Towards a cross-linguistic tagset

Drs. Jan Cloeren
TOSCA Research Group for Corpus Linguistics
Dept. of Language & Speech, University of Nijmegen
NL-6525 HT Nijmegen, The Netherlands
e-mail: cloeren@lett.kun.nl

April 29, 1993

Abstract

With the spread of large quantities of corpus data, the need has arisen to develop some standard not only for the format of interchange of text (an issue which has already been taken up by the Text Encoding Inititiave), but also for any information added in some subsequent stage of (linguistic) enrichment. The research community has much to gain by such standardization since it will enable researchers to effectively access and therefore make optimal use of the results of previous work on a corpus. This paper provides some direction of thought as to the development of a standardized tagset. We focus on a minimal tagset, i.e. a tagset containing information about wordclasses. We investigate what criteria should be met by such a tagset. On the basis of an investigation and comparison of ten different tagsets that have been used over the years for the (wordclass) tagging of corpora, we arrive at a proposal for a cross-linguistic minimal tagset for Germanic languages[1].

Part I
Introduction

The last few years there has been an increasing interest in the use of corpus data, especially by those working in the field of natural language processing (NLP). This development can in part be ascribed to the fact that the scale on which such data are becoming available has increased: as publishing houses, industry, etc. switched to an electronic format for the interchange of texts this meant a dramatic increase in the amount of text that was readily available to anyone interested, while developments in hardware and in software have made it possible to manipulate large quantities of data (more) effectively.

Earlier corpus-based approaches proved to be laborious undertakings. Before the 'real' work could begin one would have to go through the painstaking process of designing a corpus, gaining permission from publishers to use (part(s) of) texts and somehow making the texts computer-readable. The size of the corpora was very much determined by such factors as cost (in terms of time and money to be invested) and availability of data. Corpora generally were compiled for particular research purposes, such as the investigation of a particular variety of the language. Typical examples here are the Brown Corpus and the Lancaster-Oslo/Bergen (LOB) Corpus which were both compiled with the intention of representing a cross-section of the language (American English and British English respectively). Compilers of corpora would each adopt his/her own conventions for representing the textual data. In a similar fashion the further processing of the data, the enrichment of (basically) raw text with some kind of linguistic information (wordclass information, syntactic, semantic and/or even pragmatic information) would largely depend on the time and money available, and also on the particular interests of the researchers involved.

Recently, with the increase in the amount of data that are becoming available, the attention of corpus compilers has been drawn to the need for a common interchange format for texts in order to make these data more readily

1. In this paper the focus is on British English, American English, Dutch and German.

accessible for third parties. The Text Encoding Inititiative (TEI) has undertaken to develop a standard for the marking up of texts that is based on the Standard Generalized Markup Language (SGML). So far we have not yet reached the point where any serious attempts are being made to standardize the linguistic information that is being added in successive stages of linguistic enrichment. Instead, researchers from different backgrounds and with different beliefs are working their own turf exploring a variety of methods for the enrichment of corpora, ranging from purely stochastic to strictly rule-based approaches, which seem to be in competition with each other. In view of the current state of the art in corpus analysis, the amount of work that has already been done in the area of tagging corpora for wordclass information and the experience that has been gained in the process, it would appear that by now the time has come to start thinking about developing some sort of standard for the encoding of wordclass information.

In the remainder of this paper we focus on the design of a minimal tagset, i.e. a tagset containing information about wordclasses, that will provide a common basis for the wordclass tagging of texts written in Germanic languages.
We first compare a number of tagsets that have been or are being used for the tagging of wordclass information in prominent corpora. On the basis of this comparison we arrive at a number of criteria that a standardized (minimal) tagset should meet. Finally put forward a proposal for a basic tagset that may be applied cross-linguistically.

