EAGLES

Expert Advisory Group on
Language Engineering Standards

EAGLES

Study of the relation between Tagsets and Taggers

EAGLES Document EAG–CLWG–TAGS/V

Version of May, 1996

Contents

1 Authors

Simone Teufel
Helmut Schmid
Ulrich Heid
Anne Schiller
Institut für maschinelle Sprachverarbeitung
Universität Stuttgart
Azenbergsstrasse 12
HD-70174 Stuttgart
Germany
Tel: +49.711.121.13.73
Fax: +49.711.121.13.66
E-mail: uli@ims.uni-stuttgart.de

2 Introduction

This document describes a few tests which allow to assess some of the complex interdependencies between corpus tagging methods, tagsets and tagging results. In particular, we assess

- the impact of different statistical tagging methods on the results, by comparing the performance of different taggers on the same texts, when using the same tagset;
- the impact of tagset modifications on the results, by using different versions of a tagset on the same texts; differences between the versions of the tagset are documented and classified, and the impact of single modifications is tested;
- the impact of perceived linguistic differences between training texts and test (or: application) texts on the results, by using texts from different text types in training and test, tagsets and taggers being unchanged.

The tests have been carried out with tagsets directly derived from the morphosyntactic specifications for German proposed by EAGLES (see section 2 for details)

The results of these tests may help to give at least partial answers to the following, more general questions:

- Is it possible at all to derive a corpus tagset from a lexical specification of the kind of the EAGLES proposal?
- How well do taggers perform if used with such a lexicon-based tagset? Are there systematic tagging errors (significantly different from those obtained with other tagsets) across tagging methods?
- How do modifications of the tagset influence tagging results, and which types of modifications lead to increases in accuracy? What lessons can we learn for tagset design?
- Which is the impact of text type differences between training and testing? Do we need specific training on, say, fairy tales if we want to tag fairy tale texts, or can we (re-)use a tagger trained on newspaper text? How useful is it to train on "mixed" corpora?[1]

[1]This point is of relevance when it comes to tagging texts for specific purposes, e.g. technical manuals for maintenance, legal texts, etc. Do we have to retrain taggers each time we switch to a new text type?

It should be made clear, at this point, that the few tests performed here are only examples of the type of tests which would have to be carried out for a full assessment. In the analysis of some of the figures obtained in our tests, below, it will become very clear that, for example, a fuller interpretation of the impact of tagset modifications is only possible when comparisons with additional data (e.g. on frequency, on different analyses of a problematic item, etc.) are available. In this sense, the present document is more a case study than a full assessment. However, we hope that some of our test setups and of the practical and methodological considerations in this document are relevant for actions in the field of tagger evaluation in the broad sense[2].

The tests described here are part of a wider effort of validating the outcome of the EAGLES morphosyntax work in practical applications. It is connected with the parallel effort of the ELSNET Reusable Resources Task Group to create small reference text corpora annotated according to guidelines and tagset specifications directly derived from EAGLES work: the texts produced in the framework of this ELSNET/EAGLES collaboration have been used for some of the tests described here. Moreover, some of the questions discussed in this paper have come up in the EAGLES/ELSNET joint work.

This document is structured as follows:

- Section 3 gives an overview of the **background, test setups**, aims and goals of the work described here;
- section 4 briefly describes the **resources and tools** used;
- section 5, contains a **definition of the test setups** used;
- section 6 contains an overview of the **results** obtained on each test and an **interpretation of these results**.
- Suggestions for additional tests, an overview of the primarily used tagset , as well as other background information, can be found in the annexes.

[2] A major effort on the evaluation of annotation tools and techniques is the French project GRACE, supported, among others, ANPELF, and organized by the CNRS.

3 Background and Test Setups

3.1 Background

The EAGLES work group on computational lexicons has produced recommendations for the morphosyntactic classification of word forms. These have been presented in the EAGLES document [Monachini, Calzolari 1994], in the form of an inventory of labels for word forms. From there, formalized specifications have been derived, for French, German and Italian, as well as, slightly less complete, for English[3].

The proposals made in [Monachini, Calzolari 1994] are intended to be applicable to different European languages and to be independent from a particular NLP-application. The language-specific documents contain typed linguistic specifications which themselves are as well not geared towards a specific single application.

However, given that corpus tagging is one of the main applications of the kind of morphosyntactic specifications which the EAGLES lexicon group has produced, evidently corpus tagging has influenced the development: the synopsis step which led to the proposals published in [Monachini, Calzolari 1994] was based on a review of potential predecessors and sources of input which included tagsets for corpus annotation; and the linguistic specifications for French, Italian and German (referred to, in the following, by ELM-FR, ELM-IT and ELM-DE), respectively[4], are influenced to some extent by experience from tagset development. Moreover, the constant interaction with the EAGLES Work Group on the linguistic annotation of text corpora has of course contributed to this orientation. This interaction has helped shape the language-specific ELM-incarnations very much[5].

The work described in this report shows that the EAGLES-based ELM-DE specifications indeed allow to derive a tagset which can be practically used for the tagging of German and which leads to acceptable results. Moreover, we could follow part of the history of the tagset and evaluate the impact of the modifications introduced.

ELM-DE has two sources: the results of [Monachini, Calzolari 1994], and a corpus tagset jointly developed by the universities of Tübingen[6] and Stuttgart, since 1993/94. This tagset, now referred to as STTS (**S**tuttgart-**T**übingen **T**ag **S**et), see [Schiller et al 1995] has mainly been used in the tests described in this paper. It is mappable with ELM-DE, and automatic tools exist which perform this mapping. Where we have analyzed the impact of tagset changes on the tagging results, the modifications tested were influenced by differences between STTS and its predecessor, the IMS-TUE tagset (also jointly developed by Stuttgart and Tübingen).

The work reported here was carried out at Rank Xerox Research Centre (RXRC), Grenoble, France, and at the Institut für maschinelle Sprachverarbeitung, Computerlinguistik, of Universität Stuttgart (STR). The text material and tagsets all concern the German language.

[3] Cf. [von Rekowski 1995] for French, [Teufel 1995a] for German, [Monachini 1995] for Italian, [Teufel 1995b] for English.

[4] ELM stands for EAGLES/Lexicon/Morphosyntax.

[5] We would like to thank in particular Prof. Geoffrey Leech (Lancaster) and his team for much help and feedback.

[6] Particular thanks go to Christine Thielen and Helmut Feldweg, of SNS, Universität Tübingen, and to the ELWIS project at SNS, for many fruitful discussions and much practical feedback on the versions of the tagset.

3.2 Test setups and objectives

We suggest three types of tests:

- "Tagger Evaluation":
  tests allowing to assess the impact of different statistical tagging methods on the results, by comparing the performance of different taggers on the same training and test data, using the same tagset;
- "Tagset Evaluation":
  tests allowing to assess the impact of tagset modifications on the results, by using different versions of a given tagset on the same texts; differences between the versions of the tagset are documented and classified, and the impact of each modification is tested;
- "Text type Evaluation":
  tests allowing to assess the impact of perceived linguistic differences between training texts and test (or: application) texts on the results, by using texts from different text types in training and testing, tagsets and taggers being unchanged otherwise.

The test setups and the objectives underlying each test are described in more detail in the following.

We suggest three series of tests:

3.2.1 Tagger Evaluation

Different taggers, one tagset: test the impact of different tagging methods on the results by comparing the results of different taggers:

- tagging errors shared by the different methods;
- differences in the tagging results.

3.2.2 Tagset Evaluation

One tagger, different versions of tagsets: test the impact of modifications of the tagset on tagging results:

- find out about useful modifications of tagsets: how to modify your tagset to achieve better tagging results;
- derive from the analysis of the results some guidelines for possible "hard" cases in tagset design ("what should be avoided in tagset design to get good tagging results").

3.2.3 Text Type Evaluation

Differences between training texts and test texts:

- it is assumed that certain text types (e.g. news vs. laws vs. technical prose) differ with respect to the distribution of syntactic constructions;
- these text-typical differences may lead to (slightly?) different statistical models, for different texts.

The objective is to find ways of assessing the impact of such differences, when they occur between training and test texts.

4 Resources and tools

The experiments have been carried out only on the basis of tools and resources for German. Even though the specific results for individual lexical classes are language-specific, the general results should be applicable to other languages as well.

Given that the work is based on the ELM-DE specification and the STTS tagset, derived from it, EAGLES compatibility of the whole test setup is fully guaranteed.

4.1 Corpora

In the following, we briefly describe the textual corpora used at STR and RXRC for the work described here. The STR corpora have been used for tagger and tagset evaluation, the RXRC corpora for text type evaluation.

- **STR**: manually tagged corpus of 75,000 tokens
  - 50,000 from *Frankfurter Rundschau*
    This corpus is the ELSNET/EAGLES DE reference corpus, and is composed of news stories about economy (12,500), politics (12,500), culture (12,500), local events (6,250) and sports (6,250).
  - 25,000 from *Stuttgarter Zeitung*
    This corpus is STR internal.
- **RXRC**: automatically tagged and manually corrected corpus of 90,000 tokens.
  - 20,000 reference corpus from SfS Tübingen
  - 30,000 news (*Spiegel, Welt, taz, spektrum*)
  - 26,000 fairy tales (*Andersen, Bechstein, Grimm*)
  - 14,000 other (reports, guidelines, ...)

4.2 German Tagsets

The point of reference for all tests described in this document is the German specification of the EAGLES morphosyntax encoding proposal, ELM-DE. The following needs to be taken into account when interpreting the test setups and the results.

- The tests take into account the technical state of the art of automatic tagging and thus concentrate on the part of speech annotation proposed in ELM-DE, leaving aside the more detailed and more complex set of morphosyntactic features. Thus, only the higher levels of the ELM-DE specification have been used here. This part corresponds to the STTS tagset for German, which distinguishes 54 part of speech tags (cf. Appendix C).
- The tagset used at RXRC is a derivative of STTS, but it omits the distinction between proper names and common nouns, and it also contains a few modifications, with respect to STTS, as far as the classification of function words is concerned. The latter differences are not particularly relevant for the present tests, and the former one will be kept track of.

- The 1995 version of STTS had a predecessor, the IMS/TUE-Tagset, of 1993. STTS is mappable to ELM-DE (it only has a few finer grained distinctions in the domain of indefinite pronouns). In the redesigned phase which led from IMS/TUE to STTS, a number of interesting questions came up which concern tagset design and its impact on tagging results. Some of these questions have been analyzed in more detail in the experiments we will report on in this document, the others have been collected in Appendix B. Thus the tagset evaluation is based on concrete questions which came up during tagset design and not just on arbitrary or otherwise motivated questions. A detailed definition of the modifications tested in the tagset evaluation work reported here is given in Appendix A.

4.3 Taggers

In the following, we give a short description of the taggers which are available for tests at STR and RXRC.

