated with part-of-speech and structural information. Agreement for part-of-speech is 98.6%, the labeled F-score for structures is 92.4%. The two annotations are used to create a common final version by discussing differences and by several iterations of cleaning. Initial and final versions are compared. We identify categories causing large numbers of differences and categories that are handled inconsistently.

## 1. Introduction

One large problem of each annotation project is consistency. This entails inter-annotator consistency (i.e., two annotators annotate the same sentence equally) and intra-annotator consistency (i.e., if an annotator encounters the same sentence, or part thereof, again, he annotates them equally). Consistency needs to be maintained during the whole annotation project and possibly for a large number of annotators. Consistency highly increases the usefulness of a corpus for training or testing automatic methods, and for linguistic investigations.

Maintaining consistency requires large efforts. During the annotation of the NEGRA corpus<sup>1</sup> (Skut et al., 1997; Brants et al., 1999), we developed very efficient interactive annotation tools. Based on graphical feedback (Brants and Plaehn, 2000), the annotator interacts with a tagger and a parser running in the background (Brants, 1999). A trained annotator needs on average 50 seconds per sentence with an average length of 17.5 tokens (around 1,300 tokens/hour) for part-of-speech plus structural annotation. Despite this very fast initial pass, we found that the total annotation requires approx. 10 minutes/sentence. The latter is the sum of the time spent by the involved annotators and includes:

- a) two independent annotations,
- b) correction of obvious errors that occur during comparison,
- c) discussion and correction of the remaining differences,
- d) the training phase of the annotator,
- e) changes to the corpus that are required because of a change in the annotation scheme.

Part a) needs less than two minutes, parts d) and e) are more or less fixed amounts of time that are restricted to the project's or annotator's initial phase. Most time is spent on

parts b) and c), hence it is due to inter-annotator disagreement. This paper investigates the differences in annotations for the NEGRA corpus. The aim of the investigation is to detect and classify the differences. This information is used to improve handling of problematic phenomena and to increase inter-annotator consistency. A side effect of this is higher annotation speed since less differences need to be eliminated.

We investigate the records of parts of the NEGRA corpus. These parts consist of 10,500 sentences with recorded changes in structural annotations and 8,500 sentences with recorded changes in part-of-speech annotations. The first annotations of both annotators, as well as the changes in each sentence are archived.

The structural annotation consists of possibly discontinuous constituents, labeled nodes (25 phrase types) and labeled edges (45 grammatical functions). For part-of-speech, we use the Stuttgart-Tübingen-Tagset STTS consisting of 54 tags (Thielen and Schiller, 1995, cf. appendix A). Figure 1 shows an example sentence and its annotation.

An experiment on the upper bound of interjudge agreement for part-of-speech tagging was presented by (Voutilainen, 1999). His experiment differs from our investigation. He used trained linguists with years of experience for the annotation, while our corpus is created by hired students. Furthermore, he used a different tagset, which avoids some decisions made in our tagset. Therefore, we expected (and actually found) lower rates of agreement for our project.

The study of (Véronis, 1998) is concerned with inter-annotator agreement for the task of word sense disambiguation. As for part-of-speech, the information is annotated at the word level, although it is a completely different type of annotation.

We are not aware of similar investigations on structural agreement.

## 2. Measures

For part-of-speech tagging, we compare two initial annotations (versions A and B) and the annotations after several steps of discussion and cleaning (version FINAL). We

<sup>1</sup>For availability, please check <http://www.coli.uni-sb.de/sfb378/negra-corpus/>

