NERC-1 CONSORTIUM

Birmingham	The University of Birmingham
Leiden	Instituut voor Nederlandse Lexicologie (INL)
Malaga	Universidad de Malaga
Mannheim	Institut für Deutsche Sprache (IDS)
Paris	Institut National de la Langue Française (INALF-CNRS)
Pisa (Coordinator)	Consorzio Pisa Ricerche, Istituto di Linguistica Computazionale (ILC-CNR), Dipartimento di Linguistica dell'Università di Pisa

The Initial Chapter (0), which is based on recommandations of all the NERC partners, has been drafted by J.Sinclair and A.Zampolli, with the assistance of an editorial committee including P.Lafon, W.Teubert, P.Van Sterkenburg.

Each of the following chapters (1-7) is the result of one or two of the 10 workpackages in which the NERC1 study was articulated.
Each workpackage was the responsibility of a specific partner, who also drafted the corresponding chapter, or section of chapter.

The table below indicates the correspondences between chapters, workpakages, and responsible partners.

Ch.	WP	
1	1, 2	IDS - Mannheim: User Needs
2	6	INL - Leiden: Corpus Design Criteria
3	3	INALF - Paris: Text Representation: Written Language
	4	University of Birmingham: Text Representation: Spoken Language
4	7	INALF - Paris: Text Acquisition and Reusability
	5	University of Birmingham: Access and Management Software Tools
5	8	ILC - Pisa and University of Pisa: Linguistic Annotation of Texts: scientific and technical problems; guidelines for harmonization
6	9	University of Birmingham: Corpus Annotation Tools
7	10	ILC - Pisa and University of Pisa: Knowledge Extraction

Chapters 1-7 have been assembled and edited by an editorial committee composed of N.Calzolari (chair), M.Baker, T.Kruyt. The editorial committee was helped by M.Monachini and P.Orsolini.

Table of Contents

Chapter 0

Implementation Plan

1 Introduction

NERC (Network of European Reference Corpora) is a feasibility study which has the objective of making recommendations to the EC about the future of language corpus provision in Europe. It began in a very small way, at a time when corpus matters were not very important in the language industries and in the field of Natural Language Processing. There were not many institutions in this field which had established corpora and there was little common ground among those who had built up experience. NERC was not part of an ongoing Plan, but an exploratory sideline.

That time was less than four years ago. During the period of the project the central value of corpora in the study of language has been recognised by more and more people of influence in the field, and corpora now figure prominently in the research of many groups, in the LRE and other programmes, and in the Fourth Framework of the EC. The NERC Workshop in January 1992 - postponed for a year because of the Gulf war - attracted a very distinguished group of participants and brought out most of the issues that are dealt with in this Report.

But even after one year the debates and concerns of the workshop are getting out of date, such is the pace of advance of people's thinking. The DG XIII report entitled "Language and Technology" dated September 1992 (EC DGXIII, 1992) provides an interim reference point. The ambitious programme envisaged there is increasingly seen to require underpinning from extensive corpus resources in many languages.

When NERC was being designed, corpora ranged in size from 1-20 million words, mostly in the lower end of the range. Open-ended and Monitor corpora were very distant prospects; now they are regarded as essential. Corpora two years ago were monolithic. They had probably been carefully put together, but their sub-components would not stand on their own as representative of genres within the corpora. Nowadays this is expected of the big corpora.

Annotation when NERC began was mainly at the level of adding codes to a text in order to facilitate first-line handling of it. A few research institutions were developing taggers and parsers, but there was virtually nothing "off the shelf", and most analytical software had problems. The last two years have seen the provision of large amounts of English text with grammatical tags attached and taggers are now available for a number of other languages.

The different languages of Europe were treated very differently with respect to corpora. English was well provided for (and for once, not because of American involvement). French was ahead of the field in the construction of a historical-literary corpus reference. German had a substantial corpus resource that was well designed for growth and progress, and a number of smaller ventures of interest. Italy had a strong stake in corpus linguistics and a multi-million word corpus. Dutch had more than one big corpus and one of the leading analytical research groups. Corpus work was beginning in Spanish. Less prominent languages had distinguished themselves with special collections - Frisian, for example. The design of each corpus bore no resemblance to any other; the

software used to handle the corpora was specific to the corpus, the institution and even the machine (and on the whole it was pretty primitive by today's standards). It was a major achievement in 1990 to exchange even a few concordance lines.

There was very little pressure from potential users or support from visionary sponsors. Lexicography was the first area where corpus work took root and proceeded beyond the limits of academic research enquiry. Even in lexicography there was some opposition to accepting the authority of corpus evidence. For other applications the scale of corpus needed was usually beyond the dimensions of a particular job, and the sophistication of the retrieval tools did not measure up to the needs of the developer. The situation was clearly one that required a medium-term policy of precompetitive provision of resources, to allow the corpus experts time and space to meet the demands that could in 1991 be foreseen.

NERC can report a considerable movement in that direction, partly through various funded projects but mainly through a major process of reconceptualisation among the people concerned. It has gradually been realised that even apparently simple applications of language technology could not reach an acceptable level of performance without fundamental corpus research. The importance of detail became recognised, and the unfolding of detail in the growing corpora gave promise of much sharper descriptions, for applications such as machine translation.

It has been necessary for many scholars to engender a new respect for the way people actually talk and write. For a generation linguistics has been dominated by a study of the potential of language at the expense of the actuality, which was thought by many to be trivial. The variability of language in use could not be captured in formal models, and the tasks that linguistic engineers were being asked to perform depended on an accurate depiction of language in use. Hence the move towards corpora, initially to inform models which were too rigid in their expectations, and now more and more as a source of new models and insights which were unobtainable in the past.

2 Proposals: Corpora

The framework for corpus development in Europe is in place; there is a general expression of approval and encouragement, and the EAGLES[1] project to set standards and conventions that will steer funded projects towards harmonised practice. The detailed recommendations of this Report can be fed in to EAGLES, and revised in the light of progress over the duration of EAGLES. There is at least a safety net for the future. NERC wishes to argue that in spite of the existing provision of programmes such as EAGLES, LRE and ESPRIT, a different sort of investment is required to ensure the proper and efficient development of language industries in Europe over the next five years.

In the first instance, the disparity in treatment of the different languages of the community needs to be addressed, and inequalities sorted out.

There should be a statement of target corpus provision for each of the official languages of the

1 EAGLES (Expert Advisory Group on Language Engineering Standards) is an LRE project funded by the EC and started on February 1993.

Community, another for each of the recognised languages, and a third for the remaining indigenous languages and any other languages which the Community decides to take an interest in. For example, in the preparations for another country joining the Community there might be an opportunity to assess the position of the languages used in that country, and their availability in other parts of the Community. There is no need to wait until a country has joined the Community to prepare its language for the place.

At this time NERC feels it can best elaborate proposals for the languages of the EC. In no way is this intended to hinder the development of the other languages of Europe, which will be substantially supported by the progress that is proposed here. NERC policy will put into the public domain all the necessary experience and software for any language to initiate corpus development, and EC programmes will we hope from time to time support stages in this pan-european movement.

The strategy proposed is in three stages, building on a very general memorandum that emerged from the NERC Pisa Workshop of January 1992, incorporated into a Strategic Briefing Paper in April 1992 (NERC Consortium, 1992, NERC-99). Delegates to that workshop included a strong American group and individuals from other European countries such as Sweden[2].

Implementation: Stage One[3] (Duration: 10 months)

In the immediate future there is a proposal which will harmonise the work of the partners, including the new partners, and prepare the ground for Stage Two. NERC is a feasibility study whose major finding is that a Network of European Reference Corpora is both feasible and desirable, and is likely to be a central platform for development of language work in the EC. The Strategic Briefing Paper's first formulation of a practical step forward envisaged two parallel thrusts:

(a) The speedy provision of corpus data in the languages of Europe;

(b) The establishment of a physical network, initially on a pilot basis.

---

2 The prefinal draft of this report, including this chapter, has been very positively evaluated by the reviewers appointed by the EC in June 93. In particular, the 3 stage approach has won the unanimous approval of the evaluators. Another confirmation of the soundness of this proposal is indirectly given by the fact that some projects, which are going to be launched by the EC in the field of Linguistic Resources, include in their programme of work, in a partial way, some of the objectives recommended by NERC. We have already ensured appropriate links with these projects in order to have the possibility of evaluating whether some of their results could be used to facilitate our program of work or to establish synergies and optimize efforts.

3 In this report, stages 1 and 2 are presented sequentially. This proposal reflects the state of affairs and the information about the possible timetable for the development of the Language Technology program we had one year ago. Given the development in the field and of our centres, it now appears that the organizing and funding of stage 1 and 2 could overlap.

To achieve these aims, the following strategy is proposed:

(a)That a multi-million word sample of each of the nine official languages of the community is made available from EC electronic files of journals and official documents.

(b) That a pilot network is established with eleven nodes, one in the institution of each of the partners in the NERC Consortium (henceforth corpus centres).

(a) EC Parallel Texts

It is convenient that the EC has prepared large amounts of text material translated into the nine official languages. This constitutes a unique text collection which can be put to immediate practical use in NERC. The opportunities for research in alignment and translation are considerable, and the collection in each language provides ample data for the establishment of a corpus centre. As a means of harmonising work in the nine languages, the EC text collection offers a lot of promise.

The tasks envisaged, with costings, are given below.

It must be pointed out that the use of parallel texts to represent a corpus is for the first, pilot, implementation stage only. There are limitations, not yet fully understood, about the use of translated material as representative of the usage of the target language. Also the genre and subject matter of EC publications is somewhat restricted and the language is inevitably specialised. However, the assembly of such parallel corpora will not be wasted as Stage Two succeeds Stage One, since an important resource will be made available for translation projects etc.

CD-ROM.

At the Pisa Workshop in 1992, it was thought that the compilation and distribution of existing material on CD-ROM was a simple and inexpensive matter; hence NERC endorsed the move to provide such material as an emergency measure. We understand that there are in fact problems of organisation, technology and finance which have introduced delays into the programme. Nevertheless, it may still be possible to produce a few interim CDs (a CD-ROM of the European Corpus Initiative has in the meantime come out) which will fill two needs:

(i)for those languages for which no corpus is easily available, some data on which to experiment

(ii)for some NLP applications which require simple data streams rather than reference corpora, more data in a variety of languages.

NERC continues to commend the production of CD-ROMs in the coming months. It is important that the sources and types of language text are made clear, particularly when data originates other than in the normal business of human communication but is derived for example from experimentation or artificial recording circumstances.