4.3.1 Taggers used for evaluation

- Xerox HMM tagger

  The Xerox tagger is based on *Hidden Markov Models* (HMM). It applies the *Baum-Welch* (or *Forward-Backward*) algorithm for training and the *Viterbi* algorithm for tagging. The tagger needs as input a finite-state *lexicon* which contains surface word forms and their corresponding tags and a finite-state *guesser* for word forms which are not in the lexicon.

  The Xerox HMM Tagger ([Cutting et al. 1992]) was designed to be trained on untagged corpora, but a more recent version ([Wilkens, Kupiec 1995]) allows initialization with untagged and/or tagged data. However, when enough tagged training material is available, it turned out (cf. [Schiller 1996]) that the best tagging results were obtained when the tagger was just initialized on a tagged corpus and no further training was added.

- TreeTagger: Decision-tree-based [Schmid, Kempe 1994]
  The TreeTagger ([Schmid 1994b], [Schmid 1995a]) is also based on a Hidden Markov Model. Only tagged data are used for training, however (supervised training). To avoid problems with sparse training data, the TreeTagger uses a decision tree to reduce the number of contextual parameters until reliable estimation of all contextual parameters is possible. Tag for unknown word forms are guessed with a suffix lexicon (see also [Cutting et al. 1992]) which is built automatically. Lexical probabilities are smoothed using ambiguity class based probabilities in order to improve tagging of rare word forms.

For our experiments we chose among the available taggers those which can easily be adapted to a specific language or tagset. This allows to use a uniform lexicon and tagset for all different methods and compare the results (cf. 5.1), and it is necessary for tests on tagset variations (cf. 5.3).

4.3.2 Other taggers

The following tagger programs were available at STR and RXRC for a further extension of the tests described. The reasons why they were not used for the actual experiments are given below.

- Bigram/Trigram Tagger

  The bigram/trigram tagger by André KEMPE ([Kempe 1994]) is available for English, French and German. It uses a (modified) version of the *Viterbi* algorithm and is based on HMM modelling, i.e. the method for tagging is basically the same as the one used by the Xerox HMM tagger. Thus, we don't expect important differences for the results.

- NetTagger

  The NetTagger ([Schmid 1994a]) uses a recursive single-layer feed-forward neural network for part-of-speech disambiguation. The network input consists of vectors of lexical tag probabilities for the words in the context of the current word (about three preceding and two following words). The output of the network is a also a probability vector[7] which is fed back into the network as the probability vector of the preceding word. The tagger is trained on tagged data using a modification of the *Backpropagation algorithm.*. Like the TreeTagger, the NetTagger uses a suffix lexicon to guess the tag probabilities of an unknown word form.

- BRILL's tagger

  The rule based tagger by Eric BRILL ([Brill 1992], [Brill 1994]) is trained in two steps. In a first phase the tagger assigns each word its most likely tag, estimated by examining a large tagged corpus. The second consists of improving the results using contextual information. Errors in tagging are recorded as triples $(tag_a, tag_b, number)$ and for each triple the contextual rule is determined which results in the greatest error reduction. For each rule the net improvement is calculated by substracting the number of errors *introduced* by this rule from the number of *corrected* errors. The rule which results in the greates improvement is added to the rule set. The rule acquisition procedure iteratively continues.

  The Brill tagger is available for English and the part-of-speech categories of the Penn TreeBank[8].

- RXRC finite-state tagger

  At RXRC, Jean-Pierre CHANOD has worked on a rule-based tagger for French ([Chanod, Tapanainen 1995]) which uses finite-state techniques for disambiguation.

  A set of finite-state rules describe permissible patterns of tags over a sentence. Composing these constraints over a morphologically analyzed input sentence eliminates most spurious tags. In the tagger, the rules are represented as finite-state transducers that

[7] The elements of the output vector are better named scores since they are not probabilities in the strict mathematical sense. E.g. they do normally not sum up to 1.

[8] It would have been necessary to adapt it for German and to the actual tagset in order to include it in our experiments. But this turned out to represent more work than what was possible in the given time frame.

are composed one after another with an ambiguous sentence. After all the transducers have been applied, the sentence has only one analysis.

The RXRC finite-state tagger is only available for French. As we restricted our experiments to German (and a tagset conforming to the EAGLES standards), we excluded this tagger from the tests. A detailed comparison of the Xerox HMM tagger and the finite state tagger for French can be found in [Chanod, Tapanainen 1994].

5 Proposals for practical experiments

In general, taggers are evaluated by computing the amount of tagging errors, for example by comparing an automatically tagged corpus to a manually tagged corpus. This error rate, however, does not only depend on the tagging method or tagset, but also on the ambiguity rate and accuracy of the lexicon that is used. Test results also may vary for different types and sizes of the test corpus.

5.1 Methodology: Statistics underlying the assessment

Tagger accuracy is, among others, influenced by parameters such as lexicon accuracy and lexical ambiguity of words. We need to have a uniform test environment for the different tagging methods; therefore, to cope with the lexicon problems, we use identical (or at least comparable) lexicons for all taggers. Comparable test conditions may be set up by using the same lexicon (apart from differences of the internal format), the same tagset and the same test corpora for different taggers.

The taggers to be evaluated are all trained on the same manually tagged corpus. The list of annotated word forms used as "tagger lexicon" for each tagger, is built from a list of word/tag pairs, which is derived from the word forms contained in the training and in the test corpus. This full form list is computed by means of a morphological analyser ([Schiller 1995]) and of mapping rules to transform the morphological categories into the appropriate tags. These mapping rules may remove or add readings for specific word forms, according to the guidelines for manual tagging ([Schiller et al 1995]); however, in the process of word list creation, no corpus-dependent modifications of the lexicon are introduced:

- The full form list is not restricted to readings which appear in the corpus, i.e. for each word the whole range of possible lexical tags is provided, not only those which occur in the actual training corpus.
- Forms which cannot be morphologically analyzed, such as proper names, foreign language material and other "gaps", are preserved, i.e. the lexicons for all taggers contain the same information, but we do not ensure full coverage of the texts.

The following definitions will be used for statistics concerning ambiguity of word forms:

- *Ambiguity class*
  = set of tags associated with a word form.
- *Ambiguity type*
  = number of tags associated with a word form.
- *Ambiguity rate*
  = average number of tags per word form in a given corpus.

We will divide the evaluation statistics for a given test corpus into two parts: corpus statistics and error statistics.

5.1.1 Corpus Statistics

The corpus statistics depends only on the test corpus and on the lexicon, and is therefore the same for all taggers, in our tests.

The lexicon, however, may be considered part of the tagger method, e.g. if lexicon *modifications* (such as omitting "marginal" readings of word forms) are part of the disambiguation method. In that case, the lexicon-dependent corpus statistics may differ even if the tested taggers rely on identical training (and testing) lexicons. Also, statistics concerning unknown word forms, depend mainly on the tagger method and are thus not identical for the tested tagger methods.

In general, all the figures given below do not depend on the disambiguation result, i.e. they can be computed *before* automatic tagging is applied, by associating the manually tagged test corpus with the lexical information and the tags for non-lexicalised word forms.

- Number of tokens in test corpus;
- Number of tags in (manually tagged) test corpus;
- Number of *lexicon gaps* (= not-lexicalized word forms);
- Number of *lexical errors* or defective word forms, i.e. word forms for which the lexicon (or the guesser method) does not include the tag which is given in the manually tagged text;
- Number of *ambiguity classes* in the corpus;
- *Ambiguity rate* of the corpus word forms.

5.1.2 Error Statistics

The error statistics are computed by comparing the tagger output with the manual tagging.

- Overall accuracy;
- Error classification by *ambiguity type*:
  = list of disambiguation accuracy for 1, 2, 3, ... tags per word form;
- Error classification by word forms:
  = list of most frequent word forms tagged erroneously;
- Error classification by tags:
  = list of most frequently confused tags, i.e. pairs of differing manually and automatically associated tags.

Example:

line	word form	manual tag	tagger tag	lex	lexical tags
1	Ilka	NE	NE	-	ADJD ADV NE NN
2	Sperber	NE	$\boxed{\text{NN}}$	+	NN
3	war	VAFIN	VAFIN	+	VAFIN
4	schon	ADV	ADV	+	ADV VVIMP
5	bei	APPR	APPR	+	APPR PTKVZ
6	den	ART	ART	+	ART PDS PRELS
7	meisten	PIDAT	PIDAT	+	PIDAT PIS
8	Stars	NN	NN	+	NN
9	zu	APPR	APPR	+	ADV APPR PTKA PTKVZ PTKZU
10	Gast	NN	NN	+	NN VVFIN VVIMP
11	,	$,	$,	+	$,
2	die	PRELS	PRELS	+	ART PDS PRELS
13	in	APPR	APPR	+	APPR
14	Berlin	NE	NE	+	NE
15	wohnen	VVFIN	$\boxed{\text{VVINF}}$	+	VVFIN VVINF
16	.	$.	$.	+	$.

For the above sample test corpus we get the following numbers:

Number of tokens:	16	
Number of tags:	11	{ $, $. ADV APPR ART NE NN PIDAT PRELS VAFIN VVFIN }
Lexicon gaps:	1	*"Ilka"*
Lexical errors:	1	*"Sperber"*
Ambiguity classes	13	{APPR}, {NE}, {NN}, {VAFIN}, {$,}, {$.}, {ADV VVIMP}, {APPR PTKVZ}, {PIDAT PIS}, {VVFIN VVINF}, {ART PDS PRELS}, {ADJD ADV NN NE} {ADV APPR PTKA PTKVZ PTKZU}
Ambiguity rate:	2.06	(= 23/16)

The following table shows the error classification by *ambiguity type* for the above sample text.

ambiguity type	number of tokens	tagger errors	accuracy
1 tag	7	1	85.7 %
2 tags	4	1	75.0 %
3 tags	3	0	100.0 %
4 tags	1	0	100.0 %
5 tags	1	0	100.0 %
overall accuracy	16	2	87.5 %

5.2 Tagger Evaluation

The tests proposed for tagger evaluation are based on a comparison of the output produced by different taggers on the same text. The lexicon statistics will be identical for all taggers, and the comparison will have to concentrate on the tagger statistics.

The following parameters will be identified, compared and the results interpreted:

- the overall error rate for each tagger,
- the tags which are most problematic for each individual tagger; this can be identified by a frequency count of tagging errors and a classification by tag;
- a comparison, across taggers, of tag confusion pairs, with the objective of identifying certain tags or tag confusion pairs which are problematic for all taggers tested; if indeed certain tag confusion pairs turned out to be very frequent, irrespective of tagging method, this result should have an impact on tagset design.

5.3 Tagset evaluation

The purpose of the tagset evaluation test is to analyze the impact of tagset variants on tagging results. Examples of such tagset variants are:

- changes in the granularity of the tagset;
- possible overlaps in tag denotation;
- changes in tag definitions.

More details on these types of modifications are given below, in section 5.3.1.

