Authors

This volume contains the Final Report of the Feasibility Study for a "Network of European Reference Corpora", a project sponsored by the DG XIII of the Commission of the European Union.
The project was carried out between the years 1990 and 1994 by a Consortium composed of the following members:
PSA: the Pisa group, comprising the Istituto di Linguistica Computazionale del CNR, the Linguistics Department of the University of Pisa, and the Consorzio Pisa Ricerche.
BIR: School of English - The University of Birmingham
MAN: Institut für Deutsche Sprache, Mannheim
LEI: Instituut voor Nederlandse Lexicologie, Leiden
MAL: Departamento de Filología Española, Facultad de Filosofía y Letras - Universidad de Málaga
PAR: Institut National de la Langue Française - INALF - CNRS, France

The Introduction to the volume has been written by A.Zampolli.
The Initial Chapter (0), which is based on the recommendations of all the NERC partners, has been drafted by J. Sinclair and A. Zampolli, with the assistance of P. Lafon, P.G.J. van Sterkenburg, W. Teubert. The Appendix on the Lexicon has been drafted by N. Calzolari and A. Zampolli.
We list here, for each chapter (or, where relevant, for each part of a chapter) the title, the author(s), and, in parentheses, the short name of the partner responsible for co-ordinating the work.

Ch.	Title, author(s) and responsible partner
1	User needs: W. Teubert (MAN)
2	Corpus Design Criteria: J.G. Kruyt, P.G.J. van Sterkenburg (LEI)
3	Text Representation: Written Language: P. Lafon, D. Vignaud (PAR) Text Representation: Spoken Language: J.M. Sinclair (BIR) (Appendixes: J.P. French)
4	Text Acquisition and Reusability: P. Lafon, F. Chahuneau (PAR) Access and Management Software Tools: J.M. Sinclair (BIR)
5	Linguistic Annotation of Texts: scientific and technical problems; guidelines for harmonization: N. Calzolari, M. Monachini, S. Montemagni, J. Sinclair, A. Spanu, W. Teubert, A. Zampolli (PSA)
6	Corpus Annotation Tools: J.M. Sinclair (BIR)
7	Knowledge Extraction: V. Pirrelli (PSA)

The editorial committee was assisted by M. Monachini and P. Orsolini. In particular, P. Orsolini made the final formatting adjustments to the entire book.

Chapter 5

Linguistic Annotation of Texts: scientific and technical problems; guidelines for harmonization

1 Introduction

1.1 *The concept of "annotation"*

In the current terminological use we distinguish between: i) a "raw" text, consisting of the electronic conversion of the original text into machine readable form (MRF); ii) an "annotated" text, also including some level(s) of linguistic description (e.g. parts of speech, immediate constituent bracketing, syntactic tree-structure, etc.).

The above distinction presents some borderline cases. In a sense, some interventions made during the pre-editing phase or during the capture of texts in MRF are already a form of annotation (for example, capitals indicating proper names vs. other capitals; disambiguation of the full stop sign (abbreviations, punctuation, etc.); identification of foreign or quoted words). For obvious reasons, the borderline is even less clear in the case of MR versions of spoken texts, where the original is not a canonical printed text but a transcription of speech. The transcription can consist of a detailed phonetic or phonological representation of speech, with or without an indication of prosodic elements (intonations, stress, expiration units, etc.). This already offers some type of "annotation". In a conventional orthographic version, transcription can be with or without an indication of elements such as pauses, repetitions, restarts, self-corrections, overlapping, etc. Both types of transcription can be done with or without normalization with reference to a standard linguistic model. However, apart from the borderline cases, the basic concept is clear: we shall use the term "annotated" to indicate a corpus with a systematic encoded representation of linguistic categories at a certain level of linguistic description and, in some cases, of their (structural) relationships.

An annotation scheme has two components: i) the set of annotation symbols (form) with a definition of their meaning (content), and ii) the guidelines for application.

1.2 *Present situation*

The majority of corpora, already collected or in progress, are "raw" corpora. Very few corpora have been annotated, but the number of annotated corpora is constantly increasing. This trend has been particularly strong in recent months and is expected to continue - obviously at different levels of speed and detail for different types of annotation. It will be influenced by the ever growing availability of more reliable and refined methods, strategies and tools (for which see Chapter 6 on Annotation Tools). We can distinguish the following main categories:

i) "Tagged" corpora: a (simple or complex) code is assigned to each word, representing grammatical information: usually, parts of speech and inflectional or morphological

categories (person, gender, number, etc.).

ii) "Lemmatized" corpora: each textual word also receives an indication of its lemma (e.g. the infinitive for verbs, the masculine singular for adjectives, etc.). A lemma is an arbitrarily chosen canonical form, under which word forms are grouped together as instances of the same headword. A lemmatized corpus is often, if not always, preferable when working on heavily inflected languages like Italian, in order to limit the dispersion of information on inflected forms[1].

iii) "Analyzed" corpora: information about "higher level" analysis is included, e.g. brackets identifying phrases of various types (nominal, prepositional groups, etc.); labelled parse-trees, etc. Analyses can be performed at different levels of linguistic description: surface syntax; deep syntax; word semantic features; semantic structures; discourse structure; pragmatics; etc.

1.3 *Tagged Corpora*

Virtually all NLP systems begin the process of analysis by classifying - i.e. tagging - the textual words of the input sentences. The tagging procedure usually consists of two logical steps:

i) look-up in a computational lexicon, and assignment to each textual word of the tag(s) provided by the lexicon;

ii) in cases where the lexicon lists more than one possible tag per word, resolution of the ambiguity.

Automatic tagging usually requires:

- a large computational lexicon;
- procedures to recognize or at least "guess" the relevant tags for "new" words;
- procedures to disambiguate grammatically ambiguous words.

Disambiguating procedures exist for English and for other languages (Italian, Spanish, German etc.). We can distinguish two main types of procedures:

a) Local, rule-based procedures which try to disambiguate by searching, in the immediate context, for specific patterns of grammatical categories which are or are not allowed to occur with each of the potential grammatical descriptions suggested for the ambiguous word.

b) Statistical procedures based on the transitional probabilities of n-consecutive grammatical descriptions preceding - or following - the ambiguous word. These procedures are usually "trained" on previously tagged corpora. The success rate reported varies between 60% and 97%, according to the language, the complexity of the tagging systems, the sublanguage to be tagged, etc.

No commonly agreed "tagging scheme" (i.e. a list of tags and a set of criteria to be applied in

[1] The advantages and disadvantages of working on a lemmatized corpus depending on different uses and purposes are discussed in (Bindi et al., 1991, NERC-103, and Bindi et al., forthcoming, NERC-177). The study of the lemmatization process is treated in (Panhuijsen et al., 1992 NERC-76).

controversial cases) yet exists, but a growing move towards convergence can certainly be noted.

1.4 *Analyzed Corpora*

1.4.1 *Syntax*

In traditional NLP systems, a syntactic component basically performs two functions:
i) to determine the syntactic structure of the input sentence (e.g. identifying the various clauses);
ii) to "regularize" the syntactic structure. Various types of structures are mapped onto a smaller number of simple canonical structures, thus simplifying subsequent processing. These structures are often intended to represent the functional relationships among the various phrases within a sentence.

In a stratificational approach, the parser produces two (or more) distinct levels of representation, namely:
- a surface or configurational syntax level,
- a deep syntax or logical form level.

In the current practice of corpus research, there are rather few examples of syntactic annotation, and these are usually at the surface level.
The term **parsing scheme** is now widely used in corpus linguistics to indicate a precise and complete definition of:
- the range of structures and categories used in parsing the corpus;
- which, among the various analyses, are considered as correct for any construction.

In exploring whether it is possible to design a common parsing scheme, we must take into account the following facts:
(a) for decades, theoretical linguistics has been concerned mainly with rival notational and explanatory models for capturing highly abstract generalizations;
(b) linguists have focussed on a limited range of phenomena and constructions selected by the research community as posing "interesting problems", relying on data obtained by introspection (i.e. provided by their personal competence as native speakers). As a consequence, linguistic theories do not generally provide a parsing scheme of sufficient coverage to cope with the language of real texts.

Even though automatic parsing has been a central issue in computational linguistics for many years, the following comments still apply:
- the definition of target analysis schemes and the extension of the linguistic coverage of parsers have not tended (with few exceptions) to be high priority tasks;
- a general agreement about the analysis the parser must provide has not been pursued, and, as a consequence, a commonly agreed parsing scheme does not exist;
- adequate parsers (i.e. parsers sufficiently "robust" to be applicable to "real-life" texts as found in a corpus) still do not exist. Particular attention must be paid to minimizing the effort and time required to train human operators to intervene in those cases in which the parser fails to operate.

1.4.2 *Semantics and Pragmatics*

The main tasks of semantic components in NLP are:
- to disambiguate ambiguous syntactic structures;
- to disambiguate homographic/polysemic words;
- to determine the general "meaning of a sentence".

The structure produced by the syntactic component is usually mapped onto a formal language, which is designed to be unambiguous and to have simple rules for interpretation and inference. In practical systems, the "meaning" of a sentence is, roughly speaking, what we want the system to do in response to our input, i.e. to retrieve data, direct robots, etc.

Disambiguating and interpreting a sentence requires more than just linguistic knowledge. It also involves accessing knowledge of the world, general and/or domain-specific, and of the specific characteristic of the communicative context (dialogue, etc.). The distinction between linguistic and pragmatic knowledge is known to be very difficult.

Current research into semantic and pragmatic analyses is not advanced, except for very restricted ad hoc NLP applications. It is only in the past years that some groups and projects have begun to work towards annotating corpora at the semantic level. Semantic annotation of words or phrases can be used, for instance, for the application of selectional restriction constraints or preference mechanisms (e.g. a verb can be "restricted" with respect to the range of items it can accept as subjects, objects, etc. In the case of competing analyses, a structure is accepted/rejected if the proposed subject/object is/is not a member of the accepted class).

1.5 *The need for annotated corpora in NLP and Lexicography*

The shortage of annotated corpora (and in particular of analyzed corpora) is not due to a lack of potential users, but to severe methodological and practical problems. Methodological problems include the inadequacy or lack of annotation schemes applicable to a real corpus; practical problems include the cost and time of manual annotation and the inadequacy of existing parsers which are not robust enough for real corpora. In fact, to extract the relevant information from a corpus, the majority of users need to perform some kind of linguistic analysis. But, very often, due to the above mentioned difficulties:

i) the analysis is performed only "mentally" and no record of the results is left in the form of annotations in the corpus. The results are therefore not reusable, and the analysis must be performed again by subsequent users;
ii) the size of the sample, the completeness and the systematicity of the analysis are drastically reduced, and the full potential of the corpus as a source of information is exploited only in a limited and inadequate way.

A linguist can work on the corpus as a source of "raw" data, and he can apply his techniques of analysis to this data. However, in order to use the categories and structures he has recognized in the corpus (e.g. to extract examples, to infer regularities, to discover new patterns, etc.) he has to be able to reuse the first order analysis, browsing and navigating through the annotated corpus, applying pattern matching or statistical procedures also on the tags, searching for co-occurrences, regularities, sorting the data according to categories, etc.

Developers of NLP systems need to use annotated corpora for several reasons, e.g.:

- to count frequencies of categories, contextual patterns, structures; to compute transitional probabilities; to create statistically-based taggers and parsers, or to complement rule-based parsers with statistical knowledge;
- to discover structures not covered or solved by the parser/grammar, and to evaluate their statistical relevance;
- to use statistical procedures in order to uncover significant co-occurrences (collocations, idioms, etc.), to enrich the computational lexicon;
- to extract categorial and structural data characterizing a given domain or sublanguage;
- to correlate structures and categories of different levels (e.g. syntactic structures and intonational patterns), etc.

The more extended and intensive analyses of corpora are performed by lexicographers, who usually analyze the contexts in which a word occurs in order to create homogeneous groupings on which to base the subdivision of a dictionary entry into different meanings, etc. Lexicographers usually limit themselves to inserting under the appropriate section of the entry some selected examples, without using their classification of the contexts, in which they have already invested a great deal of effort, to annotate the concordances and/or the corpus. Some lexicographers have now started to spread the idea that a reusable lexical knowledge base, intended as a general source from which to extract different types of lexicographical products (concise, pocket, specialized, collegiate, bilingual, learner's dictionaries), must include not only a set of entries, with the relevant linguistic information, but also an annotated corpus, where the words are linked to the relevant sections of the correspondent entry.

Annotations done by lexicographers can be immediately reused by computational linguists. Similarly, a corpus annotated by a linguist or a computational linguist provides lexicographers with distinctions, based on theoretical principles, which would otherwise escape the lexicographer[2]. Furthermore, an annotated corpus can offer the lexicographer the possibility of including in the dictionary notations on frequencies of use (in both general language and in sublanguages) of various meanings, constructions, collocates of the entry, etc.