(b) Network proposal

NERC argues the case for networking access to corpora, and has established the essential feasibility of a European network. The initial network will be the eleven partners, but each partner will enrol other users of the network as soon as the pilot stage is over. Common procedures are essential and the partnership will establish a de facto standard.

Responsibility for the work will be shared among the partners, with each carrying out the folllowing tasks:

(i)A directory to the network will be established, using X-500 protocols.

(ii)A common set of entry protocols will be established, covering centre-to-centre physical connections, centre identification, log-in conventions, passwords, data transfer protocols, log-off conventions.

(iii)A small subset of the standard query language (see Chapter 4) will be implemented in as many centres as possible.

(iv)At least one experiment will be conducted in each centre, with the aim of applying software obtained from another centre to the local corpus.

A report will be written including an evaluation of the future possibilities of (iii) and (iv), including costings. A section of the report will deal with procedures for extending the network to include non-centre users, and costings for opening up the network to the whole user community.

(c) Costs[4]

[4] We suggest that the EC meet the cost listed here for stages 1 and 2, which represent less than half of the total cost required. The NERC Centers, and through them the national authorities supporting them, will meet the other half (infrastructure, linguistic software, date, etc.).

The estimated costs of the work of Stage One are:

A. STAGE 1 (short-term)	m/m	ECU
A1. CREATION OF MULTILINGUAL CORPORA (PER LANGUAGE)		
Creation of software for conversion and formatting (from the source data in photocomposition to the common NERC format)	(1m/m)	4,000
Automatic conversion and manual checking of the converted data: 10M word parallel corpus	(6m/m)	16,000
TOTAL		20,000
A2. NETWORKING (PER INSTITUTION)		
- Acquisition of the relevant software.		2,000
Production of the documentation (in English) concerning the corpus data and the corpus access software available at each Institute.	(2m/m)	8,000
- Design of a common query language.	(1m/m)	20,000
Implementation of the common query language, i.e. software able to translate a query on the corpus from the common query language to the local access software.	(4m/m)	
TOTAL		30,000
A3. TOTAL STAGE 1 PER LANGUAGE		50,000

Implementation: Stage Two (Duration: 30 months[5]**)**

This stage is designed to achieve across a network the provision of substantial representative corpora in each of the official languages of the EC. Building on the co-ordinated experience of Stage One, and following the recommendations of NERC as modified by EAGLES, the Corpus Centres will set up reference corpora. We propose an initial target of a corpus of fifty million written and one million spoken words, carefully selected to represent the major categories in use, and following the design parameters recommended in Chapter 2. One million words will be annotated at the morphosyntactic level, according to the recommendations of Chapter 5.

- There should be a balance between public and private language, spoken and written, considered and impromptu, bearing in mind the practicalities of gathering substantial amounts of some kinds of text.

- All major genres should be represented, and definitions of them should be agreed so that comparisons across languages can be made with accuracy.

- Each text in the corpus will be identified by a substantial set of information about its origins and provenance, which will probably be stored separately and associated with a TEI header or similar.

- The initial corpus in each case should be designed so that it can be regularly extended and developed, with the aim of becoming a monitor corpus when conditions are appropriate for the language in question. See Stage Three for the continuation of this work into monitor corpora

Each corpus will be accompanied by up-to-date software to provide basic access and retrieval (see Chapter 4). The software for each language will be compatible with the others so that a user does not need to learn several sets of conventions; a standard query language is proposed in Chapter 4, which will be the basis of a common standard.

Networking will allow easy and flexible access by users to a range of languages and software, as discussed in Chapter 4.

Special expertise gained in one participant's institution will be shared among all users and, where possible, all languages.

---

5 Access to the textual data will he made possible already during the development of the reference corpus without waiting for the completion of the full corpus.

The estimated costs of the work of Stage Two are:

B. STAGE 2 (medium term)

B1. CREATION OF A CORE COMMON CORPUS (PER LANGUAGE)

- 1M words Spoken texts

Recording (20,000 ECU) and Transcription (60,000 ECU)		80,000

- 50M words Written texts

Data identification and appropriation	(12m/m)	50,000
Data acquisition and conversion	(50m/m)	160,000
Software, travel, general costs, etc.		80,000
TOTAL CORPUS CONSTRUCTION (PER LANGUAGE)		370,000

B2. ANNOTATION (PER LANGUAGE)

Automatic annotation (tagging), with the interactive checking of about 1,000,000 words

- software for annotation (tagger) and automatic annotation	(10m/m)	40,000
- interactive validation (of 1,000,000 tagged words)	(15m/m)	50,000
TOTAL ANNOTATION		90,000

B3. TOTAL STAGE 2 PER LANGUAGE	460,000
B4. Overall European Coordination	300,000

Implementation: Stage Three

In the volatile world of corpus linguistics it is difficult to plan ahead in detail. Several of the assumptions made may well be challenged by technological breakthroughs or research findings in the months between planning and implementation. But in this instance we believe that there is a natural progression of events to which we can point. The primary aim of the third stage is to enable the Corpus Centres to achieve stability over time, coping with the immense amounts of reusable data that will become available, maintaining and upgrading their corpus holdings, offering a steady, quality service to users and acquiring authoritative status as centres of expertise, advice and support for all kinds of corpus work. This is expressed in summary as follows:

To establish a monitor corpus in each of the Corpus Centres for the official languages of the EC.

First of all, the central corpus will be gradually surrounded by more recent material coming in; it will be updated and enlarged from these accretions on a regular basis, following a policy that will be co-ordinated across the participating languages.

A Monitor Corpus is a corpus in dynamic mode (Sinclair 1991, Clear 1988). While it will frequently be necessary to identify a particular set of texts as a corpus for an application, the flow of language from electronic sources will increase dramatically for most languages over the next decade. A corpus provider will not be able to ignore this, and indeed should welcome the change from the days when all material had to be keyboarded for entry to a computer.

However, it is unlikely that the sources of electronic data will provide just the right quantities of material to maintain on a daily, weekly etc. basis a balanced perspective on the language as a whole, or on any sub-corpus that is required from time to time. A monitoring routine will be required to maintain awareness of the input and select what is needed. The rest need not be abandoned, because there are some applications where quantity is all important.

The flow of available data is expected to exceed the ability of the software to process it with the sophisticated analyses that are expected to be standard in the coming years. Hence one set of tools that will be required are *filters*, applied as the data enters the system, which deal with new word forms, combinations etc., and upgrade statistical records.

In summary, then, the targets above fifty million words will not be expressed as a finite size but as a rate of flow. At the points of reception of new material the beginnings of a monitor corpus will take shape, and some analysis will be done on-line as text comes in, thus giving two complementary pictures of the language to users of the corpus resource. The balance will slowly shift to the flow of

language coming in, and the design parameters of the corpus will change from defining the size of finite blocks of text to defining the rates of inflow of material. At any one time there will be a large static corpus available, frequently updated from the monitored flow.

The concern of NERC is with the official languages of the EC. However it is hoped that support and provision will be made for the other languages, to preserve the linguistic heritage and maintain regional cultural identity. Our current proposals do not disadvantage the other languages, but set up systems which can be used by the other languages in due course. The networking proposals of Stage One provide a core which is extendable to include many other languages. The corpus creation of Stage Two and the access software are valid for any other language. The technology of Stage Three will be applicable to any language, and some important regional languages may move towards a monitor corpus with considerable speed.

Costs: These cannot be worked out in detail yet for Stage Three. Further reference is made to the cost of monitor corpora and regional languages below.

Proposals: Costing Principles

We must consider how all this is to be achieved, politically and organisationally. Who will pay for it, who will co-ordinate it, who will evaluate it? What is the role of national governments and how do nations which share languages sort it out?

First some guiding principles:

1. Corpora and associated software must be made available to all who have a professional use for them.

- Every effort must be made to overcome difficulties about such matters as copyright, and trade practices that reduce the availability of language material.

- Language text in machine readable form is both too valuable to throw away or keep from circulation, and also too trivial to attract large payments.

- The proposed solution is to strive for a set of changes in attitude to text, charging for it and handling text in the production of documents.

- The originators of text should understand that the linguistic need for their texts is not for what is normally considered of value in a text (the content, the style, the story, the recommendations etc.) but merely for instances of language in use by writers or speakers of normal fluency. A code of "good practice" with regard to intellectual property rights should be promoted.

- The fees for access to such material can only be small per million words, and perhaps should be built into authorial contracts. It should be understood that the provision of raw data and structured data have a different commercial value.

- Those who make electronic copies of texts using a keyboard or scanner should be encouraged to adopt work practices that facilitate the reuse of the texts in corpora. In some cases this can be handled by contract again, and in other cases by very modest training.

We look to both, EAGLES and RELATOR, to make firm proposals in this area. Already the heavy costs of cleaning up typesetters' tapes are forcing corpus providers to return to scanning, which was the fashion about ten years ago, and far too much keyboarding is done of material which already exists in machine readable form.

2. Access to corpora by users should be financed by fees charged to users.

- The fees would not reflect the interest of the originators, nor the start-up costs nor the development costs, which would be funded from other sources. They would simply reflect the time and trouble taken to maintain a service at an agreed standard.

- Grants or other special support would enable some classes of users (notably students) to afford a realistic level of cost.

3. START-UP and INFRASTRUCTURE. EC should meet the costs involved in Stage One and Stage Two. These proposals essentially concern the co-ordination of work at European level, the harmonisation and standardisation of the provision of corpus access and facilities, and the support of the Community for languages that have not been developed as corpus resources. Although they involve some investment in language of national importance that investment is only the minimum necessary to establish the common networks, protocols and procedures. In order to create a network of Corpus Centres, there must be a non-trivial corpus resource available at each Centre, strictly comparable with the others. The other facilities are the provision of the specific software and hardware necessary for mounting and updating the facilities in the various languages, the co-ordination of networks and application of standards, and the smoothing of inequalities in stages of development and in currency fluctuations among the language communities.

In Stage One the corpus will be provided by the EC. In Stage Two the proposals are for a common design which can be implemented at a low cost because it requires only very small quantities of material which is expensive to acquire.

4. Ultimately we must expect that some of the cost of providing corpus material in a language (excluding the software, hardware and networking costs) should be met by the community of prime users of the language - that is to say, those who will benefit most from the acquisition of knowledge about the language, its easy translateability into other languages, its availability in a wide range of products of the language industries. Hence we regard the full specification of Stage Three and its costing principles as something to be approached after consultation with national governments. Some notes are appended to indicate NERC's position.