Such changes typically appear when a tagset is developed over a certain time frame, and when "tuning work" is carried out, to improve the tagging results. An example is the development of STTS out of IMS/TUE and the development of the current RXRC tagset, as well out of IMS/TUE.

5.3.1 Relevant tagset modifications

Below, we describe the types of tagset modifications we will test in the practical part of the tagset evaluation. Some of the changes were motivated by linguistic reasons, others by practical reasons (tagging accuracy). Many of the changes included changes to ambiguity classes for a word form or a group of word forms. A list of phenomena is given below, in section 5.3.2, and more details (including a description of the "old" vs. "new" practice) can be found in appendix A.

The different kinds of tagset modifications are defined as follows:

Type I: Granularity Changes

Type Ia: Simple granularity changes: Two or more tags for ambiguous word forms are merged, or an ambiguous tag is split into one or more unambiguous tags. As mentioned above, a splitting is only advisable if the word forms have a different distribution. Total tagging accuracy is regarded after the change.

Type Ib: More complex granularity changes, overlapping tags. Two ambiguous tags, A and B, are sometimes split into three classes, where the new, third class has overlapping ambiguities with either of the original classes. Such is the case with personal and reflexive pronouns. The constellations for tests in this case are more numerous:

1. merge the forms into a single ambiguous tag (AB);
2. split the forms, such that there is a single tag for the two unambiguous forms, and a tag for all ambiguous forms (A, B, AB);
3. merge the ambiguous forms with either of the unambiguous tags (2 possibilities: *"A, AB"* and *"AB, B"*).

Type Ic: External impact of granularity changes: Granularity changes or definition changes not only affect the immediately involved (internal) ambiguity classes, but may also have an influence on other ambiguous classes. It might thus be worth splitting a non-ambiguous class, if a positive impact of this splitting on the disambiguation of other classes can be expected.

Type II: Changes in the assignment of wordforms to tags. The borderline between two tags is shifted (in guidelines and in the annotation of the reference text) but tags included in these changes are otherwise still the same. An example for German: it is very difficult to distinguish lexicalized participles used adverbially (ADJD) from participles used within the verbal complex (VVPP). We tried several different definitions, depending on the syntactic constructions the ambiguous wordforms appear in.

5.3.2 Overview of phenomena and proposed tests

Tests run in this exercise (for details, see appendix A)

- **Disambiguation of proper names and common nouns.** (NN/NE)
  Proper names and common nouns have very similar distribution, their distinction is mainly semantic. Furthermore in German (unlike in other European languages), there is no orthographical distinction between proper and common nouns.

  There is also interaction/ambiguity with foreign language material.
  (Type Ia)

- **Separable verb prefixes** (PTKVZ)
  Verbal prefixes may occur as separated units and are then ambiguous with ADV, ADJD, APPR and APPO. We are most interested in tests with adjectival and adverbial prefixes as they have a similar distribution as predicative adjectives or adverbs.
  (Type II, Ic?)

- **Verb infinitives** (VINF)
  All infinitives are ambiguous with finite 1st and 3rd Plural verbs. Sequences of infinitives

are only possible with auxiliaries or modals. Therefore we expect external impact from verb type disambiguation (lexical distinction VM, VA, VV).
Modal auxiliaries: different distribution (perfect constructions with infinitive forms).
(Type Ia, Ic?)

- **Disambiguation of reflexive and personal pronous** (PRF/PPER)
  Ambiguous forms are *mich, dich, dir, mir, euch, uns.* reflexive and personal pronouns have identical distribution; the decision about reflectiveness depends on (verbal) subcategorization, but automatic taggers have no access to criteria used by humans (reflexitivity of verb; case, person, number of subject and pronoun). Are unambiguous 3rd person personal pronouns (er, sie, es) and reflexives (sich) enough to train taggers to distinguish ambiguous forms (mich, dich, ...)?
  (Type Ib)

Phenomena suggested for further testing (cf. appendix B):

- **Functional vs. adverbially used participles** (VVPP/ADJD)
  Difficulty: copula constructions and stative passive have identical formation. Stepwise change to definitions in guidelines, by syntactic constructions.
  (Type II)
- **Indefinite vs. adjective-like determiners** (PD-/PI-/ADJ-)
  Merging of demonstrative and indefinite pronouns/determiners? Membership to demonstr./indef. is lexically determined!
  (Type Ic?)
- **Indefinite determiners with or without articles** (PIDAT/PIAT)
  Context dependent or lemma dependent?
  (Type Ia, Ic?)
- **adverbially used adjectives vs. non-adjectival adverbs** ADJD/ADV
  Differences in semantics.
  (Type II)
- **wordforms** *doch, aber, denn* (KON/ADV)
  This word forms are ambiguous between coordinating conjunctions and adverbs, depending on position within sentence
  (Type Ib)
- **wordforms** *wie, als* (KOUS/PWAV/KOKOM/KON)
  These word forms have a variety of functions and therefore tags. Which functions can be merged into one tag easiest? *(Type I, Ic?)*
- **proper names vs. foreign material** (FM/NE)
  Very problematic: dependent on semantics. Interaction with NE/NN test.
  (Type I)

- **Punctuation $(, $., $,**
  Is there external impact of behavour of comma on relative pronouns, articles, subordinating conjunctions?
  (Type Ic?)

5.4 Text type evaluation

Probabilistic taggers rely basically on similarities between training and application data.

It seems obvious that a tagger which is trained on a "standard" corpus will be less accurate on a text with highly irregular sentence structures (for example technical maintenance manuals, etc.[9]). The question is, however, how big the structural (syntactic) differences are in the corpora used in "real-life applications", and how much these differences really influence the tagger accuracy.

We cannot provide a well defined measurement of text type difference and thus the experiments are based on a subjective choice of texts which we assume are different with respect to their syntactic structure.

The experiment consists in producing several "specialised" taggers, i.e. each tagger is trained on one specific text type only. The resulting taggers are then evaluated with test corpora which also correspond to different text types.

- Test:
  - Training on a given text type
  - Run the trained tagger on this same text type;
  - Run the trained tagger on another text type
- Second test setup:
  - Training on a mixture of two (supposedly quite different) text types;
  - Test on only one of the two text types taken in isolation (do so for either text type). Compare results.

[9]Future tests should be devoted to testing very "deviant" material of this type; in the present exercise, we restricted ourselves to testing fairy tales vs. newspaper texts

6 Results

All tests for tagger evaluation (cf. section 5.2) and for tagset evaluation (cf. section 5.3) are based on the same training, test corpora and word lists provided by IMS. Two word lists are used to build the training and testing lexicons:

1. The *regular lexicon* contains only those tags which result from a morphological analyser and mapping rules.
2. The *augmented lexicon* contains all entries of the regular lexicon, but also all word forms from the training corpus with the corresponding manual tags.

For each test setup of tagset experiments (cf. section 5.3), we modified the training and testing corpora with respect to the tags to be evaluated and adapted the associated lexicons as well.

The tests on text type evaluation (see section 6.3) were run on the RXRC corpus collection, which covers different types of texts, and uses a tagset which is very close to STTS.

6.1 Tagger evaluation

As a base line, we tested two different taggers (the *TreeTagger* at STR, and the *HMM-Tagger* at RXRC, see section 4.3.1) on the standard STTS tagset as documented in [Teufel 1995a] and [Schiller et al 1995]. This test run serves as a comparative value for all subsequent tagset evaluation tests in section 6.2.

For every test, the following types of tables are given:

1. *Corpus statistics* (cf. section 5.1.1)

   The corpus statistics table shows numbers of tokens, tags and ambiguity classes in the training and test corpus, which depend on the lexicon and on the guesser or suffix lexicon.

2. *Error statistics* (cf. section 5.1.2)

   - *Accuracy by ambiguity type*
     The first table contains the number of errors in the test corpus, classified by the number of ambiguous tags and the type of errors (LE = lexical error; DE = disambiguation error).
   - *Most frequent errors*
     There are two tables for the most frequent errors: errors classified by wrongly tagged word forms, and errors classified by tags (across word forms). In both cases, we compare the tags asigned manually with the result of automatic taggers..

6.1.1 TreeTagger: Standard test

For the tests with the TreeTagger the *extended lexicon* was applied. Therefore the training corpus does not contain any non-lexicalized word form.

Corpus statistics	training	test
Tokens	62860	13416
Tags	51	46
Lexicon gaps	0	240
Lexical errors	0	62
Ambiguity classes		
Ambiguity rate		1.58

Error statistics								
ambiguity	tokens	in %	correct	in %	LE	in %	DE	in %
1	8138	60.6	8088	99.4	50	0.6	–	–
2	3275	24.4	3084	94.2	7	0.2	184	5.6
3	1533	11.4	1432	93.4	2	0.1	99	6.5
4	424	3.2	376	88.7	1	0.2	47	11.1
5	14	0.1	9	64.3	0	–	5	35.7
6	9	0.1	6	66.7	0	–	3	33.3
7	15	0.1	9	60.0	2	13.3	4	26.7
8	6	0.0	4	66.7	0	–	2	33.3
9	0	–	0	–	0	–	0	–
10	2	0.0	0	–	0	–	2	100.0
total	13416	100.0	13008	97.0	62	0.5	346	2.6

Most frequent errors (by word form)			
number	word	correct tag	tagger tag
13	um	KOUI	APPR
9	werden	VAINF	VAFIN
9	Osthold	NE	ADJD
8	Aber	KON	ADV
8	der	PRELS	ART
8	das	PDS	ART
7	dem	PRELS	ART
6	Stänner	NE	NN
6	Brück	NE	NN
5	die	PRELS	ART

Most frequent errors (by tags)		
number	correct tag	tagger tag
112	NE	NN
24	VVINF	VVFIN
21	VVFIN	VVINF
21	PRELS	ART
18	KON	ADV
17	KOUI	APPR
12	VAINF	VAFIN
11	VVFIN	VVPP
11	NN	NE
11	NE	ADJD

6.1.2 Xerox HMM-tagger: Standard test

To compare the effects of using different lexicons for training and testing, we chose two test setups for the Xerox HMM-tagger: one with the *regular lexcion* only, and a second with the *extended lexicon*, which is the same as for the test of of the TreeTagger.