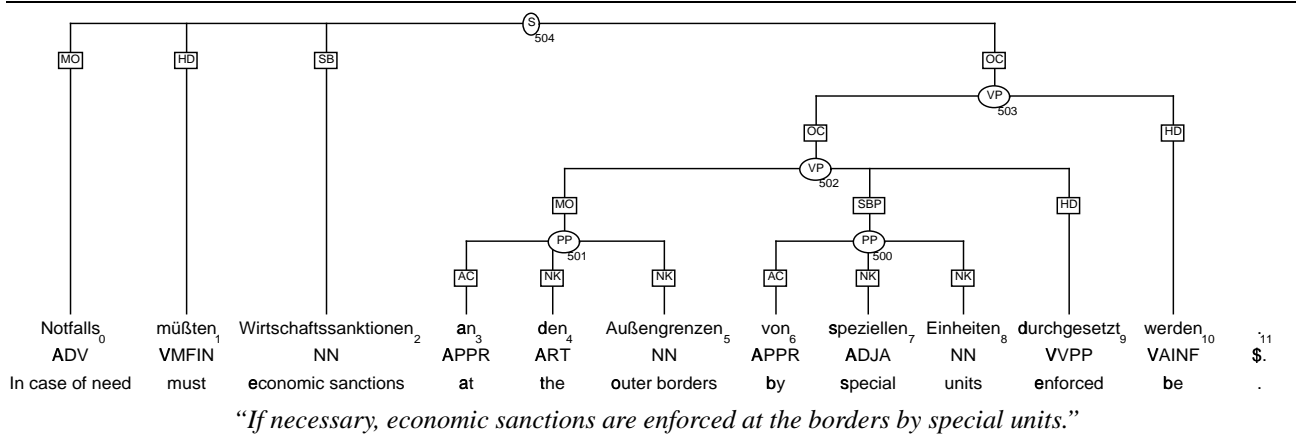


Figure 1: Example sentence with part-of-speech and structural annotation.

use the same measure that is used for indicating tagging accuracy of automatic taggers. For each word, the annotator performs a full disambiguation (i.e., exactly one tag is assigned to each word), and we determine for two tagged versions  $X$  and  $Y$  of one corpus:

$$\text{accuracy}(X, Y) = \frac{\text{number of tokens tagged identically}}{\text{number of tokens in the corpus}} \quad (1)$$

We also make a comparison for the structural annotation. The standard measures of recall and precision can be used. They have a slightly different interpretation, since comparing version A and B does not involve a “correct” annotation. When comparing two annotations  $X$  and  $Y$ , these are

$$\text{recall}(X, Y) = \frac{\text{number of identical nodes in } X \text{ and } Y}{\text{number of nodes in } X} \quad (2)$$

$$\text{precision}(X, Y) = \frac{\text{number of identical nodes in } X \text{ and } Y}{\text{number of nodes in } Y} \quad (3)$$

F-Score is the harmonic mean of both:

$$F = \frac{2PR}{P + R} \quad (4)$$

Note that  $\text{recall}(X, Y) = \text{precision}(Y, X)$ . Since there is no identified correct version when comparing structures of two annotators, the F-score is probably the most appropriate one of these three measures for our purposes.

Phrases in the NEGRA corpus can be discontinuous. Testing for identical nodes therefore requires to check more than just the phrase boundaries. We adapt an approach of (Calder, 1997) who uses the terminal yield to align context-free trees. Extending this to discontinuous annotations, two nodes in two different annotations are identical if they have the same terminal yield.

In addition to overall agreement rates, we list those part-of-speech tags and phrase types which are involved in large numbers of differences, and those with very low F-scores. Recall and precision for part-of-speech tags are calculated using the same formulas as in the structural case. Just

replace “node” with “tag”: we take into account the frequency of a particular tag in annotation A, its frequency in B, and the number of identical annotations in A and B.

Categories which cause large numbers of differences are good candidates for improving inter-annotator agreement. A better handling of these categories has the potential of eliminating large numbers of differences. On the other hand, a very low F-score identifies categories which are handled inconsistently. These categories do not need to be very frequent (as we will see in the results section). Nevertheless, a better handling of categories with low F-scores improves consistency of the corpus and makes it more useful for the investigation of infrequent phenomena.

For this investigation, we do not identify annotations of particular annotators. Instead, we compare two independently annotated versions A and B of our corpus. In total, six annotators are involved in this comparison. Corpus data was incrementally assigned to the annotators in portions of a few hundred sentences. It was (more or less) random whether an annotator worked on version A or B for a particular portion. Therefore, this investigation reports on the overall agreement of annotations, averaging over different “styles” of the annotators, and averaging over annotators that match very well or very poorly.

The annotations of two annotators are not independent of each other since the same tools are used and the same tagger and parser make suggestions that are confirmed or rejected by the annotators. We could not introduce such an independence due to practical limitations. Nevertheless, the annotators were *not* allowed to discuss a portion of sentences before finishing the initial annotations A and B.