- Nowadays, as this Report shows, the costs of acquiring most types of material are small, and even the more expensive types, such as informal conversations, is not a large item in a language community's cultural budget.

- However, it is not necessary that the whole burden should fall on the taxpayers. Language is a badge of culture and commands a lot of emotional support. Projects of this kind should attract private sponsorship, including appeals, bequests and donations from the communications industry. Wealthy individuals and foundations which promote regionalism and the survival of small cultural entities should be willing sponsors of projects designed to enhance the position of their language.

- In identifying the necessary software, we wish to make a distinction between language specific software and software of general handling utility. The former is considered to be research activity and of prime concern to the community of users; hence it should be financed from national or cultural sources. The latter, which is applicable to virtually any language, is reasonably financed from centralised sources.

- Multilateral development agencies, and possibly some bilateral ones as well, should be approached for support in technological development for those language communities who are not as advanced in corpus work, and/or who find technological imports expensive. The support budgets should include substantial provision for training. The many languages of central and eastern Europe are prime candidates for this kind of support, which should not be delayed until the users become members of the EC, but which we recommend should be built into "know-how" packages as soon as possible.

3 General Principles for the Organization of the Work

1. The organisation of the work should be entrusted to a consortium of institutions monitored by EC, and guaranteed by an advisory group of leading experts. The maintenance, development and upgrading of the corpora and lexicons would be an activity co-ordinated at a European level, to which each participating institution would have to conform.

2. The participating institutions are called LRs (Linguistic Resources) Centres in this Report. Each one should have experience in large corpus maintenance and in corpus based computational lexicon design and construction.

3. It is desirable that each LRs Centre should support major research in corpus linguistics, so that the impetus to keep developing the corpus resource is part of the institutional framework, and does not become the responsibility of the corpus sponsors.

In the wrong hands, a LRs Centre might become a block to progress, seeing itself as an archive rather than an ongoing resource. The whole experience of the NERC project tells us that any specific goals must be provisional, and that they need to be revised twice a year. The broad general resolve to maintain state-of-the-art LRs provision must form a costant commitment, and the levels of resource needed to keep the work going do not need to change frequently. The experience of the last fifteen or twenty years in corpus linguistics has been that the cost of any job or piece of equipment keeps falling sharply, while activities which were impossible come in the range of expensive but now possible. In these circumstances it is reasonable to expect level funding to provide a steadily improving sevice. The charges that users pay will also provide level funding for that aspect of the provision that will be financed from their contributions.

4. A framework of research activity is also important for the definition from time to time of what constitutes a basic access and retrieval package. It is pointed out in Chapter 4 that users will expect an ever more sophisticated service.

Those who provide LRs from public funds require to distinguish between services which are part of the general infrastructure, such as standardising and co-ordinating services, and activities which are properly part of a research effort or a commercially viable application, and which should find funding from other sources.

5. The guidance of a panel of experts is also important, and the framework of EAGLES may provide the right sort of advice and guarantee.

The need for LRs Centres may not be permanent, and after some time it may be necessary to cut their size and scope or merge them with a related activity. If they had become in the meantime powerful and autonomous institutions in a community there might be a need for strong voices of recommendation to avoid wastage. For example, it is not at all beyond possibility that students of

language in a few years time may be able to tap in to a range of electronic texts directly, and the corpus will indeed become more of an archive and less of a day-to-day resource. On the other hand it may be felt necessary after some experience to establish one or more reference points of European relevance to deal with industrial, political and academic corpus needs. No-one knows yet what the value of corpus expertise will be in general matters of content retrieval, document searching and abstracting, classification and archiving. This Report draws attention to some promising developments.

4 NERC Implementation Plan: Organizational and Managerial Aspects

4.1 *Introduction*

The current state of the NERC Consortium is of eleven institutions, each representing a member country of the EC. Six of the institutions have worked together over a period of years, to establish the viability of the network and to provide a depth of experience and documentation. The newer members have made a commitment to this work, and made a preliminary statement about their participation and their view of the NERC proposals.

NERC has been in touch with two interest groups outside the EC - those in the USA, in particular the ACL/DCI and the ARPA LDC, - and corpus linguists elsewhere in Europe. Colleagues in Sweden (Prof. Sture Allen) have already asked to join, and contacts are building up in central and eastern Europe (see Appendix 1), following the previous activities of some NERC members in the framework of the Council of Europe Expert Group on Corpora.

4.2 *Management Structure*

At present NERC is organized as is the praxis for shared cost EC sponsored project consortia. Those responsible for the various partners' Institutes meet periodically to discuss various consortium issues: Technical reports, sharing of work, etc.

It is understood by all that during the next stages of the work a more formal coordination structure will be necessary. Also, in preparation for roles of public responsibility, the partners will build up management structures and representative boards so that national, academic, and industrial interests are involved in the development of LRs Centres. During the second stage it is expected that some experience of income generation will be gained, and consortium plans and policies will be worked out on the times indicated below. This follows the identification of potential users and involvement of them in the planning of the third stage.

By the end of the second stage it is planned that all the NERC members will have achieved the right structure, resources and assurances of continuity to become national Corpus Centres.

It is proposed that the NERC 1 and NERC 2 Consortia will form a unified consortium for linguistic resources (LRs) in the next stage, although there could be a temporary distinction in their roles. As an example, it has been suggested that the NERC 1 partners could, from the beginning, act as full members, while the NERC 2 partners or, at least, those which still have to acquire some of the fundamental features required as part of the minimal profile recommended in this report (see section 4.5 below) could act as associated members, preparing the Consortium for the next stage, with the addition of (at least) Sweden (University of Gothenburg) as an Associate Member. Since the parallel corpus material that is to be the basis of Stage 1 is not available in Swedish, Gothenburg cannot be a full partner until Stage 2. The same is true of any other country, e.g. Finland, Hungary, Poland, which would be encouraged to join[6].

At the end of the first stage there will be an independent evaluation of the Associated Members in terms of their speed of development, ability to mount Stage 2 and ultimate suitability for the European Consortium for LRs.

Furthermore, each member of the Consortium, in his home country, will act as a reference point for relationships with other relevant institutions. In fact, a goal of the proposed consortium for the LRs is to involve, at the various stages in the process of LRs creation, the relevant types of actors in the different European countries, each one according to his interest and potential role, taking into consideration also the variety of resources to be created (corpora, lexica of different types, etc.).

4.2.1 *Management of the Consortium*

A) STAGE 1 and 2

It is suggested, for Stages 1 and 2 of the Implementation plan, a two level management approach.

Level 1: Steering Committee

(a) Membership

- the directors of the NERC1 Centres and representatives of the other EC countries (in principle, the NERC2 Centre Directors),

- the EAGLES coordinator,

Selected representatives of the industries, publishing houses, potential users will be invited to assist the meetings.

The SC will be chaired by a President (distinct from the coordinating partner who will be contractually responsible with respect to the CEC) assisted by:

6 As examples of Institutions which could he contacted in other countries to become associated partners, we could indicate, for example, those listed in Appendix 1.

• two vice-presidents and one secretary.

The President, the vice-presidents and the secretary will be elected by the SC members.

A full-time project manager, will be selected and appointed by the SC, and paid for out of the project funds. He will be present at the SC meeting, but will not be an SC member.

(b) Responsibilities and Duties

The responsibilities and duties of the SC will include the normal functions of the project technical committee and, in addition, formulation of the specifications for the next stages.

Level 2. Working Groups

The Steering Committee will set up ad hoc temporary working groups, to deal with specific technical scientific tasks, composed of NERC participants, and possibly selected representatives of the R and D communities (industries, research institutes, publishing houses, etc.).

Each working group will be convened and chaired by a NERC centre member, and hosted by a NERC centre.

B) STAGE 3

A three level approach is suggested.

Level 1. Advisory Board

(a) Membership,

(i) representative from national government or national funding authorities,

(ii) representative from local community,

(iii) representative from national corpus community: industries, publishing houses, research institutions,

(iv) representative from user community: industries, publishing houses, research institutions,

(v) senior administrator from host institution,

(vi) representative from EC,

(vii) President and vice presidents of the SC,

(b) In Attendance,

(i) Corpus Centre Directors,

(ii) Department heads,

(iii) Project manager,

(c) Responsibilities and Duties,

(i) to ensure external communications and support for the Centre,

(ii) to harmonise Centre business with representational affairs,

(iii) to recommend developments,

(iv) to monitor quality of services and efficiency of Consortium Coordination.

Level 2

SC as above

Level 3

WGs as above

4.3 *Services and Income Generation*

- Everything that has a function at the European level only (e.g. the common query language in a network) should be paid by the EC on a 100% basis.

- Everything that has relevance to both the EC and the national corpus institutes (e.g. transcribed spoken text corpora, corrected tagged corpora) could be paid by the EC and national funding on a basis of 50/50.

- Services to commercial purposes should be paid by the company. Services to research purposes should either be paid by the particular research institute or university, or, if this is not feasible (and often it is not, in our experience), by local or national funding authorities.

- Creation of ad-hoc resources on order, belongs to the responsibility of each particular institute.

(N.B. some partners have experience already of these activities).

Stages 1 and 2 will be guaranteed by EC funding, since the primary aim is to develop a European facility, standardized throughout the community.

In Stage 1 a start will be made on selling services, and during Stage 2 a substantial level of self-financing will be achieved.

The following areas of activity will generate income:

(a) network access. Charging either by time used or by subscription. Users control process by remote log in.

(b) Consultancy. Customers engage specialist services for non-standard requests, general advice and support.

(c) Study units. Facilities (workspace, terminal access, advisory support) provided for users who wish extended access to full facilities.

(d) Products. Software, data descriptions, analyses, reports, research publications. On sale by mail order through electronic media. Development, on request, of specialized corpora, possibly structured and linguistically analysed, and of corpus based tools and products.

(e) Sales. Data (where appropriate) and products.

During the first stage a system of royalties should be designed, experimented and established, to regulate the distribution of income generated by 'selling' products and services.

General principles and specific rules should be agreed to regulate, for example, if a part of the income should go to support common tasks and structures, or the case in which a product of centre A is sold through the activities of centre B.

Particular cases will be represented, for example, by multilingual corpora, multilingual lexica, centrally generated software, whose production will involve more than one member.

4.4 *Profile of Candidate Nodes*

The list below describes the features an "ideal" node should possess to become a member of the Consortium.

This list can serve:

- as a checklist of features to be used in evaluating the level of adequacy of a candidate node,

- as a "target" that each node should try to achieve during Stage 1 and 2.