(1) Regular training lexicon

Corpus statistics	training	test
Tokens	62860	13416
Tags	51	46
Lexicon gaps	1756	283
Lexical Errors	543	65
Ambiguity classes	263	196
Ambiguity rate	1.69	1.67

Error statistics								
ambiguity	tokens	in %	correct	in %	LE	in %	DE	in %
1	7978	59.5	7942	99.6	36	0.5	–	–
2	2663	19.9	2482	93.2	13	0.5	168	6.3
3	2078	15.5	2014	96.9	8	0.4	56	2.7
4	589	4.4	518	88.0	7	1.2	64	10.9
5	81	0.6	71	87.9	1	1.2	9	11.1
6	19	0.1	16	84.2	0	–	3	15.8
7	8	0.1	7	87.5	0	–	1	12.5
total	13416	100.0	13050	97.3	65	0.5	301	2.2

Most frequent errors (by word form)			
number	word	correct tag	tagger tag
13	DM	NN	NE
9	Osthold	NE	ADJD
6	das	PDS	ART
5	werden	VAFIN	VAINF
5	Reich	NE	NN
4	haben	VAFIN	VAINF
4	dem	PRELS	ART
4	Um	KOUI	APPR
4	Deutschland	NE	NN

Most frequent errors (by tags)		
number	correct tag	tagger tag
55	NN	NE
39	NE	NN
28	VVFIN	VVINF
16	VVFIN	VVPP
12	KON	ADV
11	ADJD	VVPP
11	ADJD	ADV
10	VVINF	VVFIN
10	NE	ADJD

(2) Extended training lexicon

Even though the Xerox HMM tagger is based on the same basic lexicon as the TreeTagger (in 6.1.1) the ambiguity class of the word forms is not always the same (as one would have expected).

This difference is due to the fact that the lexicon which is used *internally* by the TreeTagger omits marginal (i.e. very rare) readings of word forms and thus reduces the ambiguity classes of lexicalized word forms. For non-lexicalized word forms, however, the TreeTagger uses ambiguity classes containing up to 10 tags, whereas the largest ambiguity class of the tested Xerox HMM tagger is 7.

Corpus statistics	training	test
Tokens	62860	13416
Tags	51	46
Lexicon gaps	0	241
Lexical errors	0	49
Ambiguity classes	275	205
Ambiguity rate	1.64	1.69

Error statistics								
ambiguity	tokens	in %	correct	in %	LE	in %	DE	in %
1	7962	59.4	7932	99.6	30	0.4	–	–
2	2660	19.8	2481	93.3	11	0.4	168	6.3
3	2082	15.5	2019	97.0	5	0.2	58	2.8
4	480	3.6	399	83.1	3	0.6	78	16.3
5	178	1.3	161	90.5	0	–	17	9.5
6	47	0.4	38	80.9	0	–	9	19.1
7	7	0.1	5	71.4	0	–	2	28.6
total	13416	100.0	13035	97.2	49	0.4	332	2.5

Most frequent errors (by word form)			
number	word	correct tag	tagger tag
13	DM	NN	NE
9	Osthold	NE	ADJD
6	das	PDS	ART
5	werden	VAFIN	VAINF
5	Reich	NE	NN
4	haben	VAFIN	VAINF
4	dem	PRELS	ART
4	Um	KOUI	APPR
4	Deutschland	NE	NN
4	Asher	NE	ADJA

Most frequent errors (by tags)		
number	correct tag	tagger tag
42	NE	NN
41	NN	NE
28	VVFIN	VVINF
20	NE	ADV
18	NE	ADJD
15	VVFIN	VVPP
12	KON	ADV
12	ADJD	ADV
11	ADJD	VVPP

6.1.3 Interpretation

The following are a few hints and directions for the interpretation of the results produced by the different taggers.

- The percentages of lexical errors and disambiguation errors have been kept separate and should be recorded in relation to each other. Ideally, the best result which could be achieved by a tagger would be a rate of 0% of disambiguation errors. It could then still have a small overall error rate which would then be due to lexical errors. For example, if there is a lexical error rate of 0.5%, the best result which could be achieved is 99.5%.

  When comparing the Tree Tagger and the Xerox HMM Tagger, a number of overlaps in errors can be found. In particular, the distinction between NN and NE, i.e. between common nouns and proper names, is problematic for both types of taggers. Even at the level of individual word forms, there are overlaps: *Osthold, das, werden* are highly ranked in the error frequency list of both tools.

- Similarly, differences between finite and infinite verb forms, and between participles and finite verb forms are hard to identify, since indeed many such forms are ambiguous (cf. DE *er wird verraten*). Against this background, a relatively high ranking of disambiguation errors in this area is expected.

- In the ambiguity cases, where proper names are involved, there was a marked difference between the two taggers. The Tree Tagger favours NN over NE, whereas the Xerox HMM Tagger seems to equally distribute the two tags. A further analysis revealed that the TreeTagger in general tends to prefer more frequent tags. This could be traced to a difference between the tagging formula used in the TreeTagger and the standard HMM formula which was previously considered irrelevant. The formula used in the TreeTagger is equivalent to multiplying the HMM formula with the a priori probability of the tag. This explains the preference for more frequent tags. This analysis was confirmed by an experiment in which the TreeTagger was used with the standard HMM formula (everything else unchanged). The modification improved tagging accuracy by 0.5 % [10].

- The two taggers differ with respect to the amount of errors produced on items such as *DM, um, Aber* (see ranking in word form error frequency list). Again, it is quite hard to interpret the data in detail. To be able to identify reasons for the differences in the analysis of say *um*, also the overall amount (absolute number of occurrences) of *um*, of the two tags in question, KOUI and APPR would be needed, as well as the absolute number of cases of ambiguity (cf. the discussion above about verb forms). A full noise/silence-test would thus be needed.

[10] Similar improvements have been obtained with English data suggesting that it is not a language specific effect.

6.2 Tagset evaluation

6.2.1 Purpose

We assume that corpus annotation should provide as much linguistic information *as possible*, i.e. tags should correspond to a (sub)set of features which are used in further language processing, such as syntactic parsing, semantic analysis, etc.

However, there is a conflict between what linguists or "higher processing steps" need and what taggers can technically provide; this conflict must be taken into account when designing a tagset. The following section shows some of the theoretical questions involved in the design phase, some more practical test cases and a description of the phenomena which we chose for our tests.

6.2.2 Questions of relevance for tagset design, addressed in the tests

1. Which categories from "traditional" grammar, semantics, lexicon design can be used in tagsets? How much "linguistics" can there be in tagsets?

2. Which features are error-prone in the tagging of a given language, i.e. which distinctions affect the overall error rate of the taggers?

   Example: `GENDER` is not relevant for French, i.e. the results are the same, whether `GENDER` is annotated or not (cf. [Elworthy 1995] and [Chanod, Tapanainen 1995]). This question is especially interesting for applications of ELM.

3. Which lexical ambiguities can be solved by means of distributional classification?

   - within the ambiguity class
     - Example for feasible disambiguation: Imperatives (a, e) occur only in sentence initial position, whereas indicative (b, c, d) forms can also occur in second or final position.

       (a) *kauft/V_IMP das Kleid !*

       (b) *sie kauft/V_IND das Kleid !*

       (c) *weil sie das Kleid kauft/V_IND*

       (d) *kauft/V_IMP ihr das Kleid !*

       (e) *kauft/V_IND ihr das Kleid ?*

       As the distribution is different for the ambiguous forms, we may expect that taggers will disambiguate correctly on grounds of distribution. It is appropriate to use two *non-ambigous* tags.
     - Example for non-feasible disambiguation: German adjectives in predicative (a,d) or adverbial (b,c) use:

       (a) *Er ist schnell/ADJ_PRED*

       (b) *Er fährt schnell/ADJ_ADV*

       (c) *Er macht seine Aufgaben schnell/ADJ_ADV*

       (d) *Er macht das Auto schnell/ADJ_PRED*

As the context is similar for adverbial and predicative adjectives, it seems appropriate to use an *ambigous* tag which covers both forms.

- outside the ambiguity class: Sometimes it is possible to increase the overall tagger rate by changes to a class of tags, even though it is not the disambiguation within this class of tags itself that is influenced, but the marking of a different distributional context for other ambiguous word forms in the context.
  We call this phenomenon *contextual disambiguation power* or *external impact.*

  Example: A subclassification of verb forms into modals and "full" verbs does not modify the type of ambiguity of the verb forms themselves (e.g. because the membership to the class is *lexically* determined[11]), but modal verbs behave differently from content verbs (they can form passive constructions with an infinitive instead of a participle); thus they can help mark different contexts and restrict ambiguities of the context words. The ambiguity class INF??FIN (infinitives are inherently ambiguous with 1stpl or 3rdpl finite verbs) can be solved in sentence (d) (with different tags for modal and content verbs) but not in sentence (b) (with a joint class for all verbs).

  (a) *sie wollen/V_INF??V_FIN gehen/V_INF??V_FIN*
  (b) *weil sie gehen/V_INF??V_FIN wollen/V_INF??V_FIN*
  (c) *sie wollen/VM_INF??VM_FIN gehen/VV_INF*
  (d) *weil sie gehen/VV_INF wollen/VM_INF??VM_FIN*

6.2.3 Practical setups

The following tests imply changes of the tagset. For each modification we will compare the results obtained with the Xerox HMM tagger to the results of the *standard* test (section 6.1.2, using a *regular lexicon*).

For every test we will present some relevant statistics from the standard test (cf. section 6.1) compared to the results obtained with the modified tagset, especially for those tags which are concerned by the tagset modifications.

[11] As there are only 6 modal verb lemmata which are easy to distinguish lexically.

6.2.4 Xerox HMM-tagger: Noun test

The most frequent error in the *standard* test concerned common nouns (NN) and proper names (NE).

The table below shows some statistics about the frequency of NN and NE in the standard training and test corpora depending on the ambiguity of word forms with respect to NN and NE.

	training corpus		test corpus	
NN in ambiguity class	14,606		3,249	
unambiguous NN	9,861	(67.5 %)	2,240	(69.0 %)
tagged as NN			2,835	(87.3 %)
incorrect			48	(1.7 %)
instead of NE			39	**81.3 %**
NE in ambiguity class	3,621		706	
unambiguous NE	1,654	(45.7 %)	320	(45.3 %)
tagged as NE			603	(85.4 %)
incorrect			64	(10.6 %)
instead of NN			55	**85.9 %**
NN <u>and</u> NE in ambiguity class	1,912		372	
tagged as NN			94	(25.3 %)
incorrect			29	(30.9 %)
instead of NE			27	**93.1 %**
tagged as NE			273	(73.4 %)
incorrect			55	(20.2 %)
instead of NN			51	**92.7 %**
tagged NN <u>or</u> NE			367	(98.7 %)
incorrect			84	(22.9 %)
inverted NN-NE			78	**92.9 %**

These figures show that the overall error rate for common nouns (NN) is less than 2 %, whereas the automatically associated proper name tag (NE) is wrong in more than 10 % of all cases.

Both for common nouns and proper names the most frequent error is a confusion of the tags NE-NN (due to disambiguation as well as to lexical errors). Thus, we should expect an improvement of tagger accuracy if we put NE and NN together and use a single tag NOUN instead.