Part II
Tagsets: a comparison

In order to give the reader some idea of the kind of (wordclass) information covered in various tagsets, a comparison is made of the tagsets employed in ten corpora[2] the Brown Corpus (Kučera and Francis, 1967), the Lancaster-Oslo/Bergen Corpus (Johansson et al., 1978), the SUSANNE Corpus (Sampson, forthcoming), the Penn Treebank (Santorini, 1991), the IBM-Lancaster Corpus (Black et al., 1993), the British National Corpus (Leech, 1993), the ICE Corpus (Greenbaum, 1991), the Tosca Corpus (Oostdijk, 1991), the Dutch Eindhoven Corpus (uit den Boogaart, 1975), and the German IDS-Mannheim Corpus (Neumann, 1987).

As a first step in making a comparison of the ten tagsets employed in the wordclass tagging of the above corpora we start by distinguishing the wordclass categories that these tagsets have in common. There are nine categories that are (in one fashion or another) included in each of the tagsets: noun, pronoun, article/determiner, adjective, adverb, preposition, conjunction, verbal form, and interjection. Apart from these nine categories we distinguish one other category, which by definition is open-ended since it is intended to cover all tags that do not fit into any of the other categories. We shall refer to this category as "open category". Next, we make an inventory of the different tags employed by each of the tagsets. As a result we obtain lists[3] of tags for each of the wordclass categories we distinguished. For example, Table 1 lists the various tags as they occur in the ten tagsets under consideration for the tagging of conjunctions (incl. connectives).

There appears to be a great deal of overlap: while actual tags may differ (cf. CO, CI, 70 for conjunctions in general, in the ICE, Brown and Eindhoven tagsets respectively), there seems to be some consensus as to the kind of information one wants to encode or tag. Before going into this, however, we take a closer look at both the format and the nature of the tags.

Starting with the oldest tagset included in our comparison, the Brown tagset, we see that this is in a sense a very 'flat' coding scheme. It was developed with the intention of encoding general wordclass information. The tags consist of character sequences which encode the wordclass category of a word and occasionally extended information such as form, aspect or case. Thus for example we find JJ for adjective and CC for coordinating conjunction, but also VB for base form of verb, VBD for past tense of verb, VBG for present participle, etc.

The LOB tagset, but also the tagsets

2. Information regarding a detailed characterization of the corpora including size, design, research context and method of processing can be found in the corresponding literature

3. These lists are rather sizeable and have therefore not been included. They are available, however, from the author (via e-mail).

Conjunctions

Code	*Interpretation*	*Corpus*
CO	conjunction	ICE
CI	conjunction	Brown
70	conjunction	Eindhoven
CONJ	conjunct	Tosca
COAP	appositive conjunction	Tosca
COCO	coordinative conjunction	Tosca
CJC	coordinating conjunction	BNC
CC	coordinating conjunction	Brown, LOB, PENN,SUS
"	" "	IBML
CCnn	part of CC	SUS
CCB	coordinating conjunction "but"	IBML
CF	semi-coordinating conjunction "yet"	IBML
ABX	double conjunction (both ...and)	Brown, LOB
DTX	double conjunction (either ... or)	Brown, LOB
LE	leading co-ordinator(both, both..and)	SUS, IBML
LEnn	part of LE	SUS
CON	connective	Tosca
CON	connective	ICE
72	comparative connectives	Eindhoven
COSU	subordinating conjunction	Tosca
CJS	subordinating conjunction	BNC
CS	subordinating conjunction	Brown, LOB, SUS,IBML
CSnn	part of CS	SUS
BCS	before subord.conjuction(even(if))	IBML
71	subordinating conjunction	Eindhoven
73	subord. conj. matrix sent. wordorder	Eindhoven
74	introductory part of conjunctive groups	Eindhoven
CJT	conjunction "that"	BNC
CST	conjunction "that"	IBML
CSA	conjunction "as"	IBML
CSN	conjunction "than"	IBML
CSW	conjunction "whether"	IBML

Table 1: conjunctions

of Penn Treebank, IBM-Lancaster and the British National Corpus very much follow the Brown tagset. Of these tagsets, the LOB tagset is of course closest to the Brown tagset, since the corpus was compiled with the intention of comparing American and British English from the same year (1961).