Clearly preference should be given to candidates which possess the greater number of features.

It must also be borne in mind that one of the functions of the Consortium should be to promote the know-how and technology transfer among the nodes and that some functions, not specifically linked to an individual language, could be performed by the Consortium as a whole, or by some competent Nodes (principle of complementarity).

(Consider, for example, the expertise of Leiden in providing services to publishing houses, of Birmingham in creating filters for the monitor corpus, of Pisa in standards design, language processing, users analysis needs and services, lexical DB structures and knowledge extraction, etc.).

In other words, the Consortium, if well constructed, could offer more than the sum of its parts.

4.4.1 *Guaranty of Autonomous long-term Stability and Continuity* are the essential preconditions to ensure the continuity requested for the maintenance and the regular updating of Linguistic Resources (LRs), the progressive enrichment of the data to cope with the advancement of the state-of-the-art, the feasibility of the reference and monitor corpus approach.

The stability and continuity could be evaluated on the basis of the longevity and past history of a node, its institutional nature (e.g. permanent institutes vs. temporary projects or SME industries), the inclusion of the creation of basic resources and tools for LP in its institutional mandate.

4.4.2 *Permanent Public Funding is necessary in order to:*

- ensure the above mentioned stability and continuity,
- promote the creation of LRs to be made immediately available in the public domain,
- provide the national contribution to the European consortium,
- reinforce the adoption of common standards, and the international multilingual coordination,
- allow "openness" in respect to various categories of users.

This aspect could be evaluated considering the affiliation of the node, the explicit correlation of the funding with the tasks of providing basic resources and tools, the autonomy of the direction of the node in allocating funds or structured resources (manpower, hardware etc.).

4.4.3 *Know-how, Practical Experience, Broad Range R&D Activities in LP*

That profound basic differences exist between collecting data, even on a computational support, for the (traditional) lexicographic or language description work, and creating LRs for Language Processing (LP), is obvious.

These differences have clear consequences on the desired profile, in terms of experience, know-how, background multi-disciplinary formation, working habits, scientific methods, technical background.

Deep knowledge and understanding of contemporary linguistic theories, long-term familiarity

with LP problems and methods, substantial experience in producing and using a variety of LP systems and components in both R and D environments are very important prerequisites for several aspects of the creation and provision of LRs to LE: identifying relevant linguistic facts and formalising their properties; adopting and updating representative standards; being aware of new developments in CL and of new requirements in LE; updating LRs to include the results of the advancement of the state of the art; providing finalized high-level services to answer specific requests by different users; cooperating with prospective users in analysing their needs, identifying and evaluating available LRs or designing new methods and LR types, etc.

Furthermore, LP know-how and experience are important to create computational tools and use LP methods for collecting, constructing, converting, maintaining, updating, accessing, encoding, analysing, distributing, evaluating, extracting linguistic knowledge, for LRs.

This aspect could be evaluated considering:

publications in the field of LP; production of tools, components, systems for LP; participation in specialised conferences (e.g. COLING, ACL) and Associations; cooperation with industries, in particular in relevant strategic national and international LP projects; technology transfer and teaching activities; etc.

4.4.4 *Internal Structure of the Node: Organisation, Functions*

The establishment, development and running of a European network for the creation of LRs require nodes whose internal structure (offices, units, etc.) could perform the following functions.

- Acquire linguistic data, in Machine Readable form, from available sources

Identify possible sources; establish contacts; select relevant data; determine feasibility and cost of extraction and conversion of interesting information; maintain an inventory of the data available; deal with copyright problems and juridical aspects.

- Conversion of data to the standard format

Analyse the source coding system; adapt (if necessary) the conversion software to individual source encoding scheme; perform the conversion into the common standard encoding format, applying analytical tools where necessary; validate with manual checking the conversion results.

- Acquire data not available in machine readable form:

- Keyboarding; OCR.

- Recording spoken language data.

- Transcribing spoken language.

- Defining lexical specifications and creating a lexical resource according to state-of-the-art agreed standards;

- Corpus design.

Define the corpus composition;
Select the individual texts to be included in the corpus; decide the modification of the reference corpus during the monitor phase.

- Maintain the catalogue of data in MRF.

- Software design and implementation: batch and on-line data access, editors, etc.

- Relationships with users: documentation; user needs analysis; consultation; data delivery; data extraction; etc.

- Linguistic analysis of texts (annotation, tagging, parsing, etc.).

- Statistical Analysis.

- Standards and formalisms

- Distribution via different types of access and media (network facilities, CD-ROM, etc.).

A prototypical organization chart showing the essential units is appended.

Ideally, a node should already include units or personnel allocated to the tasks listed above, which already possess (and could demonstrate) the relevant know-how and expertise. This personnel should serve as a reference point for eventually newly recruited additional manpower.

This aspect could be evaluated examining the organisation of the institutions, publications and concrete results testifying past work, participation in relevant national and international projects.

If a node does not yet present all the broad range of competences listed above it is important to evaluate the flexibility of the present structure to accommodate new units, the possibility of hiring new competent personnel, or of distributing part of the work to external
contractors or to associate nodes. The cost of this integration and possible funding sources must be evaluated.

4.4.5 *Hardware*

Hardware facilities should be available for the following functions:
(OBL = obligatory; OPT = optional during the first phase).

- Mass Storage (OBL) not less than 10 GIGABYTE

To maintain an adequate quantity of data on-line, for periodical processing, monitoring, filtering, updating, for interactive access, for extraction of subcorpora, for statistical analysis; etc.

- HIGH SPEED dedicated connection to an international network (OBL) to ensure both on-line access to the data on the mass storage, and data distribution (FTP, AFS, etc.)

- CD-ROM PRESSING FACILITIES: (OBL)

for distribution of large quantity of data on demand.

- ADEQUATE COMPUTING POWER: (OBL)

for data processing, available preferably in-house; if a client-server architecture is adopted for each node, the node can reply both to communication requests and storage requirements very efficiently.

- AUTOMATIC RECOVERY FACILITIES: (OBL)

to ensure fast recovery in case of problems in the data storage.

- WORKSTATIONS (OBL)

An adequate number of workstations, linked in an internal network for internal use in: interactive text processing, manually assisted linguistic annotation,data capture and validation, etc.

External visitors for data access.

- SPECIAL PRINTING DEVICES (OPT)

Capable of both mass and high-quality printing, and of multilingual multi- printing.

- HIGH QUALITY HIGH SPEED OCR FACILITIES (OPT)

- PARALLEL COMPUTING (OPT)

For language knowledge extraction and learning procedures.

- IMAGE PROCESSING FACILITIES (OPT)

In the near future, large quantities of digitalized texts and text-images will be made available in libraries. Cooperation with libraries could be a major source of data and services opportunities, as testified by initial cooperation between corpus methodology and
advanced library projects.

The evaluation of this aspect could be easily made requesting a candidate node for the list of available hardware and related specifications, including availability for LRs storage and processing.

4.4.6 *Know-how and Expertise in Standard Design and Application*

Harmonisation is essential to ensure multilingual coordination, and standardization is needed to ensure reusability.

In the current situation, emerging standards (NERC, EAGLES, TEI, etc.) are far from an exhaustive coverage. On the contrary, they will need regular updating and enrichment to cover additional phenomena or specific needs.

The nodes of the network, which should act as reference points in the respective countries, and take a major role in the creation of LRs, should actively operate and cooperate in the standard design and establishment activities. This requires specific know-how and experience in standard design.

This aspect could be evaluated considering:

- participation in recent international standard activities (EAGLES, NERC, TEI, etc.);

- establishment (prior to this recent initiative) of national or regional standards.

4.4.7 *Linguistic Data in Machine Readable Form*

The availability of (possible) large corpora, both in the national and in the other languages, is important for several reasons:

to implement the first stage of the NERC proposal, access through network to test software, methods, procedures.

The status of the data (availability to different types of users) is relevant for:

- determining the reusability of existing data, to contribute to the construction of the initial reference corpora (in the second stage of the NERC proposal), and the initial basic lexicons,

- the possibility of giving access to external users,

- the possibility of distributing at least part of the data,

- evaluating the expertise in corpus linguistics and computational lexicology of the candidate node.

The availability of lexical data, in particular lexica for LP, is important as a preliminary tool in corpus processing, and for not starting from scratch in the construction of the lexical resource.

Existing data could be classified, for this purpose, as follows:

- balanced corpora
- specialised (sublanguage or variety corpora)
- spoken data
- corpora in other languages (processable with the same software)
- multilingual parallel corpora
- tagged corpora
- aligned multilingual corpora
- other collections of texts in MRF (machine readable form)
- general lexicon for the national language (linguistic information at different levels)
- general lexicon for other languages
- conceptual taxonomies
- specialized lexica
- structured Machine Readable Dictionaries
- lexical knowledge bases.
- etc.

This aspect could be evaluated asking for:

a list of all the textual and lexical data available and their description (e.g. number of words; encoding system: proprietary, TEI conformant; levels of linguistic description; availability for different types of users; documentation showing that data have really been distributed).

4.4.8 *Specialized Software and Lingware*

Software and lingware tools for corpus and computational lexicon work should already have been produced and in current use.

A checklist for evaluation could be:

conversion from other sources to the internal format
conversion to the TEI conformant format
traditional text processing packages (concordance, frequency, etc.)
interactive access (on-line contextualisation, etc.) in-house
interactive access through INTERNET
statistical tools
lemmatizer

computational lexicon for the national language		for other languages
computational morphology	"	"
rule-based tagger	"	"
statistical tagger	"	"
parser (syntactical)	"	"

conceptual/semantic analyser
multilingual aligner (words, sentences, paragraphs)
knowledge extraction (e.g. collocation, subcategorization)
automatic learning
editor for lexical informatiom
lexical database browser
Machine Readable Dictionary browser
definitions analyser/parser
Lexical Knowledge/DataBase formalism and management
Typed-Feature Structures formalism tools
others

The evaluation of the aspect will include:

list of software available; documentation of use of the software by other institutes/industries; list of other sites which have requested and received the software; patented software; participation in national and international projects for the creation of software/lingware.

4.4.9 *Documentation of Long-Term Corpus and Computational Lexicon Work*

Long-term engagement in corpus and computational lexicon activities could be an important witness to the:

- continuity and stability,
- development of relevant skills, tools and expertise,
- scientific interest in corpus linguistics.

While the quality of the work should be evaluated considering the data, the tools, the innovative methods produced, the number and types of users and services, the integration of corpus and lexicon

work in internal R and D activities, the "longevity" could be documented through publications, teaching activities, corpora, lexica, and tools produced in the past.