1.6 *The feasibility of a shared annotation scheme: the methodology adopted in this study*

In the scientific community, there are clearly two distinct positions with regards to the annotation of corpora. Some researchers believe that it is highly unlikely that a commonly agreed tagging/parsing scheme would satisfy the needs of various users of corpora, and also that a theory-neutral tagging/parsing scheme is not feasible. As a consequence, they suggest that, instead of investing a great deal of effort in annotating a corpus, we should concentrate on creating flexible and powerful tagging/parsing software, leaving each researcher free to devise his own scheme according to his own definition of the relevant linguistic rules. In particular, they suggest that human effort should not be spent on annotating ambiguities or difficulties that cannot be solved by an automatic tagger/parser. Other researchers feel that it is necessary to:

[2] LRE DELIS is a project aiming, among others, at defining lexical specifications based on the analysis of a carefully annotated corpus syntactically and semantically).

- try to define a commonly agreed tagging/parsing scheme;
- annotate carefully selected subsets of corpora on the basis of this scheme;
- try to reduce costs and improve the results of the taggers/parsers, combining an automatic tool with carefully optimized human interventions.

Taking urgent user demands for annotated corpora into account, the NERC feasibility study tried to assess if, to what extent, and for which linguistic levels it is possible to conceive a commonly agreed multifunctional annotation scheme, i.e. such that various categories of users may derive, through appropriate interfaces, from the annotation supplied by corpus developers, (at least part of) the linguistic information they need. Given the fact that corpora are widely recognized by the research and language industry communities as essential, shareable and reusable[3] resources, standardization in this field has become an issue of vital importance[4].

This study takes into account current practices as well as the specific needs of different types of users (and in particular: the linguistic nature and content of the required annotation, priorities in terms of annotation content and of text-type (subsets) to be annotated, optimal/minimal size of the sample, acceptability of different degrees of accuracy of annotation). An attempt is made to assess at what level current schemes overlap, and whether it is possible to identify at least a "core" set of linguistic phenomena which are commonly recognized by the various users and for which the design of a commonly agreed annotation scheme is conceivable, for NERC internal use only, or also for the use of a broader community of corpus developers and users.

This involves:

- a comparative survey of existing practices, both in corpora annotation and in some NLP systems; consultation with national and international projects on corpora; cooperation with projects dealing with problems of theory-neutral, reusable linguistic resources (e.g. EEC projects on reusability of lexical and grammatical resources);
- a detailed analysis, based on the preceding survey, of the various points of agreement and disagreement for each linguistic level.

Storage of and access to annotated information have not been dealt with in this part, but in Chapter 6 on Annotation Tools. Chapter 6 deals with issues such as: whether annotation is to be inserted in the text or stored in separate files; methods for aligning the texts and the various levels of annotation; relationships with the formalisms proposed by the TEI; typology and functions of access by various classes of users (both human and programs) to various levels of information (e.g. interrogation and browsing of tree-structures).

In the following sections, we report the main results emerging from the study at the levels of phonological, morphosyntactic, syntactic, and semantic/pragmatic annotation. At the end of each section we give a condensed summary of the main recommendations emerging from each part of the study, together with an indication of further directions for future work.

[3] The concept of "reusability", which came out at the Grosseto Workshop as one of the recommendations, has become crucial as far as large linguistic resources are concerned (Calzolari and Zampolli, 1990).

[4] EAGLES (Expert Advisory Group on Language Engineering Standards), launched by the European Community, DG XIII, in the framework of the LRE projects, in order to deal with the issue of standardization has a group dealing with corpora, which is working "towards the achievement of a proposal for operational standards" (EAGLES - Workplan, 1992).

2 Phonetic/Phonemic and Prosodic Annotation

2.1 *Introduction*

Whereas for written texts there is a clear and distinct dividing line between the concept of text representation and the concept of annotation, the distinction is not so clear for spoken texts. Any kind of transcription includes coding, i.e. adding linguistic information that is not present in the original soundwave. Even orthographic transcription involves the disambiguation of homophones, and the prosodic information in the soundwave is processed into some linguistically-based rendering of sentence and clausal structures.

Discussions among members of the NERC consortium have led to an understanding shared by all members that text representation of the spoken language refers to orthographic transcriptions of the original soundwave (see Chapter 3 part B.). After careful analysis, the NERC consortium has decided to recommend the Transcription Conventions developed by J.P. French (1992, NERC-50), and in particular the level two transcription rules, for orthographic transcripts. These Transcription Conventions are, on the whole, compatible with the TEI Guidelines but are easier to interpret by readers, since they separate the text from any header-type material. Of course, a minimum amount of information on extralinguistic features about speakers, setting and technical specifications will also have to be documented in the case of orthographic transcriptions.

But orthographic transcription does not represent the phonetic or phonemic values used by the individual speaker. Whilst we recommend that orthographic transcription should include mark-up of pauses and overlaps, we recognize that it does not represent intonation, prosody, stress, pitch and many more paralinguistic features such as hesitations, interruptions, gestures etc. There is a long tradition in linguistics of dealing with such features and successful attempts at standardization have been made even before the emergence of corpus linguistics (cf. the IPA alphabet). Phonology and, to some extent, dialectology depend on the existence of coding systems for these features. Anyone interested in the phonetic/phonemic and prosodic values of recorded spoken language needs more than an orthographic transcription. The NERC consortium has therefore decided to deal with such coding systems within the framework of the chapter on linguistic annotation schemes.

2.1.2 *Recent developments*

When the work packages of the NERC feasibility study were defined (December 1990), it was still common among linguists and in the speech community to keep the soundwave of a recording on analogous tapes. Therefore, instantaneous (real time) access to specific occurrences was not possible. Phoneticians and members of the speech community alike had to work with transcripts, and the more interested they were in phonetic or prosodic features, the narrower the transcriptions they used had to be. Phonetic and prosodic transcriptions are extremely expensive to produce, and therefore at that time speech research was concerned with areas where relatively small quantities of spoken language had to be analyzed. At the time, larger corpora of spoken language were not a major concern in speech research.

But things changed quickly. The speech community stopped working with analogous

recordings; instead they stored the digitized soundwave on CD-ROMs (or on hard discs) and thus were able to create instantaneous or real-time access to the original sound occurrence. Thus it became superfluous to study phonetic or prosodic features on the basis of narrow transcriptions. Using standard computer networks, the original sound occurrences are now available everywhere and to everyone. Transcripts are needed only insofar as they can be used to mark and identify the individual occurrence, after they have been aligned with the soundwave. Orthographic transcriptions are now entirely sufficient. Only in very few cases today is speech research still concerned with narrow transcriptions. Standardization, therefore, is a less pressing issue than it was in 1990.

Recent technical advances have also made it possible to automatically align orthographic transcripts with the original soundwave. For high quality recordings, this has already been demonstrated for English (e.g. by Roger Moore), and the development of freely available, pre-competitive, robust alignment software has been commissioned by the Linguistic Data Consortium, in US in 1992. Due to its modular design, it will also be possible to adapt this software for other languages (by processing existing pronunciation dictionaries).

As a consequence, the speech community has started to express an interest in large spoken language corpora. Even general purpose corpora of impromptu, unrehearsed, unscripted, unelicited informal conversations now seem to arouse some interest in the speech research community as such corpora can be used as test-beds for speech recognition systems. The traditional kind of speech research corpus of elicited, very short stretches of a particular sublanguage in a strictly defined setting will no longer be narrowly transcribed, but accessed directly using the orthographic transcription as an index.

The NERC consortium has therefore re-assessed the envisaged provisions for the phonetic/phonemic and prosodic annotation of spoken language corpora. Instead of advocating strict standardization, it now seems more realistic to suggest certain well designed conventions that allow easy exchange of data. In some linguistic areas where working with digitized speech data is not yet the rule, e.g. in dialectology and the study of unscripted languages, such a suggestion might be too broad to meet the need for a very narrow phonetic transcription. But this kind of research is carried out in a predominantly academic and scholarly environment; and, in the coming years, working with digitized data will make phonetic transcriptions superfluous in those areas too.

2.1.3 *The State of the Art*

The technical state of the art, the needs of the speech community in terms of recording quality, digitization, spectrographic analysis, transcription levels, machinery, software and storage options are explored in (Payne, 1992, NERC-132).

This study has taken into consideration the contributions made by members of the speech community to the Pisa Workshop, 1991 (NERC-82) namely:

- John McNaught: User needs for textual corpora in natural language processing
- Roger K. Moore: User needs in speech research
- Stig Johanson, Lou Burnard, Jane Edwards, And Rosta: Text Encoding Initiative, Spoken text work group

In addition, six projects dealing with phonetic/phonemic annotation of spoken language were

analyzed in a report by (Scheiter, 1992, NERC-135). The six projects analyzed are:

- IBM Deutschland GmbH, Heidelberg Scientific Center/Speech Recognition in German, SPRING
- Institut für Phonetik und sprachliche Kommunikation der Universität Müchen/Phonetische Datenbank für gesprochenes Deutsch, PHONDAT
- Fakultät für Linguistik der Universität Bielefeld/Speech assessment Methodology, SAM (ESPRIT Project 2589: Multi-lingual speech input/output assessment, methodology and standardization)
- Institut für deutsche Sprache, Mannheim/Grunddeutsch - Pfeffer-Korpus (Basic German - Pfeffer-Corpus)
- Institut für deutsche Sprache, Mannheim/Schlichtungsgesprache (Mediation talks)
- Germanistisches Seminar der Universität Hamburg/ Die Entwicklung narrativer Diskursfähigkeiten im Deutschen und Türkischen im familiären und schulischen Kontext, ENDFAS (The development of German and Turkish narrative discourse skills in the family and at school)

Finally, a study by Jonathan Payne was commissioned (Payne, 1992, NERC-122). This report reflects the view held by the NERC consortium, namely that for text representation TEI conventions should be preferred wherever possible, that where TEI is cumbersome and difficult to implement or to read, TEI-compatible conventions should be employed, and that only in those instances where TEI is still inadequate or inferior, deviating but clearly defined (and therefore at least minimally compatible) conventions should be used. So far, the TEI guidelines have not offered an explicit analysis of different requirements for different levels of transcription, although there is some reference to fairly detailed transcriptions in the text.

As far as extralinguistic features are concerned (pauses, vocals, kinesics, events, writing), we suggest that each project should decide the level of specification possible in TEI. As for other extralinguistic features relevant to speakers and recording, the survey on textual data (Chapter 3, part B.) shows a consensus for at least the following categories: (speaker:) sex, age, region, dialect; (recording:) date, place, setting, recording technique.

For prosody, the TEI guidelines stress the 'paramount importance' of marking prosodic features 'in the absence of conventional punctuation', which, it seems, is to be avoided. However, the explicit provision within the guidelines for encoding prosody does not appear to be particularly well developed. Apart from pauses, there are two recommendations: (i) to use the <s> tag and (ii) to use the <shift> tag. As Payne shows, the <s> tag, as it is currently conceived, is not ideally suited for the recommended purpose of indicating tone units. Furthermore, within the TEI guidelines there is no clear distinction between the linguistic feature of tone unit and the paralinguistic feature of tonic unit, explained as 'shifts in voice quality', for which the <shift> tag is recommended.

The TEI proposals still suffer from two disadvantages. First, there has been no time to develop and modify them in response to experience. They should be tested in real practice (or better in a variety of practices) and the finalized recommendations should reflect this practical experience. Second, to ensure that TEI can be used as an exchange format between research institutions of different backgrounds, some proposals should be made as to what is to be encoded for which applications. For example, although there exists a mechanism for encoding quite subtle shifts in paralinguistic features, there is no straightforward proposal on how to encode prosodic

(as linguistic) features. Even if phonetic/phonemic and prosodic transcription today seems to constitute a less important issue than it did a few years ago, there are clear advantages to the user community in having a standard set of conventions for encoding spoken texts at this level. The TEI proposals will constitute a major move in this direction. For the time being, however, the NERC consortium agrees that, while the TEI conventions should certainly be taken into consideration, they should not be recommended as a standard.

2.1.4 *Recommendations*

For the annotation of phonetic/phonemic and prosodic features of spoken text corpora, the NERC consortium expects that final recommendations will be given in due course by EAGLES, taking into account the emerging trends in the phonological and the speech research communities. As with the representation of spoken texts, EAGLES will give further consideration to the establishment of common practice in this field of linguistic (and NLP) research.

In the meantime, the SAMPA (SAM Phonetic Alphabet, derived from the IPA alphabet according to computational requirements) and the SAMPROSA (SAM Prosodic Alphabet draft) are being suggested as conventions to be followed. They allow not only for a fairly broad phonetic transcription but also for the marking of the following features: local tone, global tone, terminal tone, nuclear tone, length, stress, pause, boundary etc. A more detailed presentation of the SAMPA and SAMPROSA conventions is contained in Scheiter, 1992, NERC-135.