### 3. Results

#### 3.1. Part-of-speech Annotations

Results for the agreement of part-of-speech annotations are shown in table 1. Inter-annotator agreement of initial annotations (without cleaning or discussion between the annotators) is 98.57%. Agreement between the initial annotations and the final annotations (after discussions and several iterations of cleaning) is 98.80% for both of the versions.

These agreements are significantly higher than accuracies of current automatic taggers. State-of-the-art result for unseen text in the domain of the NEGRA corpus using the

Table 1: Agreement of part-of-speech annotations between two different annotators, and between the first and the final annotations.

Comparison	total number of tokens	agreement between		
		A and B	FINAL and A	FINAL and B
Part-of-Speech	147,212	145,100 98.57%	145,445 98.80%	145,444 98.80%

Stuttgart-Tübingen-Tagset (Thielen and Schiller, 1995) is 96.7% (Brants, 2000). If we assume that the FINAL version is tagged correctly<sup>2</sup>, this means that a single human annotator reduces the error rate by 64% from 3.3% error rate for automatic tagging to 1.2% error rate for semi-automatic tagging.

Table 2 lists the tags that cause the highest number of inter-annotator disagreements. The table shows the tag, the number of tokens for which both annotators agree on the tag (column “ident”), the number of tokens for which only one of the annotators assigns the listed tag, the other annotator assigns a different tag (column “diff”), the percentage of differences caused by that tag (“%total”) and the F-score of the tag. Note that column “diff” sums up to twice the actual number of differences and that “%total” sums up to 200%. The reason is that each difference involves two tags and therefore is listed in two rows.

The tag involved in the highest number of differences (31.7%) is NN (common noun). All tags in table 2 have relatively high F-scores. This means that the *relative* agreement for the tag is high but it nevertheless causes a large number of differences due to its high frequency.

Table 3 lists the tags with the lowest F-scores. The tag VMPP (past participle of modal verb) is handled worst. Fortunately, its frequency is very low. The absolute frequencies of the other tags in this table are also very low, with the exception of FM (foreign material) which accounts for 5.6% of the differences.

We interpret both tables as follows. Infrequent tags in the NEGRA corpus tend to be handled differently by different annotators (low F-scores). But because of their infrequent occurrence they only cause a small absolute number of difference. On the other hand, frequent tags tend to be handled rather uniformly (high F-scores), but the sheer number of occurrences results in a high absolute number of differences.

Table 4 lists those pairs of tags ( $tag_1, tag_2$ ) with highest confusion rates, i.e., one of the annotators proposed  $tag_1$ , the other annotator  $tag_2$ . The summed frequencies of both tags are given in column “ $f_1 + f_2$ ”. The number of differences is given in column “diff”. The final column (“%total”) shows the fraction of all differences that stem from the

<sup>2</sup>This is an approximation. It is almost impossible to create a large syntactically annotated corpus without errors. But two annotations, comparison, and several iterations of cleaning bring part-of-speech annotations close to this ideal state.

Table 2: Part-of-speech tags which are involved in the highest numbers of differences when comparing annotations A and B. Note that the differences sum up to 200% since each difference involves two tags.

	tag	ident	diff	%total	F-score
1.	NN	31,331	670	31.7	98.9
2.	NE	7,553	580	27.5	96.3
3.	ADV	6,339	317	15.0	97.6
4.	ADJD	2,535	284	13.4	94.7
5.	ADJA	8,501	247	11.7	98.6
	— total —		4,224	200.0	98.6

Table 3: Part-of-speech tags with lowest F-scores when comparing annotations A and B. Note that the differences sum up to 200% since each difference involves two tags.

	tag	ident	diff	%total	F-score
1.	VMPP	1	3	0.1	40.0
2.	ITJ	6	10	0.5	54.6
3.	PTKANT	13	8	0.4	76.5
4.	FM	212	118	5.6	78.2
5.	PTKA	45	25	1.2	78.3
	— total —		4,224	200.0	98.6

Table 4: Pairs of part-of-speech tags with highest confusion rates when comparing annotations A and B.

	$tag_1$	$tag_2$	$f_1 + f_2$	diff	%total
1.	NE	NN	39,503	455	21.5
2.	ADJD	ADV	9,154	105	5.2
3.	ADJA	NN	40,297	74	3.5
4.	FM	NE	8,090	68	3.2
5.	PIAT	PIDAT	972	68	3.2
	— total —			2,112	100.0

confusion of these two tags. The most confusions involve the tags NE (proper name) and NN (common noun). Most proper names and common nouns are tagged identically by two annotators, their F-scores are 96.3% and 98.9%. But due to the high frequency of these two tags (they are assigned to 26.9% of the tokens in the final version) the confusion of these two tags accounts for 21.5% of all differences.