4.4.10 *Services to R and D Users*

As specified in the NERC proposal, the ability to provide services, in different forms, to users of different types is the central goal of the proposed network.

The Services could consist of several functions:

- To give access to the data (interactive queries)
  - in-house
  - through networking

- To distribute data outside the node:
  - on CD-ROM
  - through networking (FTP,AFS, etc.)

- To produce LRs on demand

- To design LRs for external users on the basis of needs analysis

- To distribute software for external use (both at national and international level)

Relevant facilities for providing services are:

- "Bureaucratical" capability of accounting and billing for services

- Hardware and software for distributing data (FTP, CD-ROM, etc.)

- Manpower explicitly dedicated to service activity

- Manpower dedicated to user needs analysis, specialized LR design, etc.

4.4.11 *Long-Term Experience in International Cooperation*

To take part in the proposed network for the creation of LRs, it is important that a node can document successful experience in international cooperation, to testify the organisational, scientific, technical, structural capability to take part in cooperative, centrally coordinated actions.

The following categories of international projects are particularly relevant for the field of LRs:

corpora

lexica

grammars

software tools.

4.4.12 *Cooperation with other Institutes of the same Country*

The capability of coordinating activities in the field of LRs, and proved experience and actual links inside the country, will be important features:

- to maximise synergies,
- to avoid duplication,
- to ensure sensibility in the production of LRs,
- to ensure relationships with potential users,
- to set up and manage a national subnetwork,
- to represent an 'entry' point in the country for the European Consortium for LRs.

4.4.13 *International Recognition*

For the success of the network, it is important that the various nodes enjoy recognition and prestige at the international level.

International recognition is - in general - immdediately evident for the experts of the field. It can be analysed considering:

- number of international projects coordinated,

- participation in international projects,

- role in international scientific Associations,

- contacts with foreign industries,

- participation in international Committees,

- invited papers.

4.5 *Approach to the Consortium for LRs Construction*

A gradual approach is suggested to constitute the final membership.

Initially two categories of nodes will be recognised.

Full members:

All those NERC1 and NERC2 members who already possess all the 'ideal' features, or at least satisfy the following minimal set of requirements:

- Members of permanent public institution with strong academic reputation, and permanent public funding,
- Long-term commitment of institution,
- Local coordination by senior official,
- Full facilities for corpus and computational lexicon building, maintenance and processing,
- Trained dedicated staff of at least two corpus linguists, two computational lexicologists and two computer specialists, provision being made with regard to the size of the country,
- Full communication facilities including email and internal connections,
- Secretariat, clerical and accounting infrastructure,
- Commitment to infrastructure development,
- National and international reputation for corpus and lexicon expertise,
- Publications in the field in active preparation,
- Good contacts with other corpus and lexica interest groups,
- Good contacts with potential national users.
- Corpus data and computational lexicon to share

Associated members:

All the NERC1 and NERC2 members who do not yet meet all the minimal requirements. They sit in the SC, in order to prepare their integration as a permanent centre, and give them the opportunity to promote their interest. They will take part in the Working Groups set up by the SC.

4.6 *Common Infrastructure*

As mentioned above, the target for the end of Stage 2 is to put in place the following infrastructural facilities:

(a) Publicity and documentation

(b) Extensive mailing list

(c) Access by electronic network, email, fax, phone, ordinary mail. Efficient response.

(d) Standard corpus built to EAGLES specifications, efficiently maintained.

(e) Standard query language in operation on corpus and lexicon, available to users.

(f) -500 protocols for maintaining network.

(g) Efficient liaison with other members of consortium and Coordinator.

(h) Efficient liaison with other national corpus interest groups.

(i) Advisory Board in place, structured according to guidelines.

(j) Initial nucleus of a monolingual lexicon built to EAGLES specifications (60,000 entries: see Appendix 2).

(k) Ability to provide services during specified working hours and on specified days, to be of maximum availability in the Community and world-wide.

5 Proposals: Computational Lexicons[7]

As far as Computational Lexicons are concerned, they are a type of Linguistic Resource which is strictly interrelated with large Text Corpora, in particular in the framework, nowadays largely shared both in the linguistic and lexicographic communities, of a corpus-based lexicon development.

Very large text corpora are on the one side an essential basis for developing mono- and multilingual Lexicons, and Computational Lexicons are on the other side an essential tool for tasks such as tagging, lemmatizing, parsing, finding collocations, and other corpus analyses.

Any Corpus Centre has to be able to provide the lexicological and linguistic expertise necessary

[7] At the request of the EC, the NERC Consortium has prepared a proposal in the field of Computational Lexicons for a meeting organized by the EC in preparation of the 4th Framework Programme. This proposal, which was approved by the participants to this meeting, and by the NERC evaluators, is enclosed as Appendix 2. We include here some preliminary recommendations.

for designing and building Computational Lexicons, and to offer Lexicons of different types already developed in-house.

The Centres of the NERC Consortium have this type of competence and can take part in the construction of Computational Lexicons in various forms:

a) participating to, or promoting, projects for the development of Computational Lexicons. Concerning this point, it has to be remarked that in any case it is necessary to build Lexicons for different types of Corpus analyses, and these already developed corpus-based lexicons can be made available externally, at a cost, by the NERC Centres.

b) offering lexicological and linguistic know-how and expertise, as an external service for the building of specialized or general lexicons on the basis of text corpora.

c) offering the use of various, possibly standardized, lexicogical and linguistic tools to external users, for corpus based lexical development.

For what concerns a proper proposal for Computational Lexicons development, we refer to Appendix 2.

6 Conclusion

In earlier times a language demonstrated its autonomy, permanence and maturity by developing a writing system. Then, with increasing sophistication, each language had to have a large dictionary to keep its status, and some of these dictionaries are held in reverential awe by the communities to which they belong. Now, in the electronic age, a different kind of record of a language has become possible and is becoming prestigious. It is more flexible than a reference book because it is a reference collection of the raw material from which reference books are made, and has many uses. As an archive it can be consulted for reinterpretations and historical researches in the future.

In the first instance we recommend that the EC pioneers a network of LRs in the official languages, and provides the necessary funding. Following the inaguration of this (and not by any means waiting till the end), the leadership of the EC and perhaps the Council of Europe will be helpful in alerting private and regional funding bodies to the importance of LRs, and the movement to establish LRs of all the indigenous languages of Europe will be an important political signal of the move towards decentralisation. All communities with an interest in their cultural heritage should be prepared to invest in a substantial LR when they know what it is and how it can be used.

References

Calzolari N., Bindi R. (1990): "Acquisition of lexical information from a large textual Italian Corpus", in *Coling '90*, Helsinki, vol. III, pp.54-59.

Clear J. (1988): "Trawling the Language: Monitor Corpora", *ZüriLEX '86 Proceedings*, M. Snell-Hornby (ed.), 383-389, Tübingen, Francke.

EC DGXIII (1992): "Language and Technology", EC Report, Luxembourg.

Sinclair J. (1991): *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

Walker D., Zampolli A., Calzolari N. (eds.): "On automating the lexicon", OUP, Oxford, 1994.

Zampolli A.: Foreword, in Atkins S., Levin B., Zampolli A. (eds.) "Computational lexicology and lexicography", (in print), OUP

Zampolli A. (1993): "Linguistic Reources for R & D Communities: Problems and Perspectives", in KB & KS '93 Proceeding, Tokyo.

Relevant NERC Papers

NERC Consortium (1992): "Policy for Corpus Provision for Europe", Strategic briefing paper, NERC-99.

APPENDIX 1

AlbaniaSektori i Enciklopedisë Shgiptare, Akademis e Shkencave, Tirana (Xhewat Lloshi)

BulgariaInstitute for Informatics, Bulgarian Academy of Sciences, Linguistic Modelling Laboratory, Acad. Bonchev Street, bl. 25a BG-1113 Sofia (Georgi K. Gargov)

CroatiaInstitute of Linguistics, Faculty of Philosophy, University of Zagreb, Salajeva 3, 41000 Zagreb (Maja Bratanic)

CzechoslovakiaAlgebraic Linguistics, Cathedra Numerical Mathematics, Mathematical Physical Faculty Charles University, Malostranske 25, 11800 Praga (Eva Haijcova)
Filozoficka Faculta UK, Nam. J.Palacha 2, 116 38 Praha 1 (Františck _ermák)

EstoniaDepartment of General Linguistics, Tartu University, Tiigi 78, EE 2400 Tartu (Heiki-Jaan Kaalep)

FinlandDept. of Computers Science, University of Helsinki (Fred Karlson)

HungaryHungarian Academy of Sciences, Szentháramság v.2, H-1024 Budapest (F.Kiefer)

LatviaInstitute of Mathematics & Computer Science, University of Latvia, Artificial Intelligence Laboratory, 29 Raina Blvd. (Andrejs Specktors)

LithuaniaDepartment of Lithuanian Language, Universitas Vytauti Magni, S.Daukanto 28, 233000 Kaunas (Ruta Marcinkeviciene)

NorwayDept. of English, University of Oslo (Stig Johansson)

PolandInst. Informatyki uw, Pkin p.838, Skr. Polzt. 1210, 00-401 Warszawa (Janusz St.Bien)

RomaniaLaboratory on Natural Language Processing, Research Institute for Informatics, Averescu Blvd. 8-10, Sect.1, 71316 Bucharest (Dan Tufis)

RussiaInst. Problemy Pereda_i, Informacii an SSSR, UL, Ermolasvoj 19, Moskva (Ju. D.Apresjan)
Institute of Russian Language, Russian Academy of Science, 123480 Moscow (W.M. Andrjuscenko)

SerbiaComputer Laboratory, Faculty of Science and Mathematics, University of Nelgrade, Studentski Trg 16, 11000 Belgrade (Duško Vitas)

SlovakiaJazydovedny Ustav Ludovmta Stzra, Slovenska Akadimia Vied, Panska 26, CS-813 64

Bratislava (Jan Dorula)

SloveniaInst. J.Stefan, Univerza, p.p. 199- IV, Jamova 39, 61001 Ljustana (P. Tancig)

UkraineUniversity of Kiev (Galine Chekal)

As far as non-official EC languages are concerned, we could signal, as an example, the Catalan Corpus directed by Prof. Raphael.

APPENDIX 2

Computational Lexicons

The action of building a basic set of European Lexicons and related Lexware has to be considered as an infrastructure development and applied R&D action.