The Penn Treebank tagset appears to be a reduced version of the Brown/LOB tagset. The reduction becomes manifest in that some tags are less detailed. For example, where Brown/LOB distinguish between possessive pronoun, personal pronoun, and reflexive pronoun and with each of these has a further subclassification (e.g. PP1A for personal pronoun, first person singular nominative; PP1AS for personal pronoun, first person plural nominative; PP1O for personal pronoun, first person singular accusative; etc.), Penn Treebank only has one tag, PP, to cover all personal, possessive, and reflexive pronouns. All in all the derivation of the Penn Treebank from the Brown/LOB tagset is fairly straightforward. A minor deviation occurs in the tagging of genitive markers in the case of nouns. Here according to the Penn Treebank tagging scheme genitive markers are assigned their own tag, which is not the case with the Brown/LOB tagging scheme.

The various tagging schemes that have emerged from the cooperation between IBM and the Lancaster-based Unit for Computer

Research on the English Language[4] all can be placed within what might be referred to as the Brown/LOB tradition in tagging. For our discussion we single out one particular tagging scheme, called CLAWS2a. This tagging scheme was developed to encode the information needed at wordclass level in order to be able to parse computer manuals. Although this tagging scheme clearly shows some resemblance to the Brown/LOB tagging scheme, it stands out in that with certain categories (such as verb and noun) there is a lot of extra detail that we do not find elsewhere. For example, tags such as *noun of style* and *noun of organization* are unique for this tagset. When we look at the format of the tags we find that there may be up to five characters per tag, while the leftmost characters relate the more general wordclass information and the characters that occur towards the right specify further detail. Compound lexical items can be encoded by means of tags that extend over more than one word. For example, while II would be the tag for simple preposition, the complex preposition in spite of is tagged II31 II32 II33.

The tagset that has been developed within the framework of the British National Corpus (BNC) contains only some 60 different tags[5] The reason for this must be sought in the fact that the BNC intends to incorporate a very large amount of text (approx. 100 million words). The grammatical (i.e. wordclass) tagging of the corpus is to be carried out automatically. For reasons of efficiency and also to increase the success rate of the tagging it was decided to have a rather small tagset. A comparison of the BNC tagging scheme to the one used in the IBM-Lancaster Corpus shows that the two tagging schemes are closely related. Both can be characterized as 'flat' tagging schemes. The major difference between the two, apart from their sizes, appears to be that the BNC uses more mnemonic abbreviations for otherwise similar tags[6] ; for example general adjectives get the label *AJO* instead of *JJ*.

4. Black et al. (1993).
5. The British National Corpus is a joint undertaking by Oxford University Press, Longman, and W.& R. Chambers, the universities of Lancaster and Oxford, and the British Library.
6. This was probably done in the hope to improve the readibility of the various tags. For example, the tag JJ for adjective is replaced by the tag AJO for general adjective.

So far we have been looking at the tagging schemes employed by five English language corpora. Turning away from those for a moment and shifting our attention to the German and the Dutch corpora, we see that the tagging schemes employed in these corpora do not differ all that much from what we have already seen with the English corpora. Again 'flat' tagging schemes are found, while the wordclass categories that are distinguished largely coincide with the English wordclasses. As with the tagging schemes for the English language corpora, the tagging schemes for the German and Dutch corpora each have their own degree of detail.

The 12-million-word Mannheim Corpus was tagged automatically for both wordclass information and syntactic information by means of the SATAN parser[7]. Neither level of analysis includes much detail. The wordclass tagging is rather rudimentary and includes only the most basic wordclass information needed for the syntactic analysis. Only occasionally information is added about prepositions that occur as collocates of other words and about the case that prepositions require for their complements.

The tagging of the relatively small Eindhoven Corpus is rather detailed. The tags consist of three digit codes, where the first digit indicates the wordclass, the second digit supplies information on the subclass, and the third digit carries additional information of various kinds, such as verb aspect, person, number, etc.