Relevant NERC Papers

French J.P. (1992): "Transcription proposals: multi-level system", Working Paper, COBULD, Birmingham, NERC-50.

NERC Consortium (1992): "Workshop on Textual Corpora", 24-26 January 1992, Report from the Conference, Pisa, NERC-82.

Payne J. (1992): "Report on the Compatibility of JP French's Spoken Corpus Transcription Conventions with the TEI Guidelines for Transcription of Spoken Texts", Working Paper, COBUILD, Birmingham and IDS, Mannheim, NERC-122.

Payne J. (1992): "Speaking the Same Language? Listening to the Speech Community", Working Paper, COBUILD, Birmingham, NERC-132.

Scheiter S. (1992): "Text Representation and Annotation Schemes in German Language Corpora", Technical Report, IDS, Mannheim, NERC-135.

3 Morphosyntactic Annotation

3.1 *Introduction*

The aim of this section is to explore the feasibility of proposing, as a short-term objective, a minimal standard for annotation at the morphosyntactic level, and to offer a methodology for achieving a shareable scheme. The present proposal seeks to provide a starting-point for further discussions and developments within this area (to be carried out mainly by the EAGLES Working Group on Corpora) and is not to be considered as final.

This section summarizes the outcome of two phases of work conducted within NERC, a survey phase and a standardisation phase, both described in detail in (Monachini and Östling, 1992a, NERC-60 and 1992b, NERC-61).

3.2 *The Survey phase*

The survey phase consisted of a review and a comparison of existing coding schemes at the morphosyntactic level, taking into account different corpus annotation policies for a number of European languages. The tagsets were analyzed in order to recognize, classify, and compare the morphosyntactic information encoded by different annotation practices, starting from reality as manifested in corpora, in a bottom-up or data-driven approach.

The present work consisted of two steps: i) a detailed study, for each tagset, of the actual tags used for each morphological class, leading to the discovery and classification of the linguistic phenomena taken into account in the annotation of the different corpora; ii) the identification of the core features peculiar to each morphological class. The information was synthesized and organized in synoptical tables, which represent the morphological classes as feature sets. These tables give a graphic representation of the complexity of word classes: they list the features of a class and make explicit whether or not they are marked by the tagsets. The common/shared features in each table can be seen as providing a nucleus of a de-facto standard. This study shows that some morphological classes are treated in almost the same way by most tagsets: the delimitation of the class and the recognition of its features by the various tagsets converge. Other morphological classes, however, present difficulties, often due to delimitation problems and the different boundaries between the word classes, or to different theoretical approaches underlying the classification. These obviously need further consideration before an acceptable proposal can be arrived at.

The tagsets taken into consideration are as follows[5]:

Pe	American English	Penn Treebank (Santorini, 1991, Marcus and Santorini, 1992)

[5] Due to the absence of a morphosyntactically annotated Spanish corpus, no tagset could be analyzed. The requirements for an adequate description of Spanish morphosyntactic phenomena are presented according to data supplied by personal communication from (Blanco Rodriguez, 1992, NERC-112).

BNC	British English	British National Corpus (Burnard, pers. comm.)
Go	American English	Gothenburg Corpus (Ellegård, 1978)
Br	American English	Brown Corpus (Francis, 1980, Francis and Kucera, 1982)
LOB	British English	LOB Corpus (Johansson, 1986)
La	British English	Lancaster Corpus (Garside et al., 1987)
SUC	Swedish	Stockholm Umeå Corpus Project (Ejerhed et al., 1992)
It	Italian	ILC Corpus (Calzolari et al., 1983, Monachini, 1992)
FrS	French	Uppsala and Stockholm (Östling, 1987a, 1987b; Engwall, 1974, 1984)
Par	French	Institut National de la Langue Française (Lafon, 1992, NERC-72)
Eur	Italian	EUROTRA[6] (Copeland et al., 1991)
UDB	Dutch	Uit den Boogaart (Dutilh-Ruitenberg, 1992, NERC-69)
ETW	British English	ENGTWOL Lexicon Helsinki (Karlsson et al., forthcoming, based on the two-level morphology)
GER	German	FAZSIE Siemens/München Corpus (Scheiter, 1992, NERC-124)

3.2.1 *Description of the Procedure*

The main morphological classes (listed below) were chosen on the basis of the categories observed in the corpora. The morphosyntactic phenomena - represented by the features and marked by the tags - have been classified and listed under the relevant morphological classes. In some cases, trans-categorizations and the different strategies adopted by various annotation schemes for handling ambiguous entities complicate the comparison between the various tagging strategies. This is discussed further in the relevant sections.

Main morphological classes

Nouns
Adjectives (content words)
Pronouns and Pronominal Adjectives
Articles
Verbs
Adverbs
Numerals
Prepositions and Particles
Coordinating and Subordinating Conjunctions

[6] Since there is no list or manual describing the EUROTRA morphological features, the classes and features taken into account were deduced from the feature bundles on the ECS (Eurotra Constituent Structure) level, i.e. the syntactic surface level. At this level, the coverage of morphosyntactic phenomena is only partial, because, in EUROTRA, some phenomena (e.g. comparison) are taken into account on higher levels of linguistic analysis.

Interjections
Foreign words
Letters, Symbols and Formulae

Each category is described in (Monachini and Östling, 1992, NERC-60) by means of a table which lists its features and their values. The categories were identified by reference to existing corpora, as already specified above, and also by taking into account the proposal of the Text Encoding Initiative (TEI AI 1W2, 1991, NERC-14). The work of the TEI is in some ways similar to the present one in that it attempts to define word classes and identify a core of widely recognized features which are expressed morphologically in a number of modern European languages. The main difference lies in the approach adopted, the present work being corpus-based, while the TEI is based on the competence of linguists. There are some differences between the categories, features and values presented below and those defined by the TEI. More subtle distinctions marked in some tagsets, and considered important for the complete description of the categories, have also been taken into account in the tables.

3.2.2 *Organization of the Tables*

In the table headings, acronyms of the annotation schemes considered are used as listed above.

The left vertical column indicates:
- the category
- the features (in small capitals)
- the relevant values (listed under the feature and preceded by the sign -)
- possible sub-values (listed under the values and preceded by the sign *)
- other distinctions within the class in question

When a tagset recognizes a category and has labels corresponding to the values of a certain feature, this is marked in the tables with an X.

3.2.3 *Categories, Features and Values*

The following is a complete list of the features, values and sub-values used, and the categories to which they apply. It is clear that the values are not always mutually exclusive: there is some overlap. It must be stressed that each language system uses the values which are most appropriate for it.

We present afterwards, for illustrative purposes, the synoptical table describing the morphological class of Nouns, preceded by some remarks concerning the peculiarities of the tagsets considered. This will explain the method and the detail used in the review phase of the work.

Category: nouns
TYPE-N
- proper
- common

Categories: pronouns, pronominal adjectives
TYPE-PR
- personal
- reflexive
- possessive
  * pronoun
  * adjective
- interrogative
  * pronoun
  * adjective
- relative
  * pronoun
  * adjective
- demonstrative
  * pronoun
  * adjective
- indefinite
  * pronoun
  * adjective

Category: adverbs
TYPE-ADV
- lexical
- interrogative/relative

Category: numerals
TYPE-NUM
- cardinal
- ordinal

Category: preposition and particles
TYPE-PREP
- preposition
- postposition
- particle
- inf marker

Category: conjunctions
TYPE-CONJ
- coordinating
- subordinating

Categories: nouns, adjectives pronouns, pronominal adjectives, articles, numerals, verbs
GENDER
- feminine
- masculine
- neuter
- utrum
- common

Categories: nouns, adjectives, pronouns, pronominal adjectives, articles, numerals, verbs
NUMBER
- singular
- plural
- invariant

Categories: nouns, adjectives, pronouns, pronominal adjectives, numeral
CASE
- nominative
- genitive
- accusative
- dative
- oblique
- basic

Categories: pronouns, pronominal adjectives, verbs
PERSON
- 1
- 2
- 3

Categories: nouns, adjectives, pronouns, pronominal adjectives, articles, numerals
DEFINITENESS
- definite
- indefinite

Categories: adjectives, adverbs
DEGREE
- positive
- comparative
- superlative

Category: verbs
AGREEMENT
- person
  * 1
  * 2
  * 3
- number
  * singular
  * plural
- gender
  * feminine
  * masculine
  * neuter
  * utrum
  * invariable

Category: verbs
VERB FORM
- infinitive
- gerund
- participle
- supine
- finite

Category: verbs
MOOD
- indicative
- imperative
- subjunctive
- conditional

Category: verbs
TENSE
- present
- past
- future
- imperfect
- preterite

Category: verbs
VOICE
- active
- passive

Category: verbs
VERB TYPE
- auxiliary
- modal
- lexical

Category: verbs
BASE FORM

Other distinctions are:

Special distinctions
    special marks for distinctions that are very language and/or purpose dependent

Double tag
    compound (disjunctive) tags in the case of unsolved ambiguities

3.2.4 *Nouns*

The Penn and BNC tagging schemes provide the possibility of marking actual ambiguity between nouns and other parts of speech with tag combinations. In the table below, this is marked with an X in the row for Double tag. Penn proposes two double tags: adjective/noun and noun/*-ing* form. BNC has three combinations: adjective/noun, common noun/proper noun, common noun/*-ing* form. The annotation strategy of both corpora is to include *-ing* forms functioning and behaving as nouns under this label.

In Penn, the indefinite pronouns are included in the noun category, and so is *one* when used as a noun, but this is a closed set of words which is easily extractable if one wants to give them a different tag. The BNC tagset has a special label for the word *one*, irrespective of its function.

In the Gothenburg tagset, which is very reduced and does not even distinguish between proper and common nouns, the noun tags may have the symbol of the possessive value added to them, in order to mark the possessive form, *'s*. The possessive value is signalled under 'Case', value genitive. In the Brown tagset, too, all the noun labels may be extended with the symbol of the possessive element. The Lancaster tagging scheme marks the possessive form with a separate label.

The LOB and Lancaster tagsets are the most detailed ones as far as the nouns are concerned. Due to their many distinctions, they are also the most purpose-dependent ones among the annotation schemes analyzed here.

The proper nouns in the Italian corpus are split between two tags: person names and toponyms. Foreign toponyms are incorporated in the Foreign word category. In SUC, too, foreign toponyms are kept apart from the proper names, and are included in the class of foreign words.

The Paris tagging model includes proper names in the noun category, which has subtags for numeral nouns and acronyms. A further distinction is made for the common gender feature: nouns which are either feminine or masculine receive one tag, and those where the gender is not marked receive another. The same kind of tagging strategy applies to the number feature: one tag for nouns that are either singular or plural, and a separate one for nouns that are unmarked with respect to number.

The Dutch tagset distinguishes a basic form and genitive case. Furthermore, some archaic flectional forms pertaining to case are recognized, and marks for some distinctions referring to special functions of nouns are also provided (these, on the boundary of the realm of morphosyntax proper, are listed under the heading Special distinctions).

A tagging of Spanish would include the same features and values as those applied to the ILC Italian corpus.

As regards ENGTWOL, some numerals are classified as nouns.

	Pe	BNC	Go	Br	LOB	La	SUC	It	FrS	Par	Eur	UDB	ETW		GER
Category noun	X	X	X	X	X	X	X	X	X	X	X	X	X		X
TYPE-N															
- common	X	X		X	X	X	X	X	X	X		X	X		
- proper	X	X		X	X	X	X	X	X	X		X	X		
GENDER[7]															
- feminine								X	X	X	X				X
- masculine								X	X	X	X				X
- neuter							X								X
- utrum							X								
- common								X	X	X					
- unmarked								X		X					
NUMBER															
- singular	X	X	X	X	X	X	X	X	X	X	X	X	X		X
- plural	X	X	X	X	X	X	X	X	X	X	X	X	X		X
- invariant		X						X	X	X			X		
- unmarked										X					
CASE															
- nominative							X						X		X
- genitive		X	X	X	X	X	X					X	X		X
- accusative															X
- dative															X
- oblique															
- basic												X			
DEFINITENESS															
- definite							X								
- indefinite						X									
Special distinctions:															
capitalization					X	X									
place nouns					X	X									
toponyms							X								
cardinal points				X[8]	X	X									
days of the week				X		X									
months				X		X									
collectives						X							X		
numeral nouns						X				X			x		
titles					X	X							X		
measurements					X	X									
cited words						X									
acronyms							X		X					X	
attributive use												X			
interjective use											X				
selfreferential funct.												X			
archaic flect. forms												X			
Double tag	X	X						X							

[7] The value 'common' can be exemplified by It.*insegnante (teacher)*, which can be either masculine or feminine, and by Fr. *un/une bibliothécaire (librarian)*. The value 'invariant' is used when the number is undecided: Eng. aircraft, data, It. attività (activity/-ies), Fr. gaz (gas), Sp. crisis (crisis). An example of 'unmarked' gender is 'Mitterrand', and an example of 'unmarked' number is pu in 'ils ou elles ont pu', where 'pu' does not agree neither in gender nor in number.