Test: Put NE and NN together in single class (NOUN)

Corpus statistics	training	test
Tokens	62860	13416
Tags	50	45
Lexical gaps	1756	283
Lexicon errors	355	49
Ambiguity classes	242	181
Ambiguity rate	1.66	1.65

Error statistics								
ambiguity	tokens	in %	correct	in %	LE	in %	DE	in %
1	8131	60.6	8110	99.7	21	0.3	–	–
2	2530	18.9	2385	94.3	12	0.5	133	5.3
3	2192	16.3	2119	96.7	13	0.6	60	2.7
4	493	3.7	449	91.1	2	0.4	42	8.5
5	62	0.5	59	95.2	1	1.6	2	3.2
6	8	0.1	7	87.5	0	–	1	12.5
total	13416	100.0	13129	97.9	49	0.4	238	1.8

Most frequent errors (by word form)			
number	word	correct tag	tagger tag
9	Osthold	NOUN	ADJD
6	das	PDS	ART
5	werden	VAFIN	VAINF
4	dem	PRELS	ART
4	Um	KOUI	APPR
4	Reich	NOUN	ADJD

Most frequent errors (by tags)		
number	correct tag	tagger tag
30	VVFIN	VVINF
16	VVFIN	VVPP
15	NOUN	ADJD
14	NOUN	ADV
12	NOUN	ADJA
12	KON	ADV
11	ADJD	VVPP
11	ADJD	ADV

A comparison of the statistical results as displayed here should be based on the results reported in section 6.1.2, page 22. There we have given the figures for the standard situation, on the basis of the Xerox HMM Tagger. Now, we give the figures for the same tagger, in a situation where common nouns and proper names are in a common "noun" class and thus not tagged differently.

The following changes are evident, as far as the corpus and lexicon characteristics are concerned: the corpus ambiguity rate, both for training and test corpus, is reduced (training corpus: from 1.69 in the standard test to 1.66 in the noun test; test corpus: from 1.67 to 1.65). Accordingly, the overall correctness rate increases from 97.3% to 97.9%. This result is expected: in section 6.1.2, clearly disambiguation errors at the level of NN vs. NE are a considerable part of the tagging errors. Evidently, now other disambiguation errors are most highly ranked in frequency: in this case, the ambiguities between finite and infinite verbs and between finite verbs and participles. However, still the problems encountered in the tagging of noun candidates are not solved completely. The table of error frequency by tags shows that now there are tag confusion pairs of the type NN vs. ADJD, NN vs. ADJ, NN vs. ADJA. Remember that, for example, the form *Osthold* led to NE vs. ADJD errors in the test displayed in section 6.1.2.

According to our calculations, nevertheless the amount of errors produced by the NN vs. NE confusion can be reduced to 46% of the original figure, by merging the NN and NE classes.

	training corpus		test corpus	
NOUN in ambiguity class	16,320		3,583	
unambiguous	12,174	(74.6 %)	2,713	(75.7 %)
tagged as NOUN			3,436	(95.9 %)
incorrect			25	(0.7 %)

6.2.5 Xerox HMM-tagger: Verb test

The standard tagset distinguishes *modals, auxiliaries* and *function verbs* which is, however, a mere lexical distinction and does not reflect the actual syntactic property in a given sentence. Therefore the verb forms themselves are not ambiguous, but the subclassification should contribute to a higher accuracy for the disambiguation of infinite and finite verb forms.

The
following tables show errors concerning the confusion of VAINF/VAFIN, VMINF/VMFIN and VVINF/VVFIN in the standard test. In German, all infinitives (except for *"sein"*) are ambiguous with finite verb (1st and 3rd person plural present tense).

	training corpus	test corpus	
V.INF in ambiguity class	3,084	584	
tagged as V.INF		263	(45.0 %)
incorrect		42	(16.0 %)
instead of V.FIN		40	**95.2 %**
V.FIN in ambiguity class	9,087	1,817	
tagged as V.FIN		947	(52.1 %)
incorrect		25	(2.6 %)
instead of V.INF		13	**52.0 %**
V.FIN <u>and</u> V.INF in ambiguity class	2,884	547	
tagged as V.INF		247	(45.2 %)
incorrect		41	(16.6 %)
instead of V.FIN		40	**97.6 %**
tagged as V.FIN		119	(21.8 %)
incorrect		13	(10.1 %)
instead of V.INF		13	**100 %**
tagged V.INF <u>or</u> V.FIN		366	(66.9 %)
incorrect		54	(14.8 %)
inverted V.FIN-V.FIN		53	**98.2 %**

Test: Do not distinguish VA-, VM- and VV- verbforms

Corpus statistics	training	test
Tokens	62860	13416
Tags	45	41
Lexicon gaps	1756	283
Lexical errors	542	65
Ambiguity classes	251	189
Ambiguity rate	1.69	1.67

Error statistics								
ambiguity	tokens	in %	correct	in %	LE	in %	DE	in %
1	7978	59.5	7942	99.6	36	0.5	–	–
2	2663	19.9	2474	92.9	13	0.5	176	6.6
3	2078	15.5	2013	96.9	8	0.4	57	2.7
4	589	4.4	518	88.0	7	1.2	64	10.9
5	81	0.6	71	87.9	1	1.2	9	11.1
6	19	0.1	16	84.2	–	–	3	15.8
7	8	0.1	7	87.5	–	–	1	12.5
total	13416	100.0	13041	97.2	65	0.5	310	2.3

Most frequent errors (by word form)			
number	word	correct tag	tagger tag
13	DM	NN	NE
9	Osthold	NE	ADJD
6	werden	VFIN	VINF
6	das	PDS	ART
5	Reich	NE	NN
4	haben	VFIN	VINF
4	dem	PRELS	ART
4	Um	KOUI	APPR
4	Deutschland	NE	NN

Most frequent errors (by tags)		
number	correct tag	tagger tag
55	NN	NE
46	VFIN	VINF
39	NE	NN
16	VFIN	VPP
14	ADJD	VPP
12	KON	ADV
11	VINF	VFIN
11	ADJD	ADV
10	NE	ADJD

The results of this test should again be interpreted against the figures displayed in section 6.1.2; again we have concentrated on the Xerox HMM Tagger, to make comparison easy.

In this test, the number of tags annotated both in the training and in the test corpus is reduced with respect to 6.1.2. This is expected, because we have merged the tags "`VAFIN`" and "`VMFIN`" into the class "`VVFIN`", and we have merged the respective infinite and participle form tags analogously. This accounts for a reduction of the number of tags. The corpus ambiguity rates, however, remain the same, of course.

The figures contained in the error statistics are not massively changed. The number of errors is slightly increased, but the changes do not seem to be significant.

The impact of the merging of the verb-subclasses on the overall treatment of verbs in the tagging process seems to be rather small.

	training corpus	test corpus	
VINF in ambiguity class	3,084	584	
tagged as VINF		271	(46.4 %)
incorrect		48	(17.7 %)
instead of VFIN		46	**95.8 %**
VFIN in ambiguity class	9,087	1,817	
tagged as VFIN		940	(51.7 %)
incorrect		23	(2.5 %)
instead of VINF		11	**47.8 %**
VFIN <u>and</u> VINF in ambiguity class	2,884	547	
tagged as VINF		255	(46.6 %)
incorrect		$\boxed{47}$	(18.4 %)
instead of VFIN		46	**97.9 %**
tagged as VFIN		111	(20.3 %)
incorrect		$\boxed{11}$	(9.9 %)
instead of VINF		11	**100 %**
tagged VINF <u>or</u> VFIN		366	(66.9 %)
incorrect		$\boxed{58}$	(15.9 %)
inverted VFIN-VFIN		57	**98.3 %**

Without the distinction VA-, VM- and VV- there are slightly more errors within the confusion class VINF/VFIN (58 versus 54 in the standard test).

6.2.6 Other tests

In 5.3 we defined experiments dealing with other tagset variations (eg. PRF vs. PPER, PTKVZ vs. ADJD), and we executed these tests as well. However, it turned out that the corpus material did not contain enough occurrences of the relevant tags to produce results which allow a reliable interpretation.

As we consider that our test and training corpus represents an average corpus, we might conclude that these phenomena are of less interest for overall tagger accuracy. But still, the accuracy for the concerned tags themselves should be evaluated with appropriate test data.

This is a general problem. For example, a 60,000 word manually annotated training corpus for German does not contain a single instance of an ambiguity much discussed in the linguistic literature, namely that between a relative pronoun and a demonstrative pronoun in an extraposition construction (*das Buch, das ich gelesen habe* (relative) vs. *das Buch, ja, das habe ich gelesen* (demonstrative)).

One could of course introduce such a phenomenon into a training corpus, because indeed some of the analyzed test material did contain examples of this phenomenon. However, it is not clear which impact the introduction of such "hard phenomena" would have, given that their frequency distribution seems to be very difficult to determine, when small-scale training corpora are used. From this point of view, one could argue that there is always a certain portion of phenomena which are not dealt with in training material for taggers and thus not analyzable correctly in the application of these taggers to test material. This would speak in favour of the hypothesis that there is an upper limit to the precision and correctness rate of automatic tagging which in a non-trivial way would be dependent on the size of training corpora.

6.3 Text type evaluation

The goal of the experiments below is to see what differences in tagger accuracy show up, when training and testing is done on different types of texts.

One major problem for validating tagger accuracy depending on text types is the lack of an objective measure for the difference of text types. The choice of texts for training and testing is based on the assumption that the structure of newspaper texts is different from fiction or poetry (eg. fairy tales).

For the tests we used the tagged German corpora available at RXRC. The tagset differs slightly from the STTS-tagset: The main difference is that there is no distinction between common nouns (NN) and proper names (NE), and there are some minor differences within function word classes.

Test and training corpora were collected from ECI-CD and from the WWW. In a first step they were tagged automatically with an available tagger. In an iterative process, we corrected the lexicon and the corpus tags, in order to obtain reliable data and similar conditions for the different types of text. Lexicon updates, however, did not include text specific proper names or other uncommon word forms, such as some antiquated words in fairy tales.

Finally, we assembled three different training corpora out of the manual tagged texts and used them to train three different HMM-Models. The resulting HMM-taggers were finally evaluated on each of the text (including the corpora which were used for training).

In the table below we list the texts used for training and testing. There are three training corpora TRAIN1, TRAIN2 and TRAIN3 built form the texts which are marked with the corresponding numbers. Unmarked texts are used for testing only.

Corpus statistics

- text types:
  news = newspaper, journal
  tale = fairy tale (old fashioned German)
  mix = collection of newspaper, technical reports, manuals, ...
- AW = *ambiguous word forms*:
  word forms with more than one tag in the lexicon (or guesser)
- AR = *ambiguity rate*:
  average number of tags per word form
- NL = *not lexicalized word forms*:
  number of tokens for which the guesser must provide appropriate tags.
- LE = *lexical errors*:
  number of words for which the lexicon (or guesser) does not provide the correct tag (i.e. not the one which is in the manual tagged corpus).