Returning now to the tagging schemes employed for the tagging of English language corpora, we find that in the case of the three remaining tagging schemes under consideration – the SUSANNE tagging scheme, the ICE tagset, and the Tosca tagset – they have opted for more hierarchically structured tagging schemes. Each of these tagging schemes encodes highly detailed wordclass information in a systematic fashion by introducing some sort of 'additional feature(s)' slots in their tags. The SUSANNE tagging scheme distinguishes as many as 352 distinct tags. Closer examination of these tags however learns us

7. The Mannheim Corpus has been compiled by the Institut für Deutsche Sprache (IDS), the German national institute for research on the German language. The SATAN parser was developed at the University of Saarbrücken for the purpose of machine translation.

that at the basis of these 352 items there are some 70 major wordclasses, while the addition of more featurelike information such as number, person, case, etc. leads to a total of 352 tags. Something similar appears to be the case when we consider the Tosca tagset and its less detailed derivative, the ICE tagset. Again tags for major wordclass categories have been extended so as to include additional feature information.

In summary, having compared the different tagging schemes that have been or are being employed in the tagging of wordclass information in ten different corpora, we find that

- there appears to be some consensus as to what to tag as far as wordclass information is concerned;
- even the format of the tags in different tagging schemes does not differ all that much;
- the kind of information does not vary too much from one language to another, in other words it appears feasible to have a tagset that could be applied cross-linguistically.

As to the last point, we must observe that the idea of having a tagset that can be applied cross-linguistically is not at all new. For example, the tagset employed in the multi-lingual ESPRIT-860 project [Kesselheim, 1986] was intended to comprise all the tags that would be required for the tagging of wordclass information in each of the EC languages. Since this meant that not only Germanic languages were included (Dutch, English, German), but also Romance languages (French, Italian, Spanish) and Greek, the tagset could be expected to be truly cross-linguistic. Unfortunately, however, examination of the tagset shows that it is more of a collection of various tags that any of the languages might at some stage require, rather than a thoroughly designed minimal interlingual (or cross-linguistic) tagset. It is therefore all the more surprising to find that certain wordclasses in German, Dutch and English are not accounted for[8].

8. For example, there appears to be no wordclass tag for interjections for German and Dutch, nor is there a tag for particles in English or in Dutch.

Part III
Criteria for a minimal tagset

Having compared the tagsets of ten prominent corpora we now arrive at a point where we should reflect a little on the criteria that will have to apply to a tagset to make it cross-linguistic and generally applicable. If we take a closer look at the results of our study described above we can conclude that there are mainly three types of criteria in relation to the development of a tagset; criteria of a linguistic nature, criteria for the data format of the labels and terminological criteria for the label names. Therefore the following points deserve special attention:

- coverage of major word classes
- addition of relevant feature information
- format of the tags
- representation of the labels

Let us begin with the linguistic criteria. The major word classes should minimally be covered by the tagset. During our projection of the different tagsets we came to the conclusion that there are 12 main word classes[9] that can be distinguished. As we said earlier, it became obvious that across the different tagsets variation as far as the main classes are concerned was not very large. The same was true even for the two non-English tagsets.
The specification of feature information, however, turned out to be more problematic than the determination of word classes, because the tagsets differed strongly in their degree of detail and/or their research context. From a linguistic point of view, there is a great variety of feature information. First of all, there are subclassifications of the major word classes, such as the distinction between common and proper nouns or the different types of pronoun. Furthermore, in relation to verbs, we can think of additional feature information about the degree of subcategorization (transitivity). Although this appeared not to be a very common feature in the tagsets examined, it can be of great importance to grammar-based syntactic corpus analysis[10]. Another relevant linguistic criterion is to facilitate the inclusion of morphological information, at least

9. see following section
10. see Oostdijk (1991)

for verbs, nouns and adjectives. Features containing this information are for example: number (singular, plural), person (first, second, third), gender (masculine, feminine, neuter), degree of comparison (positive, comparative, superlative) and, especially for the analysis of German texts, case information (nominative, genitive, dative, accustive). From the tagsets under survey it appeared that morphological information is desirable, but there seems to be no agreement on what sort of information should be added. This leads us to a following criterion that is important for the development of our tagset: the format of the tags.