[8] The days of the week and the directions north, north-east etc. share a separate tag and are thereby distinguished from other nouns.

3.3 *Standardization: Needs and Requirements*

The comparison of the morphosyntactic information encoded by the analysed tagsets led to the conclusion that it would be possible to propose a minimal standard scheme.

A - As regards linguistically annotated resources, there are some basic requirements that a standardized annotation must minimally fulfill:

- as far as possible cover a very large range of uses or offer the framework for multiple purposes;
  The tagsets used so far in corpus annotation practices are not multiple purpose schemes since they have been designed according to the needs and interests of the user(s) of that particular tagset.
- reflect a consensual analysis of data, i.e. one that is commonly agreed upon.
  The phase of analysis and comparison of different annotation practices used in corpora gave the following positive results:
- as to the morphosyntactic information encoded by different schemes, there are many contact points which can constitute the basis for an attempt at standardization;
- the existing differences, depending on language or different theoretical approaches, can usually be taken care of with a flexible multiple level proposal.

B - As regards criteria to follow in the design of a common scheme, two variables should be considered, as pointed out in (Leech, forthcoming):

- "annotators' points of view": speed, consistency and accuracy are basic requirements: a simple scheme (a reduced tagset) is easy to learn, apply and check for errors and consistency;
- "users' points of view": the user is mainly concerned with purpose: some uses require a high degree of delicacy in the analysis, i.e. a large and refined tagset. For other uses a cruder analysis is preferred, and a small tagset can be adopted.

A third variable also has to be taken into account: the "machine's point of view", i.e. the implications for the tagset of an analysis that is to be performed automatically (a discussion on a completely automatic analysis is presented in (Sinclair, 1991, NERC-19).

The Lancaster scheme (Garside et al., 1987) is an example of an annotation strategy where a large and refined tagset was preferred. The simplicity strategy was chosen within the Penn Project (Marcus and Santorini, forthcoming) since large quantities of data had to be tagged by several annotators: a reduced tagset seemed to be a guarantee of speed, consistency and the minimization of errors in the labelling process. The almost fully automatic Helsinki tagger also makes use of a reduced tagest (Karlsson, 1992, NERC-74).

An obvious interrelation can be seen between the size of the corpus to be tagged and the depth of the analysis, i.e. the delicacy of the annotation:

- small size corpus, rich annotation scheme;
- large size corpus, simplified scheme, i.e. reduced tagset and fewer distinctions.

In the design of a widely usable tagset, it can be argued that **simplicity** along with **flexibility** and **variable degree of delicacy** constitute essential properties and are necessary components on the way towards agreement.

In a certain sense, the simplicity strategy can be said to meet the annotators' and the users' needs, and thereby also to meet the requirements on a standard: a simple scheme is easy to learn and to follow, and allows high speed in the annotation process (annotators' needs). With a simple but flexible scheme, moreover, fine-grained theory-dependent decisions do not have to be made, a broad range of uses can be covered and a large quantity of data can be tagged (users' and machine's need). The present proposal is in line with the simplicity and flexibility strategy.

C - Two general and basic requirements on tagging (less controversial than the two at point A - above) have to be considered:

- it must be possible to separate the annotation from the raw text corpus.
  The annotations are added to the text and can be said to add a subjective element to it; since they are quite different in nature from the authentic corpus itself, the raw text corpus must always be recoverable (Leech, forthcoming; see also NERC Consortium, 1992, NERC-99).
- the annotation criteria must be described in as much detail as possible in the tagging guidelines.
  The guidelines are essential for the annotator and the user: both have to know what a tag stands for, and to which elements and according to which criteria it applies. In order to avoid misunderstandings and arbitrary decisions, detailed information is needed, and in the case of ambiguities, the guidelines must provide instructions as to their handling. Since the guidelines are of vital importance in any attempt at standardization, it follows that they have to be clear and exhaustive.

3.4 *Towards Standardization*

We summarize here some issues which are of relevance on the journey towards standardization.

3.4.1 *Methodology: A Bottom-up Procedure*

The methodology adopted in order to show the feasibility of harmonizing the morphosyntactic information added to corpora is a bottom-up approach, i.e. the means to enable a common tagging convention is looked for in the large core of agreement between various tagging practices. A way towards harmonization is also indicated for the difficult cases, and the problems are pointed out.

Since the methodology used is based on the study of established annotation practices, the present proposal can be said to be of the 'de facto' type. It is important to stress that its purpose is to suggest a starting point for further discussion and evaluation by users with different purposes in mind.

In the following, the focus will be on the content of the tags. Content and form are two sides of the same coin and are thereby linked to each other, but it was not the central objective of this study to deal with the formalism as well. This aspect is being developed for example within the

TEI by the AI Committee (Langendoen and Fahmy, 1991 and Langendoen and Zepp, 1992).

In (Monachini and Östling, 1992b, NERC-61) a first proposal towards a consensual scheme is discussed category by category. The definition and treatment of the categories are also accounted for. The problems encountered for some categories are focussed upon and a solution is proposed. For each category a set of morphological features is provided: a category is thus defined by its name and is associated with a set of features (whose first letters are capitalized) in the form of attribute-value pairs (values are in lower case, preceded by a dash).

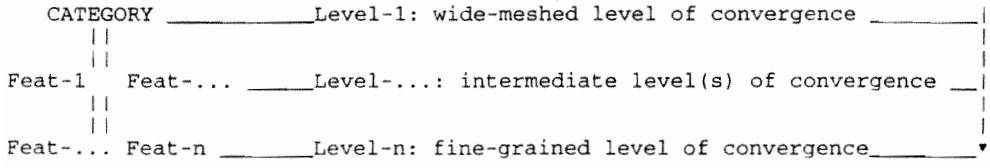
3.4.2 *Consensual Categories*

As regards the categories for which fundamental agreement emerged, no particular problems arose in the definition of a consensual set of features: those features are included which are common to various annotation schemes.

3.4.3 *Problematic Categories: different levels of granularity*

In the attempt to harmonize the information encoded by the different annotation systems and to propose a common denominator, some categories were found to be particularly problematic. In cases where there is little agreement as to the treatment of a category and a proposal based on common points cannot be made, a flexible proposal allowing for choices on different levels of standardization is explored, thereby providing separate but compatible solutions: each system will choose its most appropriate level of distinction.

- The category (PoS), if commonly recognized and defined, is the first point of convergence and can be seen as a wide-meshed level of standardization
- The features can be arranged in a hierarchy of deeper and more fine-grained levels. That is to say that all the features do not appear at the same level, but, depending on the category, some are pertinent to one level, others to subsequent level(s). The lower and deeper level (which is the level of more granular standardization) includes the relevant feature(s) of the upper level(s)[9]



Thus, a tagset which encodes only category information becomes comparable at least at the first level with tagsets which recognize a set of more granular information for the same category.

[9] It is worth reminding that the features of a lower level add new information.

3.4.4 *Transduction between Existing Tagsets and the Proposed Scheme*

For each category, it is necessary to investigate very carefully the problems regarding the transduction between existing tagsets and a common proposal. These transduction tests consist of checking the transferability between the information coded by an existing tag and that contained in the proposed common convention.

Different degrees of transferability and various problems arising from this are envisaged. If A and B stand for tags of an existing tagset and X and Y stand for categories in the proposal, the following correspondences hold:

i) A goes directly to X: there is exact correspondence. Example:
The Adjectives (A) in the Penn Treebank can be transferred to the Adjective category (X) in the proposal.

ii) A and B go to X: X is a wider category which includes A and B. There are no correspondence problems. Example:
The SUC Swedish categories Participle (A) and Verb (B) are subsumed by the category Verb (X) of the proposal.

iii) A goes to X and to Y and the different instances of A are easily extractable automatically: the correspondence is automatically retrievable. Example:
The Noun category (A) in the Penn Treebank also includes the Indefinite Pronouns, which belong to a closed set and can be listed. The Nouns can be transferred to the Noun category (X), while the elements identified as Indefinite Pronouns will go to a Pronoun category (Y).

iv) A goes to X and Y and the different instances of A are impossible to disambiguate automatically. Example:
Many tagsets do not distinguish between the pronoun and determiner functions of the Demonstratives (A). If a transfer is to be performed to Level-2 or -3 in the proposal (on which the function is distinguished: Pron (X) and Det (Y)), manual disambiguation will be necessary. Another solution would be to make the transduction on Level-1.

3.4.5 *Special Distinctions*

Distinctions that are very tagset- and/or purpose-dependent are marked as special distinctions. This is information which can not be considered in a first proposal for a minimal standard. To give an example,in the Noun category of the Lancaster tagging scheme there are special marks for the months, titles and citation forms.

In order to fulfil the flexibility requirement, it is important to retain the possibility of making distinctions according to user needs, and this factor should therefore be considered if a more articulated proposal for common morphosyntactic annotation is to be made.

3.4.6 *Double Tags*

Due to the fuzzy boundaries between categories, transcategorization phenomena occur frequently. Only some of the analyzed annotation practices allow the possibility of double tagging uncertain cases, but in order to avoid arbitrary decisions for difficult ambiguities, a standard annotation

practice should permit the recording of this uncertainty. As specified above, the guidelines must also be very clear as to the criteria for handling these ambiguities: they must be described in as much detail as possible. Annotators should be sure of the information they add, without being subject to the pressure of having to make a choice (Leech forthcoming).

3.5 *A First Proposal for a Standardized Scheme*

The proposal for a consensual annotation scheme is articulated category by category, according to the main PoS, taking into account for each of them points of convergence and divergence and drafting proposals accordingly. We summarize here, as a way of exemplification, some of the issues dealt with under the category Noun.

3.5.1 *Category: Noun*

The category Noun is recognized by all tagsets, and according to available information consensus can be achieved as to the identification of membership in the category. A particular case, however, is the Penn Treebank, which - as mentioned above - includes in the Noun category *one*, the indefinite pronouns *naught, none* and compounds of *any-, every-, no-, some-* with *-one* and *-thing*. This poses no problems with regard to the correspondence between that tagset and the one proposed here. If these elements, that belong to a closed set, are to be transferred to another category, they are easily and automatically extractable from the Nouns. This, then, is an example of correspondence of type iii) (see above, section 3.4.4.).
Noun features shared by the tagsets, and the proposed values, are the following:

Type	Gender	Number	Case
- common	- masculine	- singular	- nominative
- proper	- feminine	- plural	- genitive
	- neutrum		- dative
	- utrum		- accusative
			- basic
			- oblique

These could constitute the basic features and values of a common scheme.

Ambiguities for which double tagging should be foreseen are, minimally, Noun/Adjective and Noun/Verb-participles.

Type

The feature Type has two possible values: 'common' and 'proper'. These values are distinguished by all tagsets, except Gothenburg and EUROTRA. This means that for the last two, nouns cannot be mapped automatically onto these values.

Gender

This is a feature whose values are language-dependent: in English there is no gender distinction

This is a feature whose values are language-dependent: in English there is no gender distinction for Nouns; in the Romance languages there is the feminine, the masculine and often the common gender, while the Scandinavian languages have the genders neutrum and utrum for Nouns. It was decided to leave out the values 'common' and 'unmarked' from the proposed set of shareable values, since it can be seen as redundant information: it corresponds to the conjunction of the two single values 'masculine' plus 'feminine'.

Each annotation scheme will select from the proposed set the values pertinent to the represented language:
- Romance languages: 'masculine', 'feminine' and 'masculine+feminine'
- German: 'masculine', 'feminine' and 'neutrum'
- Scandinavian languages: 'neutrum' and 'utrum'
- English, Dutch: the feature Gender is not pertinent to English and Dutch nouns

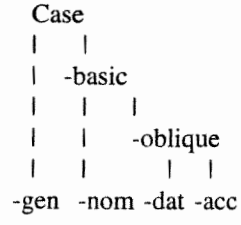
Number

All languages studied recognize the values 'singular and 'plural'. The Romance languages and two English tagsets among those analyzed mark the value 'invariant.' This last has not been included in the proposal, since it can be represented by the value 'singular + plural'.

Case

Some problems arise as to the definition of the values pertinent to this feature. As appears from the preceding phase of comparison, the values used under Case are the following: 'nominative', 'genitive', 'dative', 'accusative', 'basic', 'oblique'. Clearly not all of them are mutually exclusive: some of them overlap, being used in differently structured case systems. It should be pointed out that, given these overlappings, the values can never appear all together in one language, but a list of permitted values for each particular language has to be given. The signification of a value has to be seen in relation to the other values admitted for the same language.