Corpus	Type	Tokens	AW	in %	AR	NL	in %	LE	in %
mix1 *(1)*	mix	7216	2426	*33.6*	1.55	139	*1.9*	25	*0.4*
mix2 *(1)*	mix	18739	6394	*34.1*	1.59	512	*2.7*	48	*0.3*
europa *(1)*	mix	7293	2602	*35.7*	1.63	56	*0.8*	9	*0.1*
spiegel *(2)*	news	11719	4256	*36.3*	1.65	390	*3.3*	50	*0.4*
taz *(2)*	news	7562	2774	*36.7*	1.64	209	*2.8*	20	*0.3*
welt *(2)*	news	6080	2187	*36.0*	1.63	138	*2.3*	22	*0.4*
spektrum	news	5701	2106	*36.9*	1.64	152	*2.7*	12	*0.2*
andersen *(3)*	tale	11774	4057	*34.5*	1.59	28	*0.2*	21	*0.2*
bechstein *(3)*	tale	7479	2565	*34.3*	1.58	17	*0.2*	9	*0.1*
grimm	tale	7284	2413	*33.1*	1.55	38	*0.5*	17	*0.2*
(1) TRAIN1	mix	33248	11422	*34.3*	1.59	707	*2.1*	82	*0.3*
(2) TRAIN2	news	25361	9217	*36.3*	1.64	737	*2.9*	92	*0.4*
(3) TRAIN3	tale	19253	6622	*34.3*	1.59	45	*0.2*	30	*0.2*

Lexicon and guesser contain 53 different tags and 245 different ambiguity classes.

The training corpora do not cover all of them. The coverage is shown in the table below.

Corpus	corpus tags		ambiguity classes		resulting tagger model
TRAIN1	53	*100.00 %*	215	*87.76 %*	HMM-1
TRAIN2	52	*98.11 %*	194	*79.18 %*	HMM-2
TRAIN3	52	*98.11 %*	177	*72.24 %*	HMM-3

Error statistics

The following table shows for each tagger model the absolute number of errors (including lexical errors) and the overall accuracy rate for each test corpus.

Tests are run on each text file, including training files. In the table below, highlighted numbers represent the results for texts which were not part of the training corpus for the given HMM model.

		HMM-1		HMM-2		HMM-3	
	trained on	mix		news		tale	
Corpus	Type	errors	accuracy	errors	accuracy	errors	accuracy
mix1	mix	136	98.12 %	215	**97.02 %**	264	**96.34 %**
mix2	mix	318	98.30 %	408	**97.82 %**	555	**97.04 %**
europa	mix	108	98.52 %	155	**97.87 %**	202	**97.23 %**
spiegel	news	305	**97.40 %**	217	98.15 %	346	**97.05 %**
taz	news	181	**97.61 %**	142	98.12 %	246	**96.75 %**
welt	news	119	**98.04 %**	90	98.52 %	156	**97.43 %**
spektrum	news	152	**97.33 %**	166	**97.09 %**	188	**96.70 %**
andersen	tale	251	**97.87 %**	304	**97.42 %**	163	98.62 %
bechstein	tale	161	**97.85 %**	163	**97.82 %**	94	98.74 %
grimm	tale	150	**97.94 %**	154	**97.89 %**	134	**98.16 %**

With the chosen training and test corpora, the following tendencies can be observed; they would have to be confirmed or modified by further testing on other text genres (technical manuals, instructions, colloquial language, ...).

- Evidently, identity of text types used for training and test leads to the best results.
- Training on a mix of different types results in similar tagger accuracy for both news and tales: between 97.4 % and 98.0 %.
- The tagger trained on fairy tales shows the greatest variations in the test results: from 96.3 % to 98.4 % (training texts excluded).
- Training on mixed texts leads to slightly better results than training on the old-fashioned style of fairy tales when tagging newspaper texts.
- The taggers trained on mixed texts (HMM-1) and on newspaper texts (HMM-2) obtain almost identical results in the tests on fairy tales.
- HMM-2 (trained on news) performs better than HMM-3 (trained on tales) for tests on mixed texts.

On the whole, the differences between the three configurations (training on sets TRAIN1, TRAIN2, TRAIN3) do not lead to very important differences in the performance of the tagging on the test material.

This may be due to the fact that the differences, which we assumed for the chosen text types, may not be very important, such that the statistical models trained from the three sets may not differ all that much. Tests with other text types (eg. technical manual style as opposed to newspaper style) might be more significant.

The impact of phenomena, which are specific for a given text type, might be tested separately, by means of training texts which are particularly geared to this question (and monitored accordingly).

In general, however, the test setup seems to be successfully applicable to carry out further tests concerning the impact of text types. Future tests should be based on a qualitative and quantitative description of the perceived differences between the test and the training texts; this could, at least in part, be obtained by the use of corpus query tools: with respect to certain constructions, quantitative profiles of test and training material could be established before the experiment is run; the tagger results on these phenomena should then be checked.

References

[Brill 1992] Eric Brill: **A Simple Ruls-Based Part of Speech Tagger.** In: *Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP-92)*, Trento, Italy, 1994.

[Brill 1994] Eric Brill: **Some Advances in Transformation-Based Part of Speech Tagging.** In: *Proceedings of AAAI-94*, 1994.

[Chanod, Tapanainen 1994] Jean-Pierre Chanod, Pasi Tapanainen: **Statistical and Constraint-Based Taggers for French.** Technical Report MLTT-016, RXRC Grenoble, November 1994.

[Chanod, Tapanainen 1995] Jean-Pierre Chanod, Pasi Tapanainen: **Tagging French - comparing a statistical and a constraint-based method**. In: *Proceedings of ACL Conference*, Dublin, 1995.

[Cutting et al. 1992] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun: **A practical part-of-speech tagger.** In *Proceedings of the 3th Conference on Applied Natural Language Processing*, Trento, Italy, 1992.

[Elworthy 1994] David Elworthy: **Automatic error detection in part of speech tagging.** In: *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, 1994.

[Elworthy 1995] David Elworthy: **Tagset Design and Inflected Languages**. In: *Proceedings of the ACL SIGDAT Workshop*, Dublin, 1995.

[Feldweg 1995] Helmut Feldweg: **Implementation and evaluation of a new German HMM model for POS disambiguation**. In: *Proceedings of the ACL SIGDAT Workshop*, Dublin, 1995.

[Kempe 1994] André Kempe: **A Probabilistic Tagger and an Analysis of Tagging Errors**. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, 1994.

[Kupiec 1992] Julian Kupiec: **Robust part-of-speech tagging using a hiddem Markov model.** In: *Computer Speech and Language 6*, 1992.

[Monachini, Calzolari 1994] Monica Monachini, Nicoletta Calzolari: **Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora.** A Common Proposal and Applications to European Languages. Draft technical report, ILC Pisa, October 1994.

[Monachini 1995] Monica Monachini: **ELM-IT: An Italian Incarnation of the EAGLES Lexicon Morphosyntactic Specifications. Definition of Lexicon Specification and Guidelines.** Work in progress, ILC, Pisa, October 1995.

[von Rekowski 1995] Ursula von Rekowski: **ELM-FR: A French Incarnation of the EAGLES Lexicon Morphosyntactic Specifications. Definition of Lexicon Specification and Guidelines.** Work in progress, gsi-erli, Paris, January 1996

[Schiller 1995] Anne Schiller: **DMOR: Benutzer-Handbuch.** (Draft) Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, 1995.

[Schiller 1996] Anne Schiller: **Multi-lingual Tagger Development.** Technical Report, RXRC Grenoble (forthcoming).

[Schiller et al 1995] Anne Schiller, Simone Teufel, Christine Thielen: **Guidelines für das Tagging deutscher Textcorpora mit STTS.** Universitäten Stuttgart und Tübingen, Aug. 95, to appear.

[Schmid 1994a] Helmut Schmid: **Part-of-Speech Tagging with Neural Networks.** In: Proceedings of the International Conference on Computational Linguistics, Kyoto, Japan, 1994: 172 - 176.

[Schmid 1994b] Helmut Schmid: **Probabilistic Part-of-Speech Tagging Using Decision Trees.** In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, 1994: 44-49.

[Schmid 1995a] Helmut Schmid: **Statistische Disambiguierung von Präpositionen für den Transfer.** Technical Report: VERBMOBIL Memo, No. 87, Feb. 1995, Universität Stuttgart.

[Schmid 1995] Helmut Schmid: **Improvements in Part-of-Speech Tagging with an Application to German.** In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 172–176, Kyoto, Japan.

[Schmid, Kempe 1994] Helmut Schmid, André Kempe: **Tagging von Corpora mit HMM, Entscheidungsbäumen und Neuronalen Netzen**. Arbeitstreffen Lexikon+Text, Tübingen, February 1994.

[Schmid 1994] Helmut Schmid: **Part-of-speech Tagging with neural Networks.** Arbeitstreffen Lexikon+Text, Tübingen, February 1994.

[Teufel 1995a] Simone Teufel: **ELM-DE: A typed German incarnation of the EAGLES-TS.** Definition of Lexical Descriptions and Classification Guidelines. EAGLES report, January 1996

[Teufel 1995b] Simone Teufel: **ELM-EN: A typed English incarnation of the EAGLES-TS.** Definition of Lexical Descriptions and Classification Guidelines. EAGLES report, Aug. 95, to appear

[Thielen, Schiller 1994] Christine Thielen, Anne Schiller: **Ein kleines und erweitertes Tagset fürs Deutsche**. In *Tagungsberichte des Arbeitstreffens Lexikon + Text*, Tübingen, Germany, (to appear).

[Wilkens, Kupiec 1995] Mike Wilkens, Julian Kupiec: **Training Hidden Markov Models for Part of Speech Tagging.** Xerox Corporation, 1995.

A Description of tested phenomena/problems in detail

A.1 Proper Names vs. Common Nouns

Problem: The distinction between proper names (NE) and common nouns (NN) is sometimes only possible with semantic knowledge as there is no orthographical difference in German (lower/upper case). An additional problem is the disambiguation of proper nous and foreign language material.

Tagging practice IMS/TUE and STTS: Distinction is made on semantic grounds (systematicalized by a list of items to be regarded as proper nouns, e.g. names of geographic places, persons...). Complex proper nouns are tagged with tags that individual parts get in a non-naming context. Foreign word forms are (manually) identified by world-knowledge

Examples for use of tags NE and NN (and FM):

*Ich trinke eine Cola/***NN** *und arbeite bald bei Coca-Cola/***NE** *in Atlanta/***NE**.

*ich wohne im Hotel "Domizil/***NE***".*

*ich gehe ins "Holiday/***NE** *Inn/***NE***".*

*Deutsche/***ADJA** *Angestellten-Gewerkschaft/***NN**

*der VFB/***NE**

*der VFB-Vorstand/***NN**

*die Grünen/***NN**

*Weil/***NE** *am/***APPRART** *Rhein/***NE**

*Tai/***FM** *Chi/***FM**

*der Film "La/***FM** *mujer/***FM** *de/***FM** *Benjamin/***NE***"*

Tests: Type Ia, Ib[12]

1. NE and NN are put into one class with tag N.
2. NE, NN and FM are put into one class with tag N.
3. Lexically ambiguous word forms between proper and common nouns (that is those that have both tags in the lexicon) are tagged as N, the rest of NE and NN stay as they are.