From our study it has become clear that there is some degree of agreement on major word class information and to some extent also on their subclassification. However, with respect to additional feature information we can see a sort of growing disagreement as the detail of specification increases. This has to do with the various purposes people may have in relation to the enrichment of corpora. On the other hand language-specific items play a role in this context too, for example prepositions in postposition in Dutch[11]. So the *format* of the tags has to account for two major aspects:

- hierarchical structuring of information
- flexibility in relation to the encoding of special features and/or language specific items

Hierarchical structuring implies the ordering of information within a specific label. From left to right the degree of detail will increase, so that at the beginning we will find the major word class label, followed by a subclassification, and then several additional features.
With *flexibility* we mean that the label format should be open so that no researcher is limited by our tagset in adding special features that, according to his opinion and/or research aims, are useful or even essential. On the other hand researchers who want to make use of basic word classes only, need only concern themselves with parts of the tags and can ignore the other parts.
In our attempt to develop a cross-linguistic tagset we make use of a hierarchical data-field oriented coding scheme. The hierarchy in the labels is represented as follows: the level of detail increases from left to right and the different entries are separated by one or more unique delimiters. This way of coding also enables researchers to convert the format of the labels in a relatively easy way for their individual needs. For example in the ICE-project, where the tags form the input of a two-level grammar, labels with the format described above can be automatically converted into two-level tags. So this way of coding seems to be attractive linguistically (levels of description) as well as formally (interchangeability).

Finally we have to determine how the labels should be represented. Generally we can distinguish between two ways of coding - either a completely numeric label or a mnemonic letter-digit sequence. Although for reasons of readability mnemonic labels are preferable, numeric labels can be used as well, since it is not too difficult to transform one well-defined form into another. The advantage of numeric labels is that they are relatively compact and therefore can be stored more efficiently.
Focussing on the mnemonic way of coding, the question is: what terminology can best be used for the labels. From the tagsets examined we can conclude that there is a commonly accepted linguistic terminology with respect to word-class information, also from a cross-linguistic point of view. In order to provide codes as mnemonic as possible we are of the opinion that this terminology should also be included in our tagset.

In the following section we present a first step towards a cross-linguistic tagset as a result of our comparative study and illustrate the different major word classes, their possible subclassification, additional feature information as well as the way in which this information can be coded.

Part IV
A basic tagset for Germanic languages

In this section we present a basic cross-linguistic tagset for Germanic languages. Those familiar with the coding scheme adopted in the ESPRIT-860 project will find a number of similarities with this scheme. Thus, as we mentioned above, we have adopted the idea of a hierarchical data-field oriented coding scheme. Moreover the scheme allows for over-

11. Hij loopt het bos *in*. (engl.: He runs into the forest.)

Word Class Categories	
noun	adverb
pronoun	preposition
article	conjunction
adjective	particle
numeral	interjection
verb	formulaic expression

Table 2: word classes

as well as underspecification[12]. Unlike the ESPRIT tagging scheme, however, it is strictly datafield oriented in order to allow automatic format and addition of extra feature information. The different datafields are always separated by a # symbol, which functions as a delimiter. The hierarchical structure of the tags becomes clear when one looks at the examples we give in order to illustrate the tagging of the different wordclasses. In addition to what has been described above, the tagset also includes so-called ditto-tags which make it possible to account for lexical items that consist of more than one word, such as compound nouns or complex prepositions. Ditto-tags take the form of two numbers separated by a slash, where the first number indicates the current part of a compound item, while the second number indicates the total number of words that make up the compound. For example, the three word complex preposition *in spite of* is tagged as follows:

in PRP#...#1/3, spite PRP#...#2/3, of PRP#...#3/3.

Our basic tagset for the encoding of word class information in corpora comprises twelve major word class categories and an additional category which is intended to accommodate language-specific items. The twelve major categories that are distinguished are listed in Table 2.

The additional category we shall refer to as *open*. Each of the word class categories is discussed below.