The relationship between the values, as shown by their use in the analysed tagsets, is illustrated in the following tree:



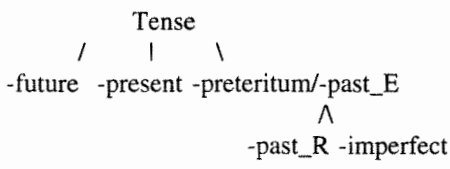
It must be stressed that each language system will use its own appropriate set of values. For the Noun category:
- German: 'nominative', 'genitive', 'accusative' and 'dative'
- Dutch and Scandinavian languages: 'basic' and 'genitive'
- English: 'basic', 'genitive'. The value 'genitive' refers to the Saxon genitive
- Romance languages: the feature Case is not pertinent to Noun (it is pertinent to Pronoun)

'Oblique' is presented here as a possible value of the feature Case, even though it does not seem to be used as a value of the category Noun. It is used in English and Swedish tagsets as a value marked for Pronouns, which present the following distinction system: 'oblique' is opposed to 'nominative', e.g. *him* vs *he*, whereas *his* marked for 'genitive' is, properly speaking, a separate Type of pronoun, the Possessive. *Whose*, on the other hand, can be regarded as the genitive case of the interrogative/relative *who*. 'Oblique' is used in the two(three)-value systems, i.e. systems which have the set of values 'oblique', 'nominative' and ('genitive'). It can be compared, as shown by the tree above, with 'accusative' and 'dative' in a four-value system, such as German. The same holds for a system like Italian, where *him* is translated by *gli* and *lo* ('dative' and 'accusative', respectively).

3.5.2 *Other categories: different problems but similar solutions*

Other problems of mismatches arising in the treatment of other categories have been dealt with wherever possible by using a flexible and multi-layered approach. This solution has been adopted, for example, for Verbs, where there are big differences in the verbal systems among the languages studied. English, which has very few inflections, is at one extreme, and the Romance languages, which have a very rich verbal morphology, are at the other. It was decided to articulate the proposal on two levels: Level-1 is the cruder one and should be easily reached from the existing tagsets, while Level-2 permits further distinctions not always made in all the tagsets.

Another problem arising in the Verb category is constituted by the fact that some values of the feature Tense are overlapping, due to the internal organization of the verbal system of each language, which groups the tenses differently. 'Present' is the only tense whose use is the same in all the languages studied. The 'preteritum' in Swedish would be split in the two values 'past' and 'imperfect', which are both pertinent to the Romance languages. The English 'past' does not have the same meaning as it does for the Romance languages: it is not opposed to an 'imperfect' value, but instead it is similar to the 'preteritum', which is opposed to the 'present'. This complex situation can be represented by the following tree:



In Romance language systems, the values 'past' and 'imperfect' are opposed and designate two different aspects of a past action, and both are opposed to the 'present' with respect to the notion they represent: 'past' is a punctual action finished in the past and 'imperfect' is a durative action initiated in the past. In order to avoid misunderstandings, a tentative solution could be to rename the Romance 'past' value 'perfect', as it is opposed to 'imperfect'.

A very basic proposal is to include all values recognized by each verbal system without trying to solve overlappings. For each language, a list with the permitted values of this system must be supplied.

3.6 *Recommendations*

Even if it is evident that a "best scheme" cannot be achieved and the recognition of a theory-neutral scheme is a controversial idea, the study has shown that it is still possible to explore the provision of a workable framework, in order to meet the needs of different users with various purposes. A consensual standard scheme, in the sense of a nucleus of tags that are broadly accepted and thereby shareable, may be proposed as a result of the observation of annotation practices. Such a scheme has to be suitable for extension, refinement and adaptation. In other words the key elements are de-facto agreement, consensual tags and a flexible scheme.

The survey of corpus annotation practices showed that it is indeed feasible to propose a minimal common scheme at the morphosyntactic level; a strategy for devising a possible tagging convention can also be formulated on the basis of this initial phase. The task is far from trivial: a major difficulty is the disagreement about the recognition, definition, and treatment[10] of some categories, depending either on differences between languages or on different linguistic traditions. In (Monachini and Östling, 1992b, NERC-61), however, it is shown that there are possible solutions to these problematic cases, and further developments are to be expected within EAGLES.

References

Calzolari N., Zampolli A. (1990): "Lexical Databases and Textual Corpora. A Trend of Convergence between Computational Linguistics and Literary and Linguistic Computing", in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, pp.76-83.

Copeland C., Durand J., Krauwer S. & Maegaard B. (eds) (1991): *The Eurotra Linguistic Specification*, Studies in Machine Translation and Natural Language Processing, Vol. I, Luxembourg, Commission of the European Communities.

EAGLES Workplan (1992, draft version): "WG - Corpora - Workplan", Pisa.

Ejerhed E., Källgren G., Wennstedt O. & Åström M. (1992): "The Linguistic Annotation System of the Stockholm Umeå Corpus Project - Description and Guidelines", Version 4.31.

Ellegård A. (1978): *The Syntactic Structure of English Texts.* Gothenburg Studies in English, 43. Stockholm.

Engwall G. (1974): *Fréquence et distribution du vocabulaire dans un choix de romans français.*

[10] Some terminological remarks can be useful.
The *recognition* of a category by a tagset means that a scheme recognizes the existence of this category.
The *definition* of a morphological class refers to which elements are included in it.
The *treatment* of a morphological class refers to which features are taken into account in a category.

Stockholm.

Engwall G. (1984): *Vocabulaire du roman français (1962 - 1968). Dictionnaire des fréquences.* Stockholm.

Francis W.N. (1980): "A tagged corpus - problems and prospects", in S. Greenbaum, G. Leech and J. Svartvik (eds), *Studies in English Linguistics - for Randolph Quirk.* Longman.

Francis W.N. & Kucera H. (1982): *Frequency Analysis of English Usage: Lexicon and Grammar.* Houghton Mifflin.

Garside R., Leech G., Sampson G. (eds.) (1987): *The Computational Analysis of English -a Corpus-Based Approach*, Longman.

Johansson S. (1986): *The Tagged LOB Corpus: Users' Manual.* Norwegian Computing Centre for the Humanities, Bergen.

Karlsson F. (1992): "SWETWOL: a comprehensive morphological analyser for Swedish", *Nordic Journal of Linguistics*, 15, pp. 1-45.

Karlsson F., Voutilainen A., Heikkilä J. & Anttila A. (eds), (forthcoming): *Constraint Grammar: a Language-Independent System for Parsing Unrestricted Text.*

Langendoen D.T., Fahmy E. (1991): "Feature structure markup for presentation at Oxford and Brown workshops", TEI AI1 W9.

Langendoen D.T., Zepp S. (1992, draft): "Encoding Linguistic Analyses Using the Guidelines of the Text Encoding Initiative", Dept of Linguistics, University of Arizona.

Leech G. (forthcoming): "Corpus Annotation Schemes", paper presented at the Pisa Corpus Workhop (24-26 January 1992), to be published in the Proceedings of the Conference, OUP.

Marcus M., Santorini B. (forthcoming): "Building very large natural language corpora: the Penn Treebank", paper presented at the Pisa Corpus Workshop (24-26 January 1992), to be 'published in the Proceedings of the Conference, OUP.

Monachini M. (1992): "Core Set of PoS Tags for Italian", Internal Report, Istituto di Linguistica Computazionale del CNR, Pisa.

Östling Andersson A. (1987a): *L'identification automatique des lexèmes du français contemporain.* Studia Romanica Upsaliensia 39. Uppsala.

Östling Andersson A. (1987b): "Une description "deux niveaux" du français écrit", UCDL-R-87-1, Center for Computational Linguistics, Uppsala University.

Renzi L. (1988): *Grande grammatica italiana di consultazione*, Il Mulino, Bologna.

Santorini B. (1991): *Part-of-Speech Tagging Guidelines for the Penn Treebank Project.*

Sågvall Hein A. (1992): "On the Coverage of a Morphological Analyser Based on "Svensk Ordbok" [A Dictionary of Swedish]". In: *Proceedings of the Scandinavian Conference in Computational Linguistics, Bergen 28-30 November 1991*, Report Series of the Norwegian Computing Centre for the Humanities, No. 56, Bergen.

Zampolli A. (1990): "Project Definition for the Constitution of a Network of European Reference Corpora", Project Proposal, Pisa.

Relevant NERC Papers

Bindi R., Calzolari N., Monachini M., Pirrelli V., Zampolli A. (forthcoming): "Corpora and Computational Lexica: Integration of Different Methodologies of Lexical Knowledge Acquisition", paper presented at the Pisa Corpus Workshop (24-26 January 1992), to be published in the Proceedings of the Conference, OUP, NERC-177.

Bindi R., Calzolari N., Monachini M., Pirrelli V. (1991): "Lexical Knowledge Acquisition from Textual Corpora: A Multivariate Statistic Approach as an Integration to Traditional Methodologies", in *USING CORPORA Proceedings, Seventh Annual Conference of the UW Centre for the New OED and Text Research*, Oxford, U.K, NERC-103.

Blanco Rodriguez M.J. (1992): "Criteria for Morphosyntactic Labelling of Spanish", Working Paper, Malaga, NERC-112.

Dutilh-Ruitenberg M.W.F. (1992): "Corpus Annotation Schemes in the Netherlands", Working Paper, INL, Leiden, NERC-69.

Lafon P. (1992): "Dictionnaires machine et lexicométrie", in *Etudes de Linguistique Appliquée* 85-86, Working Paper, Paris, NERC-72.

Monachini M., Östling A. (1992a): "Morphosyntactic Corpus Annotation - A Comparison of Different Schemes", Technical Report, ILC, Pisa, NERC-60.

Monachini M., Östling A. (1992b): "Towards a Minimal Standard for Morphosyntactic Corpus Annotation", Technical Report, ILC, Pisa, NERC-61.

NERC Consortium (1991): "Network of European Reference Corpora - Technical Annex", Pisa, NERC-99.

NERC Consortium (1992): "Workshop on Textual Corpora", 24-26 January 1992, Report from the Conference, Pisa, NERC-82.

NERC Consortium (June 1992): ''Policy for Corpus Provision for Europe", Strategic Briefing Paper, NERC-99.

Panhuijsen M., van der Voort van der Kleij J., Wagenaar P. (1992): "Automatic Lemmatization Experiment - An explorative study", Working Paper, INL, Leiden, NERC-76.

Scheiter S. (1992): "Morphosyntactic Annotation Schemes in German laungage Corpora", Working Paper, Working Paper, IDS, Mannheim, NERC-124.

Sinclair J. (1991): "The Automatic Analysis of Corpora", *Directions in Corpus Linguistics*, J. (Svartvik (ed.), 379-97, Berlin, Mouton de Gruyter, NERC-19.

TEI AI 1W2 (June 1991): "List of Common Morphological Feature for Inclusion in TEI Starter Set of Grammatical-Annotation Tags", Working Paper, NERC-14.

4 Syntactic Annotation

The issues involved in the syntactic annotation of textual corpora are so many and various that the work has to be distributed among a number of different studies. The survey which follows, of the current practices in annotating corpora at the syntactic level, was integrated, in the NERC Work Package, by the contributions of (Antona, 1992a, NERC-64, and 1992b, NERC-63, Corazzari, 1992, NERC-68, and Ruimy, 1992, NERC-65), which are case studies attempting to bridge between the experience of existing Natural Language Processing (NLP) systems and corpus linguistics annotation practices.

4.1 *Methodology*

At the syntactic level, the comparison of annotation schemes and the consequent evaluation of the feasibility of standards required an ad hoc methodology. Needless to say, the analysis and comparison of syntactic annotation schemes cannot be carried out in the same way as has been done for morphosyntactic annotation schemes (see section 3 above). There is a fundamental difference between the two. At the morphosyntactic level, the features of the linguistic structure to be coded concern (with a few exceptions) individual words, i.e. they are word-level categories. At the syntactic level, on the other hand, the linguistic structure to be dealt with is the grammatical structure of the sentence. Consequently, a comparison of syntactic annotation schemes cannot proceed directly by comparing the codes used, for instance, for each syntactic constituent; the very nature of a syntactic constituent is under discussion, given that it often differs from one annotation scheme to another. Because of the obvious specificity of comparing structures, a mapping of syntactic representations requires, in our opinion, a two stage analysis.

During the first stage, the relevant factors characterizing the different syntactic representations are identified, and the various annotation schemes are classified on this basis. In order to identify the distinctive features of syntactic annotation schemes used in corpora projects and therefore to classify them, different classes of factors are to be taken into account, from the general

grammatical model behind the parsing scheme adopted, to the treatment of ambiguities and partially recognized syntactic structures, to more "external" features such as the purpose of the annotation or the technique through which it has been produced. All these factors contribute, in different measures, to the definition of the annotation scheme.

This first stage is in turn articulated in two substeps, comprising a dissection process and a reconstruction process. The first substep, the dissection process, involves isolating the relevant features characterizing the different annotation schemes. But none of the features which are identified here is unique to one annotation scheme or another: what distinguishes each annotation scheme is the combination of features. Therefore, for a full characterization of the various syntactic annotation schemes, a reconstruction process is needed, in which the features identified during the first substep are associated with each scheme.