### 3.2. Structural Annotations

Results for the agreement of structural annotations are shown in table 5. It lists unlabeled scores, labeled scores, and labeled scores that also take into account the edge labels going up to the parent nodes. Agreement scores are shown for the two initial versions (A and B) as well as for the initial versions and the final version (FINAL).

The labeled F-score between the two annotations A and B is 92.43%, the labeled F-score between the initial and FINAL versions is significantly higher (around 95%). These results are much higher than for current automatic systems.

Table 5: Agreement of structural annotations between two annotators, and between the first and the final annotations.

	recall	precision	F-Score
<b>A vs. B</b>			
unlabeled	67850 / 72319 (93.82%)	67850 / 72478 (93.61%)	(93.72%)
labeled	66921 / 72319 (92.54%)	66921 / 72478 (92.33%)	(92.43%)
incl. edge labels	64094 / 72319 (88.63%)	64094 / 72478 (88.43%)	(88.53%)
<b>FINAL vs. A</b>			
unlabeled	69646 / 73024 (95.37%)	69646 / 72319 (96.30%)	(95.84%)
labeled	68963 / 73024 (94.44%)	68963 / 72319 (95.36%)	(94.90%)
incl. edge labels	67273 / 73024 (92.12%)	67273 / 72319 (93.02%)	(92.57%)
<b>FINAL vs. B</b>			
unlabeled	69843 / 73024 (95.64%)	69843 / 72478 (96.36%)	(96.00%)
labeled	69183 / 73024 (94.74%)	69183 / 72478 (95.45%)	(95.10%)
incl. edge labels	67477 / 73024 (92.40%)	67477 / 72478 (93.10%)	(92.75%)

Table 6: Phrase types which are involved in the highest number of differences when comparing annotations A and B.

	phrase	ident	diff	F-score
1.	NP	19594	2996	92.9
2.	VP	5623	2294	83.1
3.	PP	17863	1705	95.4
4.	S	13477	1308	95.4
5.	AP	2371	898	84.1

Table 7: Phrase types with lowest F-scores when comparing annotations A and B.

	phrase	ident	diff	F-score
1.	CCP	1	2	50.0
2.	CO	41	64	56.2
3.	ISU	2	3	57.1
4.	DL	95	92	67.4
5.	AA	13	8	76.5

Best results for context-free English structures are around 86% (Ratnaparkhi, 1997), results for German discontinuous annotations are 73% (Plaehn, 2000)<sup>3</sup>. Assuming that FINAL contains correct annotations, this means an error reduction of 64% – 81% by a single semi-automatic annotation pass (if no subsequent comparison and cleaning is applied).

Table 6 lists those phrase types that are involved in the highest number of differences when comparing annotations A and B. It lists the category, the number of identically annotated phrases of this type (“ident”), the number of differently annotated phrases (“diff”), i.e. only one of the annotators proposes a phrase of the particular type, and the corresponding F-score. The category with the highest number of differences is NP (noun phrase). The F-score of NPs is

<sup>3</sup>The F-score reported for German discontinuous phrase structures is obtained for sentences of at most 15 tokens. Results are expected to be lower if longer sentences are taken into account.

above average (92.9 vs. 92.4), but NPs account for 29.2% of all phrases in the final version. This high frequency causes a high absolute number of differences despite the good F-score.

Table 7 shows the phrase types with the lowest F-scores. Worst results are obtained for CCP (coordinated complementizer), but the absolute frequency of this tag is extremely low. As for part-of-speech tags, we find that tags with a high number of differences tend to be frequent and have a high F-score, i.e., they are handled well by the annotators but the high frequency of the category causes a high absolute number of errors. On the other hand, tags with very low F-scores tend to be infrequent. Therefore, their absolute number of differences is low. Nevertheless, cleaning categories with low F-score is very useful if one is interested in investigations on exactly these infrequent categories.