Perhaps the categories *particle*, *interjection* and *formulaic expression* are not so familiar to some readers. For this reason we give a more detailed description when presenting them.

1. Noun

   Wordclass N
   Subclass com (common)
   prop (proper)

   Additional info:
   number sg (singular)
   plu (plural)
   gender masc (masculine)
   fem (feminine)
   neut (neuter)
   case nom (nominative)
   gen (genitive)
   dat (dative)
   acc (accusative)
   compound 1/2, 2/2, etc.

   Two examples:

   N#com#plu#

   refers to a single plural common noun, for example *house*.

   N#com#plu#acc#fem

   refers to the same word class as above but contains extra case (accusative) and gender (feminine) information, for example German *Häuser* [13]

2. Pronoun

   Wordclass PN
   Subclass per (personal)
   pos (possessive)
   ref (reflexive)
   dem (demonstrative)
   int (interrogative)
   rel (relative)
   ind (indefinite)
   rec (reciprocal)
   Additional info:
   number sg (singular)
   plu (plural)
   person 1 (first)
   2 (second)
   3 (third)
   gender masc (masculine)
   fem (feminine)
   neut (neuter)
   case nom (nominative)
   gen (genitive)
   dat (dative)
   acc (accusative)

12. see Kesselheim (1986)

13. engl. *houses*

compound 1/2, 2/2, etc.

For example:

PN#per#sg#1#nom

refers to a first person singular nominative personal pronoun (*I*).

PN#rec#1/2#

refers to the first part of a reciprocal pronoun, for example each *other*.

3. Article

Wordclass ART
Subclass def (definite)
ind (indefinite)

In addition, features such as number, case and gender can be added.
For example:

ART#def#sg#acc#masc#

refers to a definite singular accusative masculine article, e.g. german *den*.

4. Adjective

Wordclass ADJ
Degree pos (positive)
com (comparative)
sup (superlative)

Subclassification of adjectives seems to be a complicated matter since many of the subdivisions we found are determined by the research context. For example, we found attributive adjectives (main, chief), nominal adjectives and even semantically superlative adjectives. So at this stage we are not able to provide some consistent subclassification.

Again, additional features such as number, case, gender, ditto (as described above) and form (for example English: -ed, -ing) can be added, for example:

ADJ#gen#pos#acc#masc#

refers to a definite general positive accusative masculine adjective, for example German *intelligenten*

5. Numeral

Wordclass NUM
Subclass ord (ordinal)
crd (cardinal)
frac (fractional)

Features like number and case can be added.
For example:

NUM#ord#

refers to an ordinal numeral, for example *fifth*.

6. Verb

Wordclass V
Subclass 1 (lexical)
a (auxiliary)
Form inf (infinitive)
part (participle
Tense pres (present)
past (past)
Mood ind (indicative)
subj (subjunctive)
Subcat different forms of transitivity such as intransitive, copular, transitive, etc.

Again features as number, person and compounding can be added.
For example:

V#1#3#past#

refers to a lexical verb (third person, past tense), for example *went*.

7. Adverb

Wordclass ADV
Degree pos (positive)
com (comparative)
sup (superlative)

Features for compounding can be added. A general subclassification of adverbs is very hard to establish because of the semantic nature of such subdivisions.
For example:

ADV#pos#

refers to a positive adverb, for example *early*

8. Preposition

Wordclass PRP
Subclass conj (conjunctive)
adv (adverbial)
post (postposition, for Dutch)
phra (phrasal)
gen (general)

Features indicating the case a preposition requires for its complement, as well

as ones for compounding can be added. For example:

PRP#phra#dat#

refers to a phrasal preposition (required by a prepositional verb) that combines with a complement with dative case.