The second stage operates instead at a more finely-grained level, that is within the classes identified during the previous stage. Syntactic annotation schemes with homologous structures are considered, and a comparison is made of shared grammatical concepts; for instance, the different kinds of syntactic constituents or syntactic functions recognized by the schemes making use of such concepts.

The study carried out in the framework of NERC, Workpackage 8.3, concentrated mainly on the first stage, while the second stage is proposed as next research step, to be performed, for instance, in the EAGLES Working Group on Text Corpora.

4.2 *The research sample*

The basis of this comparative study consists of some of the syntactic annotation schemes used for textual corpora of English. The limitation of the study to annotation schemes conceived and used for English (whether British, or American, or International) can be seen from two different perspectives. On the one hand, the choice of annotation schemes conceived for English textual corpora reflects the (un)availability of large syntactically analysed corpora of other languages as publicly available research resources. On the other hand, the very same choice makes the comparison easier: possible differences are not due to peculiarities of the different languages to which the scheme has been applied.

Obviously, the results of this study, when seen from a multilingual perspective, are partial and provisional, but they are expected to be applicable to other languages with ad hoc integrations and changes. We think that the parameter set which emerged from the survey of syntactic annotation schemes is representative of the general problems faced in the attempt to annotate corpora, at least at a surface level of syntactic analysis. Accordingly, we do not expect the analysis of annotation schemes designed for other languages to alter the set significantly, but possibly to enrich it.

The syntactically analysed corpora on which the study is based are listed in the table below. The sample composition is mainly motivated from a methodological point of view; if merely considered from the corpus angle, it appears to be very heterogeneous (see, for instance, the different corpora sizes, or the different status of the analysis, completed, under development, or still at the project stage). The reason for the selection is that we wanted the sample to reflect all possible (i.e. those emerging from the analysis of available corpora) aspects of the design of

annotation schemes for corpora; for instance, particular corpora have been included in the sample to show the advantages and disadvantages of different schemes with respect to the uses of the analysed corpus and/or the technique adopted for producing the annotation.

THE ANALYSED CORPUS	SIZE (N. OF WORDS)	VARIETY OF ENGLISH	SPOKEN/ WRITTEN	REFERENCES
Nijmegen Corpus (Nijm)	130,000	British	written	Van Halteren & Van den Heuvel 1990
International Corpus of English (ICE)	17 million (planned)	National and Regional	spoken written	Van Halteren 1992
Lancaster-Leeds Treebank (LaLe)	45,000	British	written	Sampson 1987
LOB Corpus Treebank (LOB)	144,000	British	written	Leech & Garside 1991
Lancaster/IBM treebank 1987 (La87)	70,000	British	?	Leech & Garside 1991
Lancaster/IBM skeleton treebank (Lask)	---	British	spoken	Leech & Garside 1991
Susanne Corpus (Su)	128,000	American	written	Sampson 1992b
Gothenburg Corpus (Goth)	128,000	American	written	Sampson 1992a
Polytechnic of Wales Corpus (PWC)	100,000	British	spoken	Sampson 1992a
Penn Treebank (Penn)	1,100,000	American	written	Marcus & Santorini 1992
Bank of English (Constraint Grammar) (BECG)	200 million (planned)	British American other	spoken/ written	Karlsson 1990

4.3 *Comparing syntactic annotation schemes*

A set of parameters to be used for classification purposes has emerged from the comparison of the different annotation schemes examined. These parameters, extracted through the dissection process which each annotation scheme has undergone, represent the coordinates for characterizing the syntactic annotation schemes applied to textual corpora. In what follows the parameters are listed, and for each a sketchy illustration is given (for a detailed description see Montemagni, 1992, NERC-67). In the summary table at the end of this section, each annotation scheme which has been considered has been assigned the relevant set of distinctive features.

In what follows the parameters which have emerged so far - on the basis of the annotation schemes examined - as relevant for a characterization of syntactic annotation schemes will be discussed.

A. *Constituency- vs. Dependency-based model of syntax*

The first parameter which needs to be accounted for in classifying syntactic representations concerns the syntactic hierarchy they relate to. Broadly speaking, two different notions of syntactic hierarchy can be distinguished, corresponding to a constituency model and to a dependency model of syntax. Accordingly, syntactic annotation schemes used in corpora projects can be classified on this basis, that is whether they mark constituency and/or dependency relations.

In constituency-based annotation schemes, each syntactic constituent is connected to its immediate constituents up to the ultimate constituents, which are associated with the surface text (concerning the depth of the internal structure of constituents, see parameter H). Each constituent has associated with it the linguistic information, both formal (all annotation schemes in this group mark information about the category to be assigned to the syntactic constituent under definition), and/or functional (not all annotation schemes mark functional information as well; parameter B accounts for this last point). This approach to syntactic annotation is common to most of the projects considered: Penn, Lancaster-Leeds, LOB, Susanne, Nijmegen, ICE, and Lancaster-IBM. In these projects, the resulting "parsed corpora" are also known as "treebanks", and the syntactic annotation very often consists of the syntactic bracketing task.

The other possible definition of syntactic annotation is dependency-based, used by Gothenburg and by the Constraint Grammar (for the Bank of English), which assigns flat, functional, surface labels, optimally one to each word-form.

The analytic scheme adopted by the Polytechnic of Wales Corpus is a variety of Halliday's systemic functional grammar, and for this reason has a lateral position with respect to the dichotomy constituency vs. dependency.

B. *Functional vs. Categorial labelling*

Annotation schemes can also be classified on the basis of the kinds of labels associated with each node in the linguistic structure assigned to the text, coding respectively functional and/or categorial properties. Functional labels specify the relations of constituents - words or phrases - with the constructions in which they occur (for instance, they mark subject and object relations), while categorial labels specify intrinsic properties of constituents (i.e. the syntactic category they belong to). These properties are obviously strictly related to the syntactic model behind the annotation scheme (see parameter A).

As far as categorial classifications are concerned, dependency-based annotation schemes recognize only word-level categories (which pertain to morphosyntactic annotation schemes and not to syntactic ones, and are accounted for in (Monachini and Östling, 1992a, NERC-60). On this basis, such schemes do not specify categorial labels at the phrasal level, unless they are mixed schemes, as in the case of the Gothenburg corpus. On the other hand, phrasal categories are the building blocks out of which constituent structures are built; therefore, categorial labels

are only and always used in constituency-based schemes.

Functional labels are always present in dependency-based annotation schemes, but can also optionally occur in constituency-based ones.

C. *Treatment of potential and actual ambiguities*

Although some sentences in natural languages are evidently syntactically ambiguous, most of them are disambiguated by their context, so that the ambiguity is not noticed by the reader. This is the case of possibly ambiguous syntactic constructions. But not all syntactic ambiguities can be so easily solved, giving rise - when unsolved - to actually ambiguous constructions.

From the corpus point of view, the representation of ambiguity, if allowed, can present serious problems regarding the interpretation of frequency counts. In spite of this general remark, there are parsing schemes used for annotating corpora which provide the possibility of handling corpora containing possibly as well as actually ambiguous syntactic contexts, both at intermediate stages of the corpus annotation process and in the final result (Nijmegen, Penn, and Constraint Grammar).

A first distinction can be drawn on the basis of the nature of the ambiguity, that is whether it is an assignment or an attachment ambiguity. Uncertainties of linguistic category assignment are quite frequent in the analysis of corpora: this is not due to the failure of human understanding, but to the prototypical, or fuzzy, nature of most linguistic categories. Therefore, annotation practices should aim to record uncertainties as to whether one category or another should be assigned. Moreover, as (Leech, 1992) points out, it could be very useful to assign a likelihood score to the possible assignments. The other kind of ambiguity is structurally determined, and relates to the possible nodes a given syntactic constituent may be attached to. Attachment problems are mostly problems of modifier placement, which is often uncertain (following Hindle and Rooth, the attachment of 10% of prepositional phrases is unclear in real text).

D. *Representation of partial information*

One of the principles directing the design of corpus annotation schemes is that they should provide an analysis for everything occurring in a written text, with the exception of actual misprints. This principle motivated the requirement for allowing the indication of partial information within the annotation scheme. This parameter deals with cases of unrecognized syntactic constructions, in which a label cannot be assigned to a constituent: this corresponds to the practice of so-called unlabelled bracketing, adopted in several corpora projects (Penn, Lancaster-Leeds, Nijmegen). All corpora using this practice are constituency-based.

The existence of sentences which cannot be assigned a complete representation but only chunks of grammatical structures, covering only some parts of the sentence, is another case in point. This case is foreseen only by the Penn treebank for the intermediate stages of the annotation process. In uncertain cases, only a partial structure - which is accurate for the single chunks, and corresponds to a string of trees - is provided by the parser; at this point, the annotator's task is not that of rebracketing, but that of glueing together the syntactic chunks provided by the parser. None of the other parsing schemes seems to allow this kind of partial

annotation, neither at an intermediate stage of the annotation process nor in the final result.

E. *Surface vs. deep structure*

The question "deep vs. surface analyses for corpora?" can be answered differently, according to whether the answer is based on current practices or on the desiderata of corpus users. All the schemes examined here provide analyses which are mainly surface rather than deep. On the other hand, it is obvious that deeper parses would be more useful, but deep analyses are highly contentious (see Sampson, 1987, 1991). The advantages and disadvantages of deep analyses and their feasibility with respect to real texts are discussed in (Ruimy, 1992, NERC-65).

The status of the different corpora with respect to the representation of the deep structure of sentences is the following: the analysis schemes of Susanne and the Polytechnic of Wales represent logical as well as surface grammatical form; Gothenburg includes some limited indications of logical structure whenever it differs from surface grammatical structure; in other annotation schemes, the analysis seems to be purely surface.

F. *Treatment of specific syntactic problems*

This parameter focuses on the treatment of specific syntactic problems such as null elements, discontinuities, ellipsis, and coordination. Sometimes corpus annotation schemes, specifically conceived to represent real texts, account for these linguistic phenomena in a non-standard way with respect to computational and formal grammars; sometimes they simply do not represent them. Let us consider a few examples suggested by the annotation schemes examined in this study. Unfortunately, the information available on this subject is not as systematic as in the previous cases, but we thought that in spite of its incompleteness it was worth proposing this issue as one of the possible parameters for classifying syntactic annotation schemes for corpora. What we are reporting below is only explicit evidence, derived from the descriptions of the different annotation schemes. Given the fragmentary nature of this section, we could not include this parameter in the final table, and therefore the illustration of it will be more analytical than was the case for the others.

In what follows, we first report on phenomena which are only optionally accounted for in corpus annotation schemes, such as null elements and discontinuities. Secondly, we concentrate on one of the major divergence points between formal and computational grammars on the one hand, and corpus annotation schemes on the other - that is the treatment of coordination.

In Penn, syntactic constituents as well as null elements are represented: accordingly, parses include wh-traces, large PRO, and dislocated elements. Nijmegen allows for the representation of discontinuous structures. The Susanne scheme has a tag to represent a trace marking the logical position of a constituent which has been shifted elsewhere, or deleted, in the surface structure (see Sampson, 1992b). This tag can then be assigned an index to show referential identity with other constituents of the same sentence. Moreover, indices can be generally assigned to pairs of nodes to show referential identity between items which are in certain grammatical relationships with one another. The Polytechnic of Wales and Lancaster-IBM also permit discontinuous constituents to be recognized. However, negative evidence in this respect comes from Lancaster-Leeds, whose scheme does not show the logical unity of discontinuous

constituents.

As far as the treatment of coordination is concerned, there are three annotation schemes proposing ad hoc representations for corpora: Nijmegen, Lancaster-Leeds, and Susanne.

As (Aarts and Oostdijk, 1988) point out, one of the major problems in the analysis of corpora occurs when (part of) an utterance does not constitute a single category. This phenomenon typically occurs in coordination, in particular through conjunction reduction and gapping. In the sentence "John bought a new record-player and Shirley a radio", the two noun phrases in the second conjoin ("Shirley" and "a radio") do not combine to form one sentence constituent, let alone a single category. Yet there is clearly some sort of relation between the two noun phrases which is to be expressed somehow. Most theoretical approaches to syntax attempt to describe this relation by referring to some underlying level of representation at which the second conjoin consists of a complete sentence. Even in models in which this is not the case (e.g. GPSG which deals with a single level of representation) these structures are usually regarded in terms of what is missing with regard to a superordinate node (see the slash principle in GPSG). The alternative which is being investigated within the Nijmegen corpus is closest to surface structure analysis, and consists of describing what is actually there without referring to underlying levels of representation or missing constituents, and without introducing a mother node when two constituents cannot be said a single one at a higher level. Accordingly, the analysis in this case should leave "Shirley" and "a radio" as two separate noun phrases.