1 Justification

The experience gained in the European projects, starting from ET-7, going on with Acquilex (I and II), Multilex, Genelex, ET-10, LRE and EAGLES, have created the necessary competence to allow launching a true development project in the lexical area, with industrial planning, and very well defined common guidelines for all the phases of lexicon building.

The availability of large, reusable lexica is the major bottleneck for real-life practical NLP system developments. The construction of an adequate lexicon is too expensive for an individual company to build, and duplications must be avoided. Reusability and incrementability must be guaranteed.

The proposal is extremely strategic as it is basic to all other service areas. After an initial period, results may be available to other service areas ("pilot applications" can be envisaged as testbeds after the initial phase). Results will also be disseminated. Regular interactions with the Corpus activities are envisaged, leading e.g. to the establishment of common descriptive categories for tagging, based on the acceptance of the EAGLES proposals.

Initially, an investigation should be made as to which lexica are available within each country, and, where possible, make all the necessary steps to reuse them.

In fact, after the ET-7 Study, it became very apparent that, even though a number of lexicons exist, not one of them is completely reusable. Sometimes parts of them are reusable, these parts however being very different from each other, having been built within completely different theoretical and/or applicative frameworks. An evaluation must be made of the economic effort needed to reuse existing data as they are, keeping in mind the quality of the results of this exploitation, when deciding on what can be successfully reused.

The state-of-the-art is such as to ensure the feasibility of building a large basic, truly reusable, high quality computational lexicon. The cost of populating it will be much lower than in the past, also due to the recent availability of more lexical tools.

The lexicons for all the EC languages resulting from the project will have the following characteristics (making them different from anything already existing and completely new):

- they will share a common descriptive methodology, a common architecture and common representational devices

- they will be freely reusable (criteria to ensure reusability must be specified) and have no links to

any specific application or product.

This includes the adoption of common specifications - based on EAGLES -, classification of the status of the resources (availability in the public domain), modalities for sharing costs among partners (CEC, national governments, companies, users, etc.).

Resources built within the above mentioned Lexical Projects, plus some national resources, will constitute the basis from which to build on in the action designed in the present proposal.

The bulk of the work will be, however, dedicated to the acquisition from corpora, definition and representation of syntactic, syntactico-semantic, and semantic properties, together with "collocational" information.

2 Objective and scope

The action has the objective of creating a network of large basic Lexicons for all major standard European languages, plus the relevant lexical tools. Initially only monolingual lexica will be considered.

Deliverables are grouped in three phases and/or types:

Implementation: Stage One

A starter kit comprising a minimum set of lexical items per language (about 3,000) plus relevant tools for exploitation, without restrictions of distribution.

Implementation: Stage Two

Public domain access and/or full access (subject to resolution of copyright issues) to a larger lexicon (60,000 to 80,000 per language).

Implementation: Stage Three

Extensions towards exhaustive lexicons (privately owned by the producers).

Specifications common to the Three Implementation Stages

Quantity of Information for each Lexical Entry (LE):

- Phonological level: complete (with basic phonetic information)
- Orthographical level: complete
- Morphological level: complete (including POS, simple segmentation for derivation and compounding)

- Syntactical level: complete information for subcategorization

- Syntactico/Semantic level: some information on arguments and "roles"

- Semantic level: at least a set of very basic, commonly agreed features

- Collocational level: very large data of different types, derived semi-automatically from corpora.

Terminology issues and bridges to bilingual lexica will be considered as future extensions when determining specifications.

Type of LE: simple and compound words

Methodology of work: corpus based information, with a common methodology of acquisition and a common descriptive framework (building on results of on-going projects in the lexical area).

Coding system:

- according to EAGLES specifications (available at a prefinal stage for the lexicon architecture and methodology and the morphosyntactic level of description, and being developed for the syntactic level and other levels)

- common Guidelines for acquisition, testing, representation

Development of a set of basic lexical software tools: for acquisiton, data entry, maintenance, conversion, editing, browsing and retrieval, import and export (partially
similar to the Corpora tools).

In addition, consideration will need to be given to the setting-up of an appropriate network for storing, maintaining and distributing lexica to interested parties.

3 Tasks, Resources and Costs, Duration (for Stages one and two)

A. Linguistic data Tasks

A1 - Inventory of actually available "reusable" resources
A2- Definition of the lexical entry (LE) structure and of the lexicon architecture
A3- Compilation of Guidelines for LE building in a corpus based methodology
A4 - Population of the Lexicons for the different languages.

B. Software Tools Tasks

B1 - Inventory of available "reusable" tools
B2 - Definition of the minimal/optimal set of Lexical Tools
B3 - Compilation of specifications and functionalities of the tools
B4 - Design and development of tools.

Resources required

Stage One

A1, A2, A3 3 m/y
B1, B2, B3 7 m/y

Cost 600,000 ECU

Stage Two

Resources required for one language

Task Manpower

A2 1.5 m/y
A4 20 m/y
Coord. 1 m/y

Costs for 1 language 1,290,000 ECU

Total Costs for 9 languages 12,210,000 ECU

***Duration**: 4 year project*

Implementation: Stage One

1st year:

A) Linguistic data

- inventory, evaluation, and decision on what to reuse and how to reuse it
- setting up of a common theoretical, conceptual and methodological framework

B) Software tools

- producing specifications

Implementation: Stage Two

A) Linguistic data

2nd year:

- construction of a prototype which should constitute the starter kit to be made freely available on the market
- refinement of methodology (using feedback from development of the prototype)

3rd and 4th years:

- population of large lexicons

B) Software tools

2nd to 4th year:

- prototyping and developing tools

Chapter 1

User Needs

1 Introduction

1.1 *Meeting user needs*

The main task of the Work Package *User Needs*, as viewed by the members of the NERC consortium, is to explore the variety of uses, applications and purposes of linguistically defined corpora and of corpus technology in the NLP research and development community. To this community belong the academic, not-for-profit researchers, the swiftly growing market of smaller and medium-sized software developers, multinational corporations that can afford to build up their own corpus-related resources, and finally the growing number of commercial (and academic) enterprises marketing machine- readable or electronic texts. Anyone who wishes to distribute texts on CD-ROMs, or to deposit electronic texts in libraries, or to offer access to full text data banks by networks - all these concerns are making increasing demands on corpus technology.

Within the framework of this feasibility study, we have concentrated on the user needs of the NLP research and development community. Here, European developers not only have to fight against the much more homogeneous markets of their North American and Japanese competitors, but are also put at a disadvantage by the European situation, where today nine languages enjoy equal rights. The preservation of the linguistic and cultural identity of the member nations is a high-priority goal, but it certainly increases the cost of natural language processing. This is why we consider it necessary for the European Community to support the creation of precompetitive corpus resources in terms of linguistic data, software and expertise. This will further the business interests of the NLP community and enable it to deal successfully with the particular European situation.

1.2 *The identification of user needs*

In order to identify user needs the following steps were taken:

-One session of the Pisa Workshop, January 1992 (NERC Consortium, 1992, NERC-82) was devoted to user needs.
-The literature on computational linguistics and related subjects (from 1985 until today) was evaluated with respect to the whole range of applications for corpora and corpus linguistics.
-A survey was carried out of the corpus-related services rendered by members of the NERC consortium to individuals and to academic and commercial institutions; this was complemented by a questionnaire-based enquiry as to emergent user needs.
-A comprehensive, world-wide survey on textual data was carried out with the support of U.S. partner institutions.
-A series of in-depth interviews was held with corpus providers, corpus users, and experts with a variety of backgrounds; the purpose of this was to obtain user profiles.

Details are given below.

1.3 *General observations*

In our attempts to identify user needs, we were not infrequently confronted with a lack of awareness on the part of developers of NLP software as to the importance of corpora and corpus-related resources for the quality of the products they envisage. Sometimes it seemed that the last 30 years of failures, broken promises and slow advances have in no way affected the prevailing optimistic belief that it will be possible to come up with powerful and robust NLP systems in the very near future.

Many computer linguists, particularly those with an engineering background, take it for granted that, with the expertise and tools at our disposal today, a steady improvement in NLP systems is possible to the point where, say, an operational, robust system for machine translation can be developed. They are impressed by the fact that seemingly astonishing results can be achieved with

stochastic methods necessitating rather minimal knowledge about the language involved. They believe themselves to be justified in their convictions because they look at natural languages as nothing but a set of particularly complex formal languages. But there is a generic difference. Given basic conditions, formal languages can be translated into each other. But anyone who has translated from one natural language to another knows that it takes more than just the pure linguistic knowledge involved. The translator must also understand what the text to be translated is about, and must be aware that (s)he will understand it only if (s)he has sufficient knowledge of the world. Insofar as such knowledge is a precondition for NLP systems, it will have to be integrated into our linguistic knowledge.

The linguistic knowledge available - the knowledge formulated in existing grammars and dictionaries - has to be used if the performance of NLP systems is to be improved. But even when this is understood, there remain constraints which limit the level of perfection that can be achieved. Traditional grammars and dictionaries have been devised for human users, and human users differ substantially from machines. Humans use inductive reasoning and can draw analogies easily: faculties like these are taken for granted and are reflected in existing presentations of processed linguistic data. In order to make our linguistic knowledge available to NLP systems, it has to be reorganized and rewritten entirely in terms of deductive reasoning and algebraic logic. This task is sufficiently demanding in itself. In order to carry it out we have to go back to the sources, and the source for raw linguistic data is the real and actual text in its un-annotated representation. But anyone who has gone to the sources has also experienced the problem that when we start analyzing linguistic material from scratch - when we analyze language as it occurs in a corpus - we come up with new insights. We discover that the grammars and dictionaries we have been accustomed to using give us a very biased view of language, a view that has its roots in over two thousand years of continuous linguistic theorizing. We are so accustomed to this view that we mistake it for the truth, for reality. It is true that traditional grammars and dictionaries have helped us, fairly satisfactorily, to overcome the linguistic problems we have to deal with as human beings. (But even so, we cannot depend on a dictionary for help in translating into a foreign language). The failed promises of almost all NLP systems (MT, speech recognition, expert systems, automatic abstracting etc.) have demonstrated that something must be wrong with our linguistic knowledge.