[12] As reference material, we always compare results of the single tests with the reference test, that is, the corpora tagged in the current STTS practice.

A.2 Verb Prefixes

Problem: In German, verb prefixes can be separated from the verb. These stand-alone prefixes (PTKVZ) are ambiguous with adpositions, particles, adjectives, adverbials. Dependencies are long distance. Only adjectival prefixes are an open class. Too much noise expected in cases like

*Woran machst Du das fest/***PTKVZ***?* vs.

*Er macht Kissenfüllungen immer extrem fest/***ADJD***.*

Tagging practice IMS/TUE: Only particle prefixes and else untaggable verbal prefixes tagged PTKVZ, all other like in normal contexts.

*Er stellte das dar/***PTKVZ***.*

*Sie fuhr rad/***PTKVZ***.*

but: *Er ging den Berg hinauf/***ADV***.*

but: *Sie hielten das streng geheim/***ADJD***.*

but: *Er ging der Sonne entgegen/***APPO***.*

Tagging Practice STTS: All these groups are considered **PTKVZ**.

*Er stellte das dar/***PTKVZ***.*

*Sie fuhr rad/***PTKVZ***.*

*Wie setzen wir das Auto instand/***PTKVZ***?*

*Er ging den Berg hinauf/***PTKVZ***.*

*Sie hielten das streng geheim/***PTKVZ***.*

*Er ging der Sonne entgegen/***PTKVZ***.*

Tests: Type Ic, II?

1. Definition changes: Add adjectival prefixes to ADJD class
2. A new class for adjectival verbal prefixes, ADJDPART

A.3 Finite vs. Non-finite Verb Forms

Problem: In German, there is an ambiguity between finite 1stPl and 3rdPl forms (V.FIN) and infinitives (V.INF). Modal auxiliaries introduce another form of ambiguity to infinitive-forms, that is, they can use the infinitival form also in perfect context. We suspect that the encoding of the verb type has an external disambiguation potential with the finite/infinitive ambiguation class, and with the special infinitives of modal verbs.

Tagging practice IMS/TUE and STTS: Finite and infinite forms are disambiguated. Modal infinitives in perfect context are tagged as infinitives. The distinction between auxiliaries, modal verbs and content verbs is only lemma-based.

*Er hat das nicht essen können/***VMINF***.*

*Er wird das nicht essen können/***VMINF***.*

*Er hat das nicht gekonnt/***VMPP***.*

Tests: Type Ic

1. No distinction between V,M,A; modal perfect infinitives are tagged as VINF
2. Distinction between V,M,A; modal perfect infinitives are tagged as VMPP
3. No distinction between V,M,A; modal perfect infinitives are tagged as VPP
4. Distinction between V,M,A; modal perfect infinitives are tagged as a seperate class VMINFPP.

A.4 Personal vs. Reflexive Pronouns

Problem: The following (1st and 2nd person) forms are ambiguous in German between personal pronouns (PPER) and reflexive pronouns (PRF): *mich, dich, dir, mir, euch, uns*. This makes disambiguation dependent on whether a verb is reflexive or not. Criteria for PRF are (STTS; for human taggers)

- verb must be lexicalized as reflexive verb
- Subject and pronoun in question must agree (German: person and number)
- case of pronoun in question must not clash with demanded case of lexicalized reflexive verb (dative or accusative).

The tagger lexicon lexicon does not provide syntactic information (such as subcategorization or reflexivity). Moreover, the relevant features (gender, case) are not destinguished in the tagset. Thus, automatic taggers cannot access the criteria which are necessary to disambiguate personal/reflexive pronouns. We still can test if the instances of non-ambiguous pronouns (ihn/ihr/ihm vs. sich) can provide enough statistic data to handle the ambiguous cases.

Tagging practice IMS/TUE: Special ambiguous class tag for these forms: **PPERRF**.

*Ich/***PPER** *kaufe mir/***PPERRF** *die Kette.*

*Ich/***PPER** *kaufe dir/***PPERRF** *die Kette.*

*Er/***PPER** *rasiert sich/***PRF**.

*Er/***PPER** *rasiert ihn/***PPER**.

*Wir/***PPER** *lieben euch/***PPERRF**.

Tagging practice STTS: Disambiguation is realized between the two classes.

*Ich/***PPER** *habe mich/***PRF** *rasiert.*

*Du/***PPER** *hast mich/***PPER** *auf der Straße getroffen.*

*Du/***PPER** *hast dich/***PRF** *mit mir/***PPER** *auf der Straße getroffen.*

Tests: Type Ib

1. ambiguous forms in extra class PPERRF (IMS-TUE practice)
2. ambiguous forms tagged as PRF
3. ambiguous forms tagged as PPER
4. ambiguous forms and all other PRF and PPER together in one class PPERRF

B Additional suggestions for tagset tests

B.1 Participles vs. Adjectives

Problem: In German, past participles can be used as part of the verbal complex (VVPP) or as adjectives (ADJD). As neither predicative adjectives nor participles inflect, the distinction gets ambiguous for verbs that can form a passive with *sein* (stative passive[13]). In that case, ambiguity exists with copula constructions where the participle is used as a predicative adjective.

*Sie sind von den Änderungen direkt betroffen/***VVPP**.

*Sie sind sehr betroffen/***ADJD** *und kleinlaut.*

It can also get ambiguous for constructions with "werden":

*Der Tisch wird von mehreren Möbelpackern verrückt/***VVPP**.

*Ich werde verrückt/***VVPP**.

Tagging practice IMS/TUE: All uninflected participles are tagged as V.PPF, except for special lexicalized forms (1,4,5)) and except for adverbially modifying participles (8,10).

*er ist verrückt/***ADJD**.	*(1)*
*der Tisch ist verrückt/***VVPPF**.	*(2)*
*der Tisch wurde verrückt/***VVPPF**.	*(3)*
*der alte Mann wurde verrückt/***ADJD**.	*(4)*
*er ist gelehrt/***ADJD**.	*(5)*
*er hat ihn reiten gelehrt/***VVPPF**.	*(6)*
*sie sind betroffen/***VVPPF** *und kleinlaut.*	*(7)*
*das war gekonnt/***ADJD** *gemacht/***VVPPF**.	*(8)*
*die Sache ist geplant/***VVPPF**.	*(9)*
*das gewollt/***ADJD** *verlorene Spiel.*	*(10)*

STTS: In order to get more consistency into this, we decided to tag adverbially used participles as ADJD whereever this could be decided savely, that is, also in the case of (7). Distinction between *belonging to verbal complex* and *not belonging to VC.*

Criteria for **VVPP**:

- Form of *sein* could be replaced by *werden* without too much semantic difference;
- Sentence can be put into the active;
- There is a *von*-PP denoting logical subject.

Criteria for **ADJD**:

- Above tests don't work;
- Participle could be replaced by a regular adjective.

[13] German: Zustandspassiv

Tests: Type II. We have material where ambiguous forms are manually isolated and marked. This material can be used with tags one way or the other.

1. STTS practice
2. Take different syntactic constructions into the ADJD class:
   (a) no subcategorized PPs of participle in question
   (b) subcategorized PP except *von*
   (c) *von*-PP of particle there

B.2 Indefinite vs. Demonstrative Pronouns and Determiners

Problem: The core indefinite pronouns (PIS) and determiners (PIAT) and demonstrative pronouns (PDS) and determiners (PDAT) distribute very similarly (in the place of the article, or substituting an NP, for most cases).

However, there is a problem class of wordforms somewhere between an adjective and a core indefinite pronoun: *viel-, all-, beid-*, and another class of indefinite pronouns (noninflected) that can stand *before* an article. So, indefinite pronouns and determiners as a class distribute very heterogeneously.

Tagging practice IMS/TUE: Classes for core indefinite pronouns and determiners and demonstrative pronouns and determiners form a common class, **PRO..**. Diverging items have their specific, lemma-based class **PBEID, PALL, PVIEL** and **PNFL** (for non-inflected indefinite pronouns[14]).

*das/***PROS** *weiß ich nicht*

*diejenige/***PROAT** *Person, die dasselbe/***PROAT** *Kleid trägt*

*manches/***PROAT** *andere/***ADJA** *Thema*

*manch/***PNFL** *anderes/***ADJA** *Thema*

*keiner/***PROS** *war da; in keiner/***PROAT** *Form*

*er hat viele/***PROAT** *Bücher; er trinkt viel/***PROAT** *Wein*

*er sieht vieles/***PROS** *ein*

*er ißt viel/***PVIEL***; er lacht wenig/***PVIEL**

*alles/***PALL***, was recht ist*

*all/***PALL** *diese/***PROA** *vielen/***PROA** *Leute*

*die beiden/***PBEID** *kamen gleichzeitig*

*beide/***PBEID** *waren da*

*beide/***PBEID** *Läufer waren gleich schnell*

Tagging practice STTS: Demonstrative and indefinite pronouns form seperate tags. Special, lemma-based classes are reduced to indefinite pronouns or adverbs (in adverbial readings of *viel*).

*das/***PDS** *weiß ich nicht*

*diejenige/***PDAT** *Person, die dasselbe/***PDAT** *Kleid trägt*

*manches/***PI.AT**[15] *andere/***ADJA** *Thema;*

*manch/***PI.AT** *anderes/***ADJA** *Thema*

*keiner/***PIS** *war da; in keiner/***PI.AT** *Form*

*er hat viele/***PI.AT** *Bücher; er trinkt viel/***PI.AT** *Wein*

*er sieht vieles/***PIS** *ein*

*er ißt viel/***PIS***; er lacht wenig/***ADV**

[14] In a mostly distribution-based TS, special lemma-based classes like **PBEID, PALL, PVIEL, PNFL** are not desirable. Distinction in adverbial and pronominal function should be given.

[15] The point symbolizes, that the actual form found in the corpus might have another letter there (PIDAT vs. PIAT) as indefinite pronouns are interseperated later. This distinction however is not relevant for us here, as we only consider demonstrative vs. indefinite pronouns here.

*alles/***PIS** *, was recht ist*

*all/***PI.AT** *diese/***PDAT** *vielen/***PI.AT** *Leute*

*die beiden/***PIS** *kamen gleichzeitig*

*beide/***PIS** *waren da*

*beide/***PI.AT** *Läufer waren gleich schnell*

Tests: Ia, Ic?