In the Lancaster-Leeds treebank, the treatment of coordination is assimilated to that of subordination. Coordinated noun phrases or sentences are analysed as follows: **[my mother [and my father]]**; **[John played, [Wendy sang,] [and Anne danced]]**, with the second and the subsequent conjuncts treated as subordinated to the first one. Although this approach may seem illogical (since semantically the function of coordination is to express the equivalence between the conjuncts), it is said (Sampson 1987) to be more plausible from the psychological point of view. Similarly, the Susanne scheme analyses the second and subsequent conjuncts in a coordinate structure as subordinate to the first conjunct. Thus, a coordination of the form **A, B, and C** would be assigned a structure of the form **[A, [B], [and C]]**, where the categorial tag of the entire coordination is determined by the properties of the first conjunct. The Lancaster-IBM corpus also seems to adopt a similar strategy for handling coordination.

G. *Skeletal parsing*

The skeletal parsing technique involves the bracketing of constituents above word-level and labelling them with the corresponding syntactic category, but with specific restrictions on the tags and structures allowed (the tagsets of non-terminal categories are quite reduced, less than twenty tags in all cases). The categories which have been selected are the ones considered as "canonical", that is likely to be uncontroversial and therefore to remain unaffected by differences of theory (which obviously remain among constituency-based models of syntax). These tagsets can be therefore considered as a possible basis for future studies and proposals for shared grammatical concepts.

This technique can be seen from different perspectives: it relates on the one hand to the "theory neutrality" requirement, and on the other to the training phase of stochastic grammars.

The simpler the scheme, the less likely it is to violate the presumptions of individual theories.

(Leech, 1992) reports the example of the category of noun phrases, which is broadly recogn by different theories and for which there is substantial agreement about the boundaries. disagreement is related instead to the internal structure of the noun phrase. It is there reasonable, as Leech affirms, "for a syntactic annotation scheme to distinguish the boundarie the noun phrase without being too much concerned about its constituency". Skeletal parsing g in this direction, and therefore can be seen as a possible answer to the theory neutra requirement.

Skeletal parsing is also connected with the training process of stochastic grammars. As c be noticed by examining our sample of syntactic annotation schemes, variable degrees granularity of linguistic information can be added to a raw corpus. The delicacy of the analy: should not be seen as a value in itself; instead, the more granulated the analysis the scher offers, the larger the corpora that are required in training stochastic grammars. Therefore, tl tendency to adopt more granulated analysis schemes is now being reversed at all linguistic leve of description (i.e. there is a move from more detailed annotation schemes to more simplifie ones); the skeletal parsing technique can also be seen and justified from this perspective.

Moreover, from a practical point of view, a less detailed annotation scheme helps to eliminat sources of error, inconsistency, and uncertainty in annotating, and increases the speed of bot annotation and post-editing.

The Penn and the Lancaster-IBM are the only projects in which the skeletal parsing technique is now being experimented with. Only one claim against this technique comes from the International Corpus of English, which aims for a full syntactic analysis rather than for a skeletal parsing. On the other hand, dependency-based annotation schemes seem not to be suitable candidates for the skeletal parsing technique, at least as it has been defined in this context (that is characterizing constituents by identifying their boundaries, rather than their internal structure).

H. *Flat vs. steep trees*

The skeletal parsing technique we saw above is an example of analysis reduction, on the one hand of the set of syntactic categories the analysis is based on, and on the other of the steepness of the analysis, which is flat. The situation of the annotation schemes under consideration with respect to these two possible ways of simplifying the analysis is different: while the number of syntactic categories varies considerably across the different annotation schemes, the trees are almost always flat.

The dichotomy "flat vs. steep" trees can be applied only to constituency-based annotation schemes. The general tendency in the sample examined is that of assigning flat rather than steep analyses: there is only one annotation scheme making use of steep trees, the Lancaster-IBM 1987 treebank. This is the result of an experiment in reducing the sparse statistics problem arising when using syntactically annotated corpora for training stochastic grammars.

According to (Leech and Garside, 1991), in the grammar derived from the Lancaster-Leeds treebank, using flat trees, a large proportion of the rules occurred only once. A possible way of reducing the problem of sparse statistics was to represent the syntactic structure by means of steeper annotations. The Lancaster-IBM 1987 treebank is the result of this experiment. In this treebank, the parsing scheme has been designed in such a way as to create steep parse trees, by introducing intermediate nodes. While the noun phrase in a flat representation has determiners,

adjectives, noun heads, and other possible modifiers as its immediate constituents, in a steep representation (like the one proposed by grammars modelled on X-bar syntax) it has at least one intermediate node (N'), and often several, between itself (N'') and its constituent words. But after about 70,000 words of annotated text, the project was abandoned: the time required for annotation was unacceptable; moreover, the open-endedness of the grammar of whatever language showed that steep trees were not the appropriate answer to the problem of sparse statistics.

I. *Treatment of specific phenomena to real text*

Adopting a corpus-based paradigm for syntax is to be confronted with the gap between language as described by grammatical theories and as attested by real-life usage. It is widely recognized that there is only a partial overlapping between the structures actually observed in corpora and those usually described by grammatical theories and dealt with by natural language processing systems. The existence of a massive range of phenomena which rarely or never crop up in theoretical literature imposes a revision of the syntactic annotation schemes which are heavily committed to one or another grammatical theory. If we want to deal with language as it is really used, this gap has to be filled.

There are areas of language, usually neglected in theoretical and computational as well as traditional grammatical descriptions, which are specific either to written language or to speech. Items such as postal addresses, sums of money, dates, weights and measures, bibliographical citations and other comparable phenomena occur quite frequently in written language, and have their own characteristic "syntax" in different languages. Although such constructions are almost always considered outside the domain of the language proper, they are very important from the point of view of practical language processing applications, and need to be appropriately dealt with in order to be represented as part of the linguistic structure. Still at the written language level, there is another area, that of punctuation, which is normally excluded from grammatical analysis despite its significance, which is equal to that of grammatical words such as prepositions. The same holds, in spoken language, for the so-called "speech repairs", linguistic constructions whose role at the discourse level (for instance in maintaining the topic of the discourse) is not accounted for in standard linguistic structures.

Real texts are full of idiosyncracies, but very few of the annotation schemes considered in this survey attempt to account for such phenomena.

The analytical scheme of the Lancaster-Leeds treebank attempted to specify an unambiguous analysis for any phenomenon occurring in authentic written English, including not just discursive text but items such as addresses, sums of money, bibliographical citations, and purely orthographic phenomena such as punctuation. With respect to the latter, the Lancaster-Leeds treebank, and the closely related parsed LOB corpus, treat punctuation marks as parsable items on a par with words. These parsing schemes include detailed rules for the placement of punctuation marks in parse trees: the closing full stop is treated as a sister to the S node as an immediate constituent of the root; commas surrounding a constituent like an adverbial phrase are represented as daughters of the same mother node, since they balance one another logically.

Negative evidence in this respect comes from the Gothenburg and the Polytechnic of Wales corpora. In the Gothenburg corpus, punctuation, with other orthographic details of the original text such as case distinction, has been thrown away. Similarly, in the Polytechnic of Wales

corpus, which is the only spoken corpus considered in this survey, items such as "oh" or "mm" have been excluded from the parse trees as "non verbal".

J. *Types of representation*

The type of representation used for recording and/or displaying the analysis is another factor which could contribute to the classification of the annotation schemes. Here, a first rough distinction can be drawn between vertically and horizontally organised corpora analyses.

The first case is represented by the so-called "one-word-per-line" format where each line, containing the information for one word, is in turn segmented into different fields: each field is assigned a different kind of information, going from the reference to the text and cross-references to other corpora, to the wordform and the respective lemma, to the morphosyntactic and/or syntactic analyses.

The second case is represented by the horizontal format in which the text words and the analysis, usually expressed by means of labelled brackets, are interspersed on a single line; in this format, each text word can be optionally followed, after an underline character, by its part of speech tag.

It should be pointed out that the labelled bracketing representation is implied by the constituency-based model of syntax. Usually constituency is represented in the form of tree diagrams or of labelled bracketing (encoding the same information as a tree, but presenting it linearly). Therefore dependency-based annotation schemes will not be likely to use this kind of representation. The labelled bracketing representation is implied by the horizontal format, but can also be used in the vertical format.

4.4 *Corpus annotation schemes as property bundles*

In the table below, each annotation scheme is described as a bundle of features; the features used for this definition are the parameters identified at the previous stage as relevant for annotation scheme classification, and were briefly illustrated in the section above. Unfortunately, the documentation on which this study is based does not always provide the necessary information, and so it has not always been possible to present an exhaustive description of the different annotation schemes with respect to the single parameters examined.

	Nijm	ICE	LaLe	LOB	La87	Lask	Su	Goth	PWC	Penn	BECG
Const	+	+	+	+	+	+	+	-+	-	+	-
Depen	-	-	-	-	-	-	-	+	-	-	+
Categ	+	+	+	+	+	+	+	+	+	+	-
Funct	+	+	-	-	-	-	+	+	+	-	+
Ambig	+	(+)	?	?	?	?	?	?	?	+	+
Unlab	+	(+)	+	?	?	?	?			+	
Deep	(-)	(-)	-	-	(-)	(-)	+	+	+	(-)	-
Skel	-	-	-	-	-	+	-			+	
Flat	+	+	+	+	-	+	+			+	
Real	?	?	+	+	?	?	?	-	-	?	?
Horz	*	?	-	+	+	+	-	-	+	+	-
Brlab	-	?	+	+	+	+	+	-	?	+	-

+ the feature is included in the annotation scheme
- the feature is not included in the annotation scheme
* lateral position of the annotation scheme with respect to the parameter
( ) no explicit information with respect to the parameter; the information between parentheses has been inferred from the observation of excerpts of the analysed corpus
? neither explicit nor implicit information with respect to the parameter empty cell the parameter cannot be applied to the annotation scheme

Rows labels:

A	Const	constituency-based representation
A	Depen	dependency-based representation
B	Categ	categorial labelling
B	Funct	functional labelling
C	Ambig	ambiguity representation
D	Unlab	unlabelled bracketing
E	Deep	deep structure representation
G	Skel	skeletal parsing
H	Flat	flat trees
I	Real	treatment of specific phenomena to real text
J	Horz	horizontal format representation
J	Brlab	labelled bracketing representation

4.5 *Related issues*

At this point it is worth referring to two issues which emerged during the survey of the parameters proposed for classifying syntactic annotation schemes. They are not directly related to classification and comparison, but we think that they contributed indirectly to the final characterization of the syntactic annotation schemes. They concern on the one hand the methods adopted for annotating corpora, on the other the uses of syntactically annotated corpora: it is unquestionable that these two issues affected one way or another the resulting scheme of annotation.

4.5.1 *Methods adopted for annotating corpora*

From the methodological point of view, annotations may be added automatically (with a rule-based or a probabilistic parser) with manual post-editing, or inserted manually with varying degrees of interactive help. Even if this issue is not directly relevant in this context, we think that the technique used for producing the annotation is more or less closely linked to some of the peculiarities of the parsing scheme adopted; not all the parsing schemes can be easily handled by the parsing systems, especially when the analysis is to be performed on real texts.

In about half of the corpora examined, the syntactic annotation was produced manually, and not as the output of an automatic parsing system. This holds for Gothenburg, Susanne, Lancaster-Leeds, Lancaster-IBM 1987, Lancaster-IBM, and the Polytechnic of Wales. In Penn, Nijmegen, and LOB, on the other hand, annotations were added automatically with manual post-editing; in the first two by rule-based systems and in the latter by a stochastic one. For the Constraint Grammar, it is obvious that the parsing scheme described here corresponds to the output of the parsing system. In the International Corpus of English, which is still at the project stage, most of the work will be done interactively, by having a computer produce all the options; the final decisions will have to be made by humans.

4.5.2 *Uses of syntactically annotated corpora*

One of the main goals of the construction of syntactically annotated corpora concerns the development of statistics-based automatic parsing techniques. As pointed out with respect to the parameter H, not all the parsing schemes are equivalent in terms of these techniques. Therefore, when evaluating and classifying annotation schemes, the purpose of the annotation should be taken into account. Behind different uses there are conflicting needs: detailed linguistic analyses require finely-grained annotation schemes; coarse-grained annotations, on the other hand, are better suited to the training phase of probabilistic grammars. As Leech points out in this respect (Leech, 1992), "it is important, in one's general approach to annotation schemes, to allow for variable delicacy as one aspect of descriptive variability of annotation schemes".

4.6 *Towards standardization: recommendations and directions of work*

The aim of this part of the study was to make a description and comparison of actually existing

syntactically annotated corpora, and of the underlying approaches. This phase laid the foundations for evaluating the feasibility of proposing standards for this level of linguistic description, a task which could be further carried out in the EAGLES project.