9. Conjunction (incl. connectives)

   Wordclass CON
   Subclass sub (subordinating)
   cor (coordinating)
   con (connective)

   Also, ditto-tags can be added. For example:

   CON#sub#

   refers to a subordinating conjunction, for example *because*

10. Particle, Interjection, Formula

    Wordclass PRT (particle)
    INT (interjection)
    FOR (formulaic expression)

    Interjections are normally referred to as words that do not enter into syntactic relations and that do not have a clear morphological structure. Very often they are of a onomatopoeic nature. Examples of interjections are: *aha, hm, wow, psst, oops.*
    Formulaic Expressions are fixed expressions used as formulaic reactions in a certain dialogue contexts. Examples are: *all the best; excuse me; dank u wel; Danke, gut.*
    Particles are morphologically fixed words that do not belong to any of the word classes described above and that can function in many ways in a sentence, for example as introducing element of the subject of an infinitival clause (for example: I am waiting *for* the meeting to begin), or they function as fixed answers to questions (for example: *yes, no, ja*[14]
    Ditto-tags can be applied to the different elements of the tagged item.
    For example: *Good* FOR#1/2# *Morning* FOR#2/2#

11. Open
    Wordclass : O (open)
    The subclasses to be distinguished within this wordclass category may vary, depending on the specific language the tagset is used for. For English the genitive marker belongs in this category; the same goes for the German verb particle. For example:

    O#GM#

    refers to a genitive marker.

14. for more detailed information about particles, see Engel (1988)

Part V
Conclusion

In this paper we have sketched the way in which linguistic enrichment of corpora could be standardized. We have reported on our efforts to standardize the word class tags. In addition, we compared the tag sets of a number of prominent corpora. The differences between these sets encouraged us to proceed towards a standardized cross-linguistic tagset. This set could contribute to improved access and exchange of analyzed corpora. In addition to a standardized tagset it might be interesting to determine if and how a standard annotation of linguistic information on higher levels of description (syntax, semantics, pragmatics) can be established.

We are working on these issues and we hope to encourage other (academic and industrial) researchers in the field of corpus linguistics to participate in the discussion about common guidelines for the linguistic annotation of corpora in the future.

Part VI
References

- Aarts, J. and Th. van den Heuvel (1985): *Computational tools for the syntactic analysis of corpora*, in: Linguistics 23: 303-35.
- Black, E., R. Garside and G. Leech (1987): *Statistically-Driven Computer Grammars of English: The IBM-Lancaster Approach*. Internal report.

- Boogaart, P.C. uit den (1975): *Woordfrequenties.* Utrecht: Oosthoek, Scheltema & Holkema.
- Engel, U. (1988): Deutsche Grammatik. Heidelberg: Julius Groos Verlag.
- Greenbaum, S. (1991): *The development of the International Corpus of English.* in: Aijmer, K. and B. Altenberg (1991): English Corpus Linguistics: 83-91. London: Longman.
- Greenbaum, S. (1992): *The ICE Tagset Manual.* London: University College London.
- Johansson, S. et al. (1978): *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with digital Computers.* Oslo: Dept. of English, University of Oslo.
- Kesselheim (1986): *Coding for word-classes,* in: ESPRIT-860 Final report [BU-WKL-0576].
- Kučera, H. and W.N. Francis (1967): *Computational Analysis of Present-Day American English.* Providence, Rhode Island: Brown University Press.
- Leech, G. (1993): *100 million words of English,* in: English Today 33: 9-15.
- Marcus, M.P. and B. Santorini (1991): *Building Very Large Natural Language Corpora: The Penn Treebank.* CIS-report: University of Pennsylvania.
- Neumann, R. (1987): *Die grammatische Erschliessung der Mannheimer Korpora.* Mannheim: Institut fuer Deutsche Sprache. Internal report.
- Oostdijk, N. (1991): *Corpus Linguistics and the Automatic Analysis of English.* Amsterdam: Rodopi.
- Sampson, G. (1992): *The Susanne Corpus.* E-mail report.
- Sampson, G. (forthcoming): *English for the Computer.* Oxford: Oxford University Press.
- Santorini, B. (1991): *Part-of-Speech Tagging Guidelines for the Penn Treebank Project.* Internal report.
- TEI (1991): *List of Comon Morphological Features.* TEI document A11W2.