This is why, however cumbersome and expensive it may be, it is absolutely necessary to analyze language from scratch. The fact remains, however, that language has to be described in a way that will be appropriate for NLP systems. In the Council of Europe corpus-based project, the *Multilingual Dictionary Experiment* (project leader John Sinclair, with participants from England, Italy, Hungary, Germany, Sweden and Yugoslavia) it has been demonstrated that monolingual and bilingual dictionaries are of no use when it comes to automatically translating a word from one language into another in cases where there is more than one alternative to choose from. A close analysis of the problems involved in the translation of nominalizations between German, French and Hungarian (also corpus-based) has also shown that all the descriptions available in dictionaries and grammars are inadequate, incomplete, and ultimately useless (see Teubert, 1992). To reduce the cost of this indispensable language analysis, particular tools have to be developed which arrange the facts (using statistics-driven devices for context analysis) and which even process them (with some human intervention) into algorithmic linguistic knowledge unbiased by theoretical preconceptions. Perhaps this will result in the finding that traditional categories like *noun, verb,* and *adjective* do

not, after all, reflect ontological categories.

Many of the more complex NLP systems available today either have an extremely narrow range of application or they are more like toys, like the pocket translation devices to be found everywhere. Many of the NLP applications available do not even use existing traditional linguistic knowledge. In the short term they can and must be improved, drawing on the sources available. Thus taggers and parsers can certainly be made more powerful (as this feasibility study shows), and this line has to be pursued for as long as there is no alternative.

But with future generations of NLP systems that can draw on newly processed corpus-based linguistic data, there will be a leap up to a higher level of quality. These will have a broader range of application, and they will be able to operate as robust systems in professional environments.

The European language industry in general, and smaller and medium-sized software developers in particular, have to assert themselves against powerful North American and Japanese competitors in a tight market that is quickly becoming one of the most important sectors of the economy. The European language industry needs support. In this study its immediate and emergent needs are identified: these are needs which must be served without delay. Corpora, processed language data, corpus-related resources, tools etc. have to be made available wherever they are needed, but because they are so expensive they cannot be developed by each country independently.

The NERC consortium therefore proposes the setting up of a strong network by the important corpus centres which already exist in the countries of the European Community. Every one of these national corpus centres must have immediate access to the whole array of corpus and corpus technology resources available in the whole Community.

The national corpus centres must be responsible for obtaining and preparing corpora, for processing linguistic data, and for tailoring software components to the specific needs of a given project. They can give advice in the preparatory stages of projects, and they can even identify market opportunities and propose new projects. They can pass on their combined expertise and their experience by organizing custom-tailored workshops and by offering training programs.

At the same time, the national corpus centres will continue to work on their longer-term tasks. They will build up comparable and shareable corpora and design tools, and process linguistic data for a new generation of NLP systems. Setting up a strong network, making use of the synergic effect by joining forces in developing corpus resources and in serving the NLP research and development community - these are steps long overdue. In the U.S., the Linguistic Data Consortium (LDC), funded by DARPA, began operation in 1992, and similar institutions are working in Japan. In this feasibility study, the members of the NERC consortium present their view that forming a network of national corpus centres in Europe should have a high priority.

2 A Description of Information Sources

2.1 *The NERC Workshop in Pisa, January 1992*

The Pisa Workshop incorporated a session on user needs, chaired by Nicholas Ostler, London (rapporteur: Wolfgang Teubert).

The following papers were presented in the user needs session:

(1) Roger K. Moore: *User Needs in Speech Research*

Moore mentions the following user needs: more annotated data; richer annotation (levels of transcription); standardized mark-up conventions and dictionary formats; conventions to cover all speaker-generated sounds; extensions to cover simultaneous acoustic/non-acoustic events; conformance to agreed standards/formats; considerations to do with the use of standards, formats and standard DBMS; and finally, annotations *not* embedded in the data.

(2) John McNaught: *User Needs for Textual Corpora in Natural Language Processing*

Textual corpora are of strategic importance for NLP. Some of the reasons for using corpora are that textual corpora provide good lexical coverage; they can be used as test-beds; and they can be used for constructing advanced NLP models. The goals of the NLP community are wider than those of the theoretical linguist. NLP research is interested in odd areas, including deviant language and particular sublanguages. Current NLP systems often perform badly because they are not based on processed corpus-based data. Requirements include human support in corpus processing; tools for skeletal analysis, including statistical techniques; a concentration on sublanguages; and an increase in authoring aids.

(3) John M. Sinclair: *Lexicographers' Needs*

Even a finite corpus of 100 million words will not be sufficient to satisfy the needs of lexicographers if the goal is a general purpose dictionary. Neology is an important aspect of vocabulary, involving not only new words but also new compounds and new meanings of existing words. A dynamic concept is needed: a corpus open-ended in size, reflecting the open-ended flow of language. Corpus analysis should be free from unchecked linguistic hypotheses. Tools should therefore come up with comparable results regardless of theoretical predilections.

(4) Henry S. Thompson: *Unscripted Spoken Corpora: Resources for Real Language Systems*

In order to create real language systems, there has to be a revolution in (theoretical) linguistics of the kind which phonetics has seen over the last 20 years. Not introspective competence, but real life language, in all its diversity, has to be dealt with. Rule-based grammars are not able to cope with this kind of natural language, unless complemented by stochastic models. The source of information for such models, and their test-bed, can only be large speech corpora, and these have to be transcribed orthographically. The reusability of theory-based linguistic annotation of corpora should be approached with scepticism.

For all NLP systems, large corpora are needed. Assembling such corpora is expensive, and the EC Commission will have to be convinced that corpus resources should be provided as public domain. Public funding will also ensure the application of the standards necessary for data exchange.

2.2 *Relevant literature on textual reference corpora 1985 - 1993*

An assessment of monographs, anthologies, textbooks, journals and conference proceedings on computational linguistics and related subjects published between 1985 and today has led to an indexed bibliography of about 700 titles, stored and implemented on an ORACLE database at the Institut für deutsche Sprache, Mannheim (Liebert, 1992, NERC-126). About 100 titles from this database were then selected and summarized as a second step towards a description of the user needs emerging from the literature covered. In every abstract the information about the corpora used (wherever available) is given; references to similar approaches are given at the end of a series of abstracts (Liebert, 1992, NERC-128).

On the basis of this bibliography and the collection of abstracts, user needs as emerging from the literature covered were analyzed and evaluated (Liebert, 1992, NERC-125). This report gives an overview of trends in today's corpus linguistics which reflects the recent growing interest in semantics, pragmatics and text analysis. It then identifies classes of users of corpora in general and of spoken language corpora. Central users of corpora and corpus-related resources are:

-the NLP research and development community
-the speech research community
-machine translation
-research and development of parsers
-lexicography (including the elaboration of lexicon components for NLP systems)
-computer-aided language learning (CALL)
-theoretical linguistics
-the corpus community itself.

The varying needs of these users are discussed. The analysis of these needs is then reflected in the recommendations given in this feasibility study in the context of the applicable work packages. The needs are evaluated in the light of the creation of a European network of national corpus centres.

In addition, the needs of `peripheral' users of corpora, and of corpus technology in particular, are discussed. Among those peripheral corpus/corpus technology users we count commercial and academic institutions interested in knowledge extraction from large full text data bases. The basic needs of this growing community of corpus users can be fulfilled with precompetitive/public domain software for text acquisition, text representation, basic access function and some more or less sophisticated tools for annotation (with special emphasis on spoken language texts). In addition, this group of users is also interested in software supporting the standardization required for data exchange. This group of peripheral users are associated, among others, with the following subject areas:

-Folklore documentation/oral history
-Psychotherapy/psychoanalysis

-Theology
-Social science/market research/content analysis
-Discourse analysis
-History/historical linguistics
-Medical language processing
-Information retrieval in insurance companies
-Legal informatics
-Terminology
-Multimedia/hypertext developments

Corpora are used by researchers and developers in many different domains. Some classes of users are fairly recent (e.g. those involved in medical language processing), while the growth of some known user groups could not be confirmed by the literature, e.g. translators, technical writers, advertisers, and pollsters. It is to be expected that some of the peripheral corpus users will become important clients of a European corpus network, when their particular needs concerning corpus design and corpus tools are examined in greater detail.

2.3 *A synoptic study of the needs of corpus users*

All the members of the NERC consortium have a long history of providing services, to both individuals and to commercial and academic institutions, relating to corpora and corpus-related resources. In the case of some of the members, it was possible to obtain the relevant data, i.e. for the Institut für deutsche Sprache, Mannheim, for the Instituut for Nederlandse Lexicologie, Leiden, and for the Istituto di Linguistica Computazionale, Pisa. Other NERC members contributed to this study by conducting surveys on the needs of potential users, namely the Institut de la langue française, Paris, the University of Malaga, and the ILC, Pisa. Birmingham contributed a short outline of actual users and user needs with regard to the COBUILD Corpus. The result of this investigation is a synoptic report (Endres and Wagner, 1992, NERC-119).

It is clear, then, that the methods used by the various partners are quite different. In some instances the requests of actual users are analyzed, in others the projected needs of potential users. This makes it difficult to come up with generalizations. This study analyzes the different kinds of users (institutional affiliations, disciplines (subject areas), and interests), attempting to provide more clearly delineated user profiles.

The other main topics of this study are the applications and domains, for which the help of the consortium members was sought (or expected).

Unfortunately, the question as to which institutions the users belong to cannot be answered satisfactorily. The data collected by Malaga, Leiden and Paris suggests that commercial users were not taken into account (or did not reply to the survey). In the case of Mannheim, most users are affiliated with non-commercial institutions. In Italy, on the other hand, over one third of identified (and potential) users have a commercial background.

As for the favourite subject areas which the (actual and potential) clients can be identified with, these are: *theoretical linguistics* for Mannheim, *computational linguistics* for Leiden, and

social and cultural sciences for Paris. As for Malaga, *computational linguistics* and *social and cultural sciences* are the top two disciplines. In the case of Pisa, there is a fairly equal distribution among *theoretical linguistics, computational linguistics, neurolinguistics, social and cultural sciences* and *language teaching*. The available evidence suggests that lexicology/lexicography is less represented among clients than was formerly assumed. Looking at the projects for which support is sought, it seems that about half of the projects are envisaging commercial applications (including dictionaries) even if they are carried out in not-for-profit institutions.

Applications, both actual and potential, for which support was sought were classified into 11 domains, ranging from *basic linguistic research*, to *NLP,* to *lexicograpy/lexicography* and to *neurolinguistics*. A surprising result is that for Paris *content analysis* constitutes about one third of all the instances of interest given, whereas neither Pisa nor Leiden mentions this. Other important domains are, as is to be expected, *NLP* (practically everywhere*), basic linguistic research* (Malaga, Mannheim and Paris), *sociolinguistics* (Paris), and *speech research* (Pisa). Outside of Paris and Mannheim, *lexicology* and lexicography are apparently not very attractive.