1. Granularity tests: putting STTS classes (PI.AT and PD.AT) together and (PIS and PDS).
2. impact on PRELS/ART?

B.3 Indefinite Determiners

Problem: However, the class *indefinite attributive pronouns* (PIAT) or *indefinite determiners* was very heterogeneous in distribution now, especially in respect to articles in the surrounding, as the following examples illustrate:

> **manch** *ein Mensch*
>
> *die* **beiden** *Verwandten*
>
> *ein* **jeder** *Mensch*
>
> **jeder** *Mensch*
>
> *ich habe* **wenig** *gelacht*
>
> *ich habe nur ein* **wenig** *gegessen*
>
> *ich habe* **etwas** *gesehen*
>
> *ein* **bißchen** *Schokolade*
>
> *er sah* **etwas** *gequält aus*

Tagging practice IMS/TUE: As described in B.2

Tagging practice STTS: We decided to distinguish the group into indefinite determiners that can occur together with another (non-indefinite) determiner (an article, a demonstrative or possessive pronoun), and into those, that must occur on their own. This property should be lemma-based, i.e. only depending on the ability of the given indefinite determiner to combine with a determiner (independent of the question if there is a determiner in the actual context). Adverbial readings should be tagged as adverbs, not as indefinite pronouns.

> *solch/***PIDAT** *ein Mensch*
>
> *in keiner/***PIAT** *Form*
>
> *er sah etwas/***ADV** *gequält aus*
>
> *etwas/***PIAT** *Schokolade*

Tests: I, II

1. Granularity changes: put PIDAT and PIAT into one class (as is in ELM-DE).
2. Change in definition: context depencency of PIAT vs. PIDAT as compared to lemma based disambiguation (as it is now).

B.4 Adjective vs. Adverb

Problem: There are certain adverbially used word forms that share the same surface form as predicatively used adjectives, but there is a semantic difference between the uses.

Tagging practice IMS/TUE: The test for manual taggers was: if there is a different semantic from attributive or predicative use and adverbial use, then check if the word form can be used predicatively (with the same semantics as the adverb). If this is not the case, the word form is tagged as **ADV**, else **ADJD**.

> *sie waren äußerst/***ADV** *gewitzt*
>
> *zur äußersten/***ADJA** *Not*
>
> *längst/***ADV** *vergessen*
>
> *früher/***ADV** *ist er nie so spät gekommen.*
>
> *es ist gerade/***ADV** *14 Uhr*

Tagging practice STTS: We have a list of ambiguous word forms that can be adverbs. If one of them is found in a sentence adverbial meaning, modifying an adjective or adverb, then the tag **ADV** is given. Else (or in same meaning as attributively): **ADJD**.

> *die Frage ist nämlich/***ADV***, ob ...*
>
> *die nämliche/***ADJA** *Frage*
>
> *er ist heute früher/***ADJD** *gekommen*
>
> *früher/***ADV** *ist er nie so spät gekommen.*
>
> *die Straße ist eben/***ADJD**
>
> *das ist eben/***ADV** *die Frage.*
>
> *der Garten ist weit/***ADJD** *angelegt*
>
> *es sind weit/***ADV** *mehr als 100 Gäste*

Tests: II, Ic. To find candidates: create list (morphology system: doubly analysed forms ADJD/ADV). Run granularity/definition change tests on them:

1. Create special tags for them
2. Context/syntactic construction dependent changes
3. put these forms to the ADV
4. put these forms to the ADJD

B.5 Conjunction vs. Adverb

Problem: The word forms *doch, aber, denn, jedoch* can be used as coordinating conjunction:

er war wütend, denn er hatte die Zahnpasta vergessen.

but also purely adverbial:

wo hatte er sie denn liegengelassen?

In a varity of cases, there is a reading between adverb and conjunction, but because coordinating conjunctions don't change the position of the verb in German, the position of the verb cannot be takene as a test.

Tagging practice STTS/IMS-TUE: *doch, aber, denn, jedoch* are tagged as **KON** when they appear as first word of a sentence or as first word after a comma in a complete coordinated sentence. In all other situations, they are tagged as **ADV**.

Tests: Ib

1. tagging always as ADV
2. special tag KONADV

B.6 Function Words "wie", "als"

Problem: The word forms "wie" and "als" fulfil a variety of functions.

Examples for "wie":

- Comparision
- Coordinating conjunction
- Subordinating conjunction
- Apposition/enumerations
- direct or indirect questions

Most of the distinctions between these functions are long distance or semantically determined.

Tagging practice IMS-TUE: Distinction between all these functions is drawn. The tag KOKOM is given by semantical criteria.

*Wie/***PWAV** *groß Du geworden bist!* *(1)*

*Wie/***PWAV** *groß bist Du geworden?* *(2)*

*er fragte nach, wie/***PWAV** *es mir so gehe.* *(3)*

*Wie/***KOUS** *das Presseamt mitteilte ...* *(4)*

*so schnell wie/***KOKOM** *du.* *(5)*

*Männer wie/***KON** *Frauen strömten herbei.* *(6)*

*so schnell wie/***KOKOM** *du niemals rennen wirst.* *(7)*

*viele Länder, wie/***KON** *Spanien, zogen nach* *(8)*

*Sie arbeitet wie/***KOKOM** *eine Besessene* *(9)*

Tagging practice IMS-TUE: Distinction purely syntactic:

- Direct questions: "wie" is PWAV (2)
- wie is followed by V-Final sentence: KOUS (3,4,7)
- wie is followed by any other phrase (NP, AP...): KOKOM (1,5,6,8,9)

Test:
1. Granularity test: put PWAV and KOUS together
2. Definition test: borderline KOKOM/KOUS.
   Comparisons of sentences tagged as KOKOM or KOUS.

B.7 Foreign Material

Problem: There is more foreign material (FM) and foreign names of entities (more than one item long) in the corpus than expected. Some of this material forms part of the German syntax (names), some don't (film titles or translations). Proverbial sayings from Latin, French or English are inserted and have a function as a whole, but a single function cannot be given to its parts (because we are not describing Latin, French or English parts of speech). Is the length of the inclusion relevant? What does the language model do with chains of FM?

IMS/TUE: All foreign material had to be sorted into a special class. In analogy zu z.B./ADV: last/ADV but/ADV not/ADV least/ADV

Tagging practice STTS: Names of cities, persons, institutions are tagged as proper names, where one is sure. Foreign common nouns which are already lexicalized in German (Yoga, Joghurt, Jeans) are to be tagged as NN. All other foreign material to be tagged as FM.

*Der beliebte Film " A/***FM** *fish/***FM** *called/***FM** *Wanda/***NE** *".*

*per/***FM** *se/***FM** *ist das kein Problem.*

*der berühmte "dedazo/***FM***" (Fingerzeig) funktioniert noch immer.*

*Sie essen heute a/***FM** *la/***FM** *carte/***FM**

*er war damals schon lange persona/***FM** *non/***FM** *grata/***FM***.*

*er ist primus/***FM** *inter/***FM** *pares/***FM**

*Der spanische Titel " Mi/***FM** *querido/***FM** *Tim/***NE** *Mix/***NE** *"*

*und dann, last/***FM** *but/***FM** *not/***FM** *least/***FM***, gibt es ein gutes Mittagessen.*

Tests: I

1. simple granularity test
2. inclusion in Test on NN/NE.

C Overview of German tagset STTS

Tag Description Examples

Tag	Description	Examples
ADJA	attributive adjective	*[das] große [Haus]*
ADJD	adverbially or predicatively used adjective	*[er fährt] schnell [er ist] schnell*
ADV	adverb	*schon, bald, doch*
PAV	pronominal adverb	*dafür, dabei, deswegen, trotzdem*
APPR	preposition or left part of circumposition	*in [der Stadt], ohne [mich] von [Anfang an]*
APPRART	preposition with article	*im [Haus], zur [Sache]*
APPO	postposition	*[ihm] zufolge, [der Sache] wegen*
APZR	right part of circumposition	*[von jetzt] an*
ART	definite or indefinite article	*der, die, das, ein, eine*
CARD	cardinal number	*zwei [Männer], [im Jahre] 1994*
FM	material of a foreign language	*[Er wird mit "] A big fish [" übersetzt]*
ITJ	interjection	*mhm, ach, tja*
KOUI	subordinating conjunction with "zu" infinitive	*um [zu leben], anstatt [anzufragen]*
KOUS	subordinating conjunction with a sentence	*weil, daß, damit, wenn, ob*
KON	coordinating conjunction	*und, oder, aber*
KOKOM	comparative conjunction	*als, wie*
NN	common noun	*Tisch, Herr, [das] Reisen*
NE	proper noun	*Hans, Hamburg, HSV*
PDS	demonstrative pronoun	*dieser, jener*
PDAT	demonstrative determiner	*jener [Mensch], dieses [Buch]*
PIS	indefinite pronoun	*keiner, viele, man, niemand*
PIAT	indefinite determiner	*kein [Mensch], irgendein [Glas]*
PIDAT	indefinite modifier with determiner	*[ein] wenig [Wasser], [die] beiden [Brüder]*
PPER	irreflexive personal pronoun	*ich, er, ihm, mich, dir*
PRF	reflexive personal pronoun	*sich, einander, dich, mir*
PPOSS	possessive pronoun	*meins, deiner*
PPOSAT	possessive determiner	*mein [Buch], deine [Mutter]*
PRELS	relative pronoun	*[der Hund ,] der*
PRELAT	relative determiner	*[der Mann ,] dessen [Hund]*

Tag	Description	Examples
PWS	interrogative pronoun	*wer, was*
PWAT	interrogative determiner	*welche [Farbe], wessen [Hut]*
PWAV	interrogative or relative adverb relative adverb	*warum, wo, wann, worüber, wobei*
PTKZU	"zu" before an infinitive	*zu [gehen]*
PTKNEG	negation particle	*nicht*
PTKVZ	seperated verbal particle	*[er kommt] an, [er fährt] rad*
PTKANT	answer particle	*ja, nein, danke, bitte*
PTKA	particle with adjective or adverb	*am [schönsten], [zu] schnell*
TRUNC	first (seperated) part of composition	*An- [und Abreise]*
VVFIN	finite content verb	*[du] gehst, [wir] kommen [an]*
VVIMP	imperative content verb	*komm [!]*
VVINF	infinitive of content verb	*gehen, ankommen*
VVIZU	infinitive of content verb with incorporated "zu"	*anzukommen, loszulassen*
VVPP	past participle of content verb	*gegangen, angekommen*
VAFIN	finite verb, primary auxiliary	*[du] bist, [wir] werden*
VAIMP	imperativ, primary auxiliary	*sei [ruhig !]*
VAINF	infinitive, primary auxiliary	*werden, sein*
VAPP	past partiple, primary auxiliary	*gewesen*
VMFIN	finite verb, modal auxiliary	*dürfen*
VMINF	infinitive, modal auxiliary	*wollen*
VMPP	past participle, modal auxiliary	*gekonnt*
XY	non-word	*1.2.2., H2O, D2XW3*
$,	comma	,
$.	punctuation at end of sentence	**. ? ! ; :**
$(	other punctuation, sentence internal	**- [ ] ( )**