## 4. Conclusions

We presented inter-annotator agreement for part-of-speech and structural syntactic annotations in the NEGRA corpus. Measures that are used to determine the accuracy of automatic tagging and parsing systems can also be applied to semi-automatic annotations. The agreement rates for human annotations are much higher than accuracies of current systems. A single semi-automatic pass reduces the error rate on the part-of-speech and phrase level by 64 – 81% over fully automatic processing. We used two annotations, comparison, discussions of the annotators, and several iterations of cleaning to further reduce the error rate.

Analysis of tags that cause high disagreement rates reveals that categories causing high absolute number of differences do *not* coincide with categories causing high relative numbers of differences (low F-scores). The first tend to have relatively high F-scores, and are thus handled very consistently by the annotators, but their high frequency causes a high number of differences. The latter tend to be infrequent, so even a very low F-score does not result in a high absolute number of differences. We found this effect for part-of-speech tags (at the word level) and for phrase categories (at the structural level).

The next step of analysis, which is beyond the scope of

this paper, is to analyze the differences in more detail, starting with those categories listed in the tables of the results section. What is the exact reason for the high number of differences or the low F-score? Do we need to change or improve the annotation scheme, or do the annotators need more training in order to improve inter-annotator agreement?

## Acknowledgements

The work presented in this paper was carried out in the DFG Sonderforschungsbereich 378, project C3 NEGRA. The annotation is being continued in the DFG project TIGER.

My special thanks go to the annotators of the NEGRA corpus who spent a lot of time in fruitful discussions to find the correct annotations.

## 5. References

- Brants, Thorsten, 1999. Cascaded Markov models. In *Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics EACL-99*. Bergen, Norway.
- Brants, Thorsten, 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing ANLP-2000*. Seattle, WA.
- Brants, Thorsten and Oliver Plaehn, 2000. Interactive corpus annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens, Greece.
- Brants, Thorsten, Wojciech Skut, and Hans Uszkoreit, 1999. Syntactic annotation of a German newspaper corpus. In *Proceedings of the ATALA Treebank Workshop*. Paris, France.
- Calder, Jo, 1997. On aligning trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP-97*. Providence, RI, USA.
- Plaehn, Oliver, 2000. Computing the Most Probable Parse for a Discontinuous Phrase Structure Grammar. In *Proceedings of the 6th International Workshop on Parsing Technologies*. Trento, Italy.
- Ratnaparkhi, Adwait, 1997. A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing EMNLP-97*. Providence, RI.
- Skut, Wojciech, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit, 1997. An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*. Washington, DC.
- Thielen, Christine and Anne Schiller, 1995. Ein kleines und erweitertes Tagset fürs Deutsche. In *Tagungsberichte des Arbeitstreffens Lexikon + Text 17./18. Februar 1994, Schloß Hohentübingen. Lexicographica Series Maior*. Tübingen: Niemeyer.
- Véronis, J., 1998. A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*. Herstmonceux Castle, England.

Voutilainen, Atro, 1999. An experiment on the upper bound of interjudge agreement: the case of tagging. In *Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics EACL-99*. Bergen, Norway.

## Appendix A: Tagsets

This section contains descriptions of tags used in this paper. These are *not* complete lists.

### A.1 Part-of-Speech Tags

We use the Stuttgart-Tübingen-Tagset. The complete set is described in (Thielen and Schiller, 1995).

ADJA	attributive adjective
ADJD	predicatively used adjective
ADV	adverb
APPR	preposition
ART	article
FM	foreign material
ITJ	interjection
NE	proper noun
NN	common noun
PIAT	attributive indefinite pronoun
PIDAT	attr. indef. pronoun with determiner
PROAV	pronominal adverb
PTKANT	answer particle
VAFIN	finite auxiliary
VAINF	infinite auxiliary
VMFIN	finite modal verb
VMPP	past participle of modal verb
VVPP	past participle of main verb

### A.2 Phrase Categories

AA	superlative with <i>am</i>
AP	adjective phrase
CCP	coordinated complementizer
CO	coordination of different categories
DL	discourse level constituent
ISU	idiosyncratic unit
MPN	multi-word proper noun
NP	noun phrase
PP	prepositional phrase
S	sentence
VP	verb phrase

### A.3 Grammatical Functions

AC	adpositional case marker
HD	head
MO	modifier
NK	noun kernel
OA	accusative object
OC	clausal object
SB	subject
SBP	passivized subject