The goal of defining a common interchange standard for syntactic annotation has a peculiar characterization, differing from e.g. the morphosyntactic annotation level where it was possible to identify a core of features common to all the annotation schemes examined (see Monachini and Oestling, 1992b, NERC-61). For the syntactic level an integration of the different annotation schemes (as they are now) within a single, unvarying framework compatible with all of them is, in our opinion, an almost impossible objective, given the situation set out in the previous sections. The factors contributing to the definition of the syntactic annotation schemes are too many to be inserted simultaneously into a single, coherent framework without conflicts or mutilations for one or another of the annotation schemes.

From this perspective, the direction to be followed for the definition of standards is that of verifying the compatibility of the different annotation schemes, rather than their conformity. This means that the research should be directed towards the evaluation of whether and how the different annotation schemes are intertranslatable, rather than trying to build a unique coherent framework into which all of them are subsumed. The only explicit indication of the possibility of translating one annotation scheme in terms of another comes from (Karlsson, 1990) who points out that a constituency-based representation can be easily derived from the Constraint Grammar annotation scheme, which is dependency-based. It should be noted that this indication is restricted to one of the parameters which have been taken into account, the syntax model behind the annotation scheme, and that - in our opinion - it is doubtful whether the reverse is also true. Nevertheless, it can be seen as an encouraging step in the direction of standards as compatible representations.

Defining a standard as an overall framework of compatible representations is related with a crucial issue, the theory neutrality requirement. One of the maxims for annotators proposed by Leech (namely, the fifth one) claims that "annotation schemes should preferably be based as far as possible on 'consensual', theory-neutral analyses of the data" (Leech, 1992). Here, the theory neutrality requirement acquires a broader meaning. As said before, the research into a theory neutral representation of core grammatical phenomena, mediating between different annotation schemes (in their turn inspired to different grammar theories), is a controversial and almost impossible objective. The idea of a standard, as proposed here, is theory neutral in the sense that it includes all primitive basic features representing the building blocks of different annotation schemes, inspired by different grammatical theories. Such a representation, in spite of being theory neutral, could not still account for peripheral constructions. Therefore, in corpus-based research, a theory neutral representation has also to fill the gap between language as ideally drawn by grammatical theories and as actually attested by real-life usage. The representation at this level, not mediated by any theoretical model, should stick to the actual phenomena, and in this sense be "neutral" with respect to theories.

The compatibility of representations can be verified - according to the methodology set up here - by dissecting them and finding out the basic features they make use of. From this perspective, a redundancy check directed towards verifying the relatedness of the features individuated, and particularly their mutual implications, is a crucial step in the standard definition, which could help to reduce the features of the standard to the essential ones only. We

tried to answer this point, whenever possible, in the course of the study.

At the present stage of research, a standard over annotation schemes modelled on different families of theories (mainly constituency- and dependency-based) seems to require the identification of the primitive basic features - or parameters - starting from which the single schemes could be reconstructed, with their individuality. The first step can consist in assessing the feasibility of reducing annotation schemes belonging to different families to a set of primitive features; the configuration of the features to be activated varies according to the model behind the annotation scheme.

Obviously, a standard can be more easily defined over annotation schemes modelled on the same kind of grammar theory: different but homologous annotation schemes vary mainly as to the number and type of syntactic constituents they recognize, but the representations are expected to be compatible in the end. A classification of shared grammatical concepts could be seen as the next research step towards the definition of standards.

Encouraging results in this direction emerged from the activity of the Group on Evaluation of Broad-Coverage Grammars of English, whose documentation has been kindly provided us by Mark Liberman. The research project of this group - Parseval - aims at developing criteria, methods, measures and procedures for evaluating the syntax performance of different broad-coverage parsers/grammars of English (see Harrison et al., 1991, Abney et al., 1992). This project has been motivated by the difficulty of comparing different grammars because of divergences in the way they handle various syntactic phenomena, such as the employment of null nodes by the grammar, the attachment of auxiliaries, negation, pre-infinitival "to", adverbs and other kinds of constituents, as well as punctuation. What is of interest in our context are the methodologies they developed in order to make the different analyses comparable, based on the systematic elimination from the parse trees of such problematic constructions.

As far as the syntactic labelling is concerned, the kind of labelling - categorial and/or functional - depends on the syntactic model behind the annotation scheme. Optimally, in a standard both of them are required; anyway, the standard should also provide the possibility of selecting just one of the two. A standard should also provide the possibility of handling ambiguities and partial analyses, both during intermediate analysis stages and in the final result, that is within the annotated corpus.

With respect to the granularity of the analysis, in the standard definition it should be taken into account that the optimal degree of delicacy is application dependent, since the purpose of an application can require distinguishing particular information which may not be relevant for other applications. Two opposite tendencies have been recognized in this respect: skeletal parsing (that is using a minimal set of basic syntactic categories) vs. detailed annotation schemes. The first approach, while satisfying the theory-neutrality requirement, improves the consistency and speed of the annotation process, and speeds up the training phase of stochastic grammars. On the other hand, the second one is better suited to cover and distinguish the variety of linguistic phenomena usually occurring in real text. Therefore, variable delicacy should be allowed in the standard according to application requirements. This implies using variable parsing schemes, ranging over skeletal and detailed representations.

The same variability should also be allowed with respect to the depth of the analysis; whenever needed, deep representations should be associated with surface representations. Obviously, the standard should also provide a suitable representation for phenomena specific to

real text, such as punctuation, postal addresses, money sums, dates, weights and measures, bibliographical citations and other comparable phenomena.

The framework which emerged from this survey of the current practices in annotating corpora at the syntactic level can also be seen as the background to the negative conclusions with respect to a direct use of existing NLP annotation systems in corpus-based research, proposed by the case studies by Antona and Ruimy (see Antona, 1992a, NERC-64, Ruimy, 1992, NERC-65). These studies, taking into account the analysis schemes adopted by the Eurotra project for machine translation, try to assess the feasibility of their direct use in corpus research. Such analysis schemes, when exported as they are, show all the limitations typical of grammar models when confronted with unrestricted text. In any case, we think it would be very useful to consider in this context the theoretical investigations carried out in NLP projects, even though they are not directly exportable to cover unrestricted text phenomena. Their results, for instance, could be exploited in devising the annotation of particularly problematic constructions (see Antona, 1992b, NERC-63).

What came out from this phase of research is a restricted and rough set of guidelines which can be used as a starting point for further studies assessing the feasibility of standards for syntactic annotation, and proposing actual directions for further work. The fact that we have limited and heterogeneous information, and that we are operating on schemes conceived only for English makes these guidelines partial and incomplete, but at least they constitute a core to start with.

References

Aarts J., Van Der Heuvel T. (1984): "Linguistic and computational aspects of corpus research", in Aarts J., Meijs W., (eds), *Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English Language Research*, Amsterdam, Rodopi, pp. 83-94.

Aarts J., Van Den Heuvel T. (1985): "Computational tools for the syntactic analysis of corpora", *Linguistics*, 23, pp. 303-335.

Aarts J., Oostdijk N. (1988): "Corpus-related research at Nijmegen University", in Kyto M., Ihalainen O., Rissanen M., (eds), *Corpus linguistics, hard and soft*, Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora (ICAME 8th), Rodopi, Amsterdam, pp. 1-14.

Abney S., Black E., Flickinger D., Gdaniec C., Grishman R., Harrison P., Hindle D., Ingria R., Jelinek F., Klavans J., Liberman M., Marcus M., Roukos S., Santorini B., Strzalkowsky T. (forthcoming): *A quantitative evaluation procedure for English grammars.*

Atwell E. (1987): "Constituent-likelihood grammar", in Garside R., Leech G.,Sampson G., (eds), pp. 57-65.

Atwell E. (1988): "Transforming a parsed corpus into a corpus parser", in Kyto M., Ihalainen O.,

Rissanen M., (eds), *Corpus linguistics, hard and soft*, Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora (ICAME 8th), Rodopi, Amsterdam, pp. 61-69.

Garside R., Leech G. (1987): "The UCREL probabilistic parsing system", in Garside R., Leech G., Sampson G., (eds), pp. 66-81.

Garside R., Leech G., Sampson G., (eds) (1987): *The computational analysis of English. A corpus-based approach*, London, Longman.

Harrison P., Abney S., Black E., Flickinger D., Gdaniec C., Grishman R., Hindle D., Ingria R., Marcus M., Santorini B., Strzalkowsky T. (1991): *Evaluating Syntax Performance of Parsers/Grammars of English*, Proceedings of the Workshop on Grammar Evaluation, ACL 1991.

Hudson R.A. (1980): "Constituency and dependency", *Linguistics*, 18, pp. 179-198.

Karlsson F., Voutilainen A., Anttila A., Heikkila J. (1991): "Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text, with an Application to English", in *Natural Language Text Retrieval. Workshop Notes from the Ninth National Conference on Artificial Intelligence*, American Association for Artificial Intelligence, Anaheim, Cal.

Karlsson F. (1990): "Constraint Grammar as a Framework for Parsing Running Text", in Karlgren H., (ed.), *Proceedings of the XIIIth Conference on Computational Linguistics*, Helsinki, Vol. 3, pp. 168-173.

Karlsson F. (1992): *Lexicography and Corpus Linguistics*, Opening Address at 5th Congress of Euralex, Tampere, August 4, 1992.

Leech G., Garside R. (1991): "Running a grammar factory: the production of syntactically analysed corpora or 'treebanks'", in Johansson S., Stenstrom A.B., *English Computer Corpora: Selected Papers and Research Guide*, Berlin, Mouton de Gruyter, pp. 15-32.

Leech G. (1992): *Corpus Annotation Schemes*, Paper presented at the "Workshop on Textual Corpora", Pisa 24-26 January 1992.

Marcus M.P., Santorini B. (1992): *Building very large natural language corpora: the Penn Treebank*, Paper presented at the "Workshop on Textual Corpora", Pisa 24-26 January 1992.

Sampson G. (1987): "The grammatical database and parsing system", in Garside R., Leech G., Sampson G., (eds), pp. 82-96.

Sampson G. (1987): "Evidence against the 'grammatical' / 'ungrammatical' distinction", in Meijs W., (ed), *Corpus Linguistics and beyond*, Amsterdam, Rodopi, pp. 219-226.

Sampson G. (1989): "How Fully Does a Machine-Usable Dictionary Cover English Text?", *Literary and Linguistic Computing*, Vol. 4, No. 1, 29-35.

Sampson G. (1991): Needed: *a grammatical stocktaking*, Paper presented at the "Workshop on Textual Corpora", Pisa 24-26 January 1992.

Sampson G. (1992a): "Analysed Corpora of English: A Consumer Guide", in Pennington M.C., Stevens V., (eds), *Computers in Applied Linguistics*, Clevedon, Multilingual Matters, pp. 181-200.

Sampson G. (1992b): *The Susanne Corpus*, Release 1, 6th September 1992.

Van Den Heuvel T. (1987): "Interaction in Syntactic Corpus Analysis", in Meijs W., (ed), *Corpus Linguistics and beyond*, Amsterdam, Rodopi, pp. 235-252.

Van Den Heuvel T. (1988): "TOSCA: An Aid for Building Syntactic Databases", *Literary and Linguistic Computing*, Vol. 3, No. 3, 147-151.

Van Halteren H. (1992): "Syntactic Markup in the ICE project", in *Conference Abstracts and Programme* of the 19th International Conference of the Association for Literary and Linguistic Computing (ALLC) and the 12th International Conference on Computers and the Humanities (ACH), Christ Church, Oxford, April 1992, pp. 33-35.

Van Halteren H., Van den Heuvel T. (1990): *The linguistic exploitation of syntactic databases. The use of the Nijmegen LDB program*, Amsterdam, Rodopi.

Voutilainen A., Heikkila J., Anttila A. (1992): *Constraint Grammar of English. A Performance-Oriented Introduction*, University of Helsinki, Department of General Linguistics, Publication n. 21.

Relevant NERC Papers

Antona M. (1992a): "A Comparison of Eurotra ECS Grammars", Working Paper, ILC, Pisa, NERC-64.

Antona M. (1992b): "The treatment of subordinate clauses in Eurotra. An overview", Working Paper, ILC Pisa, NERC-63.

Corazzari O. (1992): "Phraseological Units", Working Paper, ILC Pisa, NERC-68.

Monachini M., Östling A. (1992a): "Morphosyntactic Corpus Annotation - A Comparison of Different Schemes", Technical Report, ILC Pisa, NERC-60.

Monachini M., Östling A. (1992b): "Towards a Minimal Standard for Morphosyntactic Corpus Annotation", Technical Report, ILC Pisa, NERC-61.

Montemagni S. (1992): "Syntactically annotated corpora: comparing the underlying annotation schemes", Technical Report, ILC Pisa, NERC-67.

Ruimy N. (1992): "The Argument Structure in Eurotra: General Principles and Applications", Working Paper, ILC Pisa, NERC-65.