As for requests coming from the language industry, software developers and publishers, an increase of instances can be reported. However, in many cases the results of these requests cannot be processed. One reason for this lies in considerations of copyright, while another is the unsatisfactory state of corpus resources available today at the national corpus centres. There seems to be a growing demand for annotated corpora that cannot so far be met. This same applies to multilingual corpora, which are also not yet available.

2.4 *A survey of textual data*

2.4.1 *The organization of the survey*

2.4.1.1 *Goals*

Within the NERC framework, the *Survey of Textual Data* constitutes an independent Work Package. It is therefore presented in more detail here than are the other fact-finding activities associated with *User Needs*. To stress the importance of this survey and to underline our conviction that user needs should become the permanent concern of a future European Corpus Network, some recommendations for future work are given at the end of this subsection.

The main goal of this work package has been a comprehensive collection and evaluation of data on corpora and corpus technology, based on a worldwide survey. These data and their evaluation are intended to contribute to the overall goals of the NERC project by providing a sound basis on which the recommendations of this feasibility study can be founded.

The carrying out of the survey was devised in the form of a questionnaire. In the design of the questionnaire, special emphasis was laid on an adequate representation of the work packages of the NERC study. The survey was intended to provide a view both on the state of the art and on the emergent needs of actual and potential corpus users.

2.4.1.2 *The preparation of the questionnaire*

Drawing upon former experience, the design of the questionnaire was developed through close cooperation between Mannheim and Pisa, and the design was then thoroughly tested with the NERC partners. This led to alterations in the design; a new layout; the rewording of many questions; and the development of a guidance sheet for the addressees. The questionnaire was then distributed to Pisa, Susan Hockey (then of the Oxford Text Archive) and Donald Walker (Bellcore, Morristown, N.J.) for further comments.

2.4.1.3 *Cooperation with the CETH*

In January 1992 it was agreed that Susan Hockey, then Director of the Center for Electronic Texts in the Humanities (CETH), would participate in the survey, possibly co-operating with the newly established Linguistic Data Consortium (LDC), set up by DARPA. CETH and LDC represent, respectively, the scholarly and the NLP aspects of U.S. corpus activities, and to include them as partners in the survey was felt to be necessary in order to improve the return rate of the U.S. addressees. However, the inclusion of the CETH and its associates necessitated another redesign of the questionnaire, leading to a serious delay in the distribution of the survey.

In September 1992 the CETH sent out the questionnaires to holders of corpora, text collections and single electronic texts based on a list of addresses established by earlier research and updated by both the original and the new members of the NERC consortium. By mid-January 1993, about 40 filled-in questionnaires had been returned to the CETH. Copies were immediately sent to Mannheim, which was responsible for the evaluation within the NERC framework.

2.4.1.4 *Evaluation*

Due to the tight NERC schedule, evaluation was carried out on the basis of only 34 returned questionnaires. The data were interpreted and confirmed in the light of additional information on the corpus holders and other corpus sources. This knowledge was extracted and substantiated by informal contacts and by a survey of the relevant literature. The results of the evaluation are presented in (Rettig, 1992, NERC-136).

2.4.2 *Critical observations*

Several of the respondents commented on the questionnaire, and various institutions which had been addressed but had not responded were also asked for their comments. It was agreed that the questions covered all the relevant topics concerning machine-readable texts. But the need to extract as much relevant information as possible led to a rather lengthy questionnaire with a complicated structure that might have alarmed and discouraged possible respondents, in spite of the help and guidance offered. Another problem is that in many cases the detailed information pertaining to different domains does not reside with one person, but - especially in larger institutions - with a team of collaborators. Computer scientists may know little about corpus design, and linguists cannot describe the software they are using. As there is usually only one person per institution to work on questionnaires, the answers tend to have a number of shortcomings.

For owners of large corpora or text collections that have grown over a long period of time, there is often a huge variety of representation and annotation schemes employed; there is also a diversity of software tools used for encoding, accessing and manipulating the texts; and even of the operating systems used on a wide array of hardware, reflecting years of technological development. All this means that it is simply not possible for these organizations to provide a full picture of their operation within a reasonable amount of time. Finally, it might have been disadvantageous to send the same questionnaire to holders of large corpora and of single electronic texts. Some commentators felt that this led to a hybrid presentation of the questions, which was not conducive to eliciting the necessary care and effort required for an adequate response.

However, the institutions contributing to the survey design and the corpus holders who were contacted agreed that the data concerned have to be collected and have to be made available to all interested parties. Because this information has been lacking, progress has been slower than it would otherwise have been: unnecessary delays have occurred. A general consensus has emerged that the survey should not be viewed as a once-off initiative but rather as an ongoing service to the whole corpus creator and corpus user community. Recommendations concerning such a service are given below (section 2.4.4.).

2.4.3 *Evaluation*

It has to be clearly stated that for the reasons given above none of the large, well-known corpus centres has so far (January 5th) responded to the questionnaires. This will change over the next months. With respect to several important topics, the inclusion of these centres will change the picture considerably. Therefore, a second evaluation is planned for May 1993.

Our evaluation of the survey (based on 34 respondents, as mentioned above) takes into account the work packages and domains of the NERC feasibility study. The areas explored are:

- composition and design (2.4.3.1.)
- software tools (2.4.3.2.)
- annotations schemes (2.4.3.3.)
- text representation (2.4.3.4.)
- acquisition and reusability (2.4.3.5.)
- user needs (2.4.3.6.)

The evaluative report (Rettig, 1992, NERC-136) was passed on to the members of the NERC consortium immediately after completion. Thus it was possible to base the NERC recommendations also on the findings of the survey.

2.4.3.1 *Composition and design*

Characteristics of corpora

closed: 5
open-ended: 8

synchronic: 11
diachronic: 3

balanced: 10
not balanced: 2

all textual material stored: 10
monitor corpus: 1

core and shell organization of data: 4
core and periphery organization of data: 1

Future perspectives

The large majority of respondents state there is an urgent need to assemble multifunctional general language corpora. Only two respondents favour assembling specialized, task or sublanguage oriented corpora. As to the minimum size of a general language corpus, there seems to be little agreement, the answers ranging from 100,000 (!) words to 10 million words (large corpus centres are not yet represented in these results). Apparently there is an increasing demand for specialized corpora, but there seems to be little agreement as to the text type parameters to be used for the definition of sublanguages. Everyone agrees that multilingual corpora are useful, so sets of monolingual corpora built up in different countries should be integrated into a multilingual corpus. These multilingual corpora should also contain parallel texts.

With regard to written texts, the genres contained in the collections and corpora vary from case to case. About half of the respondents serve a research interest demanding only a very small number of genres, while the rest hold corpora including more then ten genres. For most respondents, *subject matter/topic* is the main or only selection criterion; and in most cases the complete text is included.

With regard to spoken texts, the most frequently identified genres are *conversation* and *debate/discussion*. The selection criteria vary. The most frequent one again is *subject matter/topic*, and a second group of relevant criteria concerns a variety of speech features.

Only two respondents so far have larger amounts of parallel texts (GILLBT Texts [African languages] and the ATR Dialogue Database [Japanese, English]).

2.4.3.2 *Software tools*

There is still a wide variety of access options among corpus holders. The majority of respondents use frequency software; sometimes this is commercially available but more often it has been developed by the corpus holding institutions. Almost every respondent now uses concordancing and/or indexing software (mostly of their own design). A good half of the access systems work interactively.

Annotation software is available at 13 institutions, and it includes dictionary look-up, morpho-

logy, and lemmatization. Most of the taggers used are rule-based (only two are stochastic). Of the five parsers, one is stochastic, while the others are rule-based. At three institutions, there are tools for semantic disambiguation. Ten annotation systems require manual intervention, while four do not.

Taggers and morphological analyzers are high on the list of future developments. The majority of corpus holders still seem to prefer their own developments, while three institutions are looking for software from other sources.

2.4.3.3 *Annotation schemes*

Part-of-speech tagging is carried out by 13 respondents and planned by two more. Five corpus holders tag manually, another five semi-automatically, and only three fully automatically. In ten cases the annotation scheme has been developed by the institution itself; but the answers do not permit satisfactory assumptions on the linguistic theories employed. 13 institutions have already lemmatized some of their texts, another three plan to do it, whereas 19 make no claims. Only two institutions have lemmatized 80 % to 100 % of their texts. In the majority of cases a semi-automatic mode of lemmatization is employed; and only one seems to be able to lemmatize fully automatically.

As for parsing, encoded structures have been specified as *phrasal structures, functional roles, logical/surface grammar, adjunct categories* and even *speech management structure*. As far as spoken corpora are concerned, levels of analysis are specified as: *repairs, referential domain and discourse function, turn-taking, communicative function* etc. Frequently encoded extralinguistic features are: *sex, age, region, dialect, date,* and *place and setting of recording*. Intonation and prosody features mentioned include: *pause*, *rising/falling contours, stress, duration,* and *pitch*. Other paralinguistic features (*gestures*, etc.) are encoded in some instances.

Most of the respondents agree that linguistic annotation is useful or even necessary. Among the applications and purposes of annotated corpora, we find listed: *stylistic description and classification, the study of grammatical features, variation, text types, linguistic research, building intelligent text processing software, register research, functional grammar, discourse research, exploratory research for psycholinguistics* and even *computational modelling*.

There is no consistent tendency as to the level of annotation aimed at. But a desire for a consensus on categories and their standardization for the various levels was frequently expressed.

2.4.3.4 *Text representation*

With the exception of two instances, the representation systems are neither SGML-based nor TEI-conformant, which reflects the long history of corpus linguistics. As for written texts, in most cases structural subdivisions and front matter are encoded. Also encoded are: *new page* (14)*, new line* (13), *font/style shifts* (8), *indentation* (8), and *hyphenation* (6). Among encoded linguistic features we find: *sentence boundaries* (6), *proper nouns* (3), *quotation*s (8)*, notes or marginalia* (3), and *editorial emendations* (3), to name the more frequent ones. The following multifunctional features are disambiguated: *upper/lower case letters* (13), *hyphenation* (5), and *periods* (5).

Spoken text is represented in *orthographical* (5), *enriched orthographical* (4), *phonological*

(3), and *phonetic* (1) forms. Punctation is often inserted. Also marked are: *word repetition* (9), *false starts* (6), *interrupted words* (5), *overlapping* (5), *interruptions* (5), *unintelligible speech* (4), and *omission of words* (4).

2.4.3.5 *Acquisition and reusability*

Texts are entered into machine-readable form by: *word processor* (15), *text editor* (10), *OCR* (7) *typesetting tapes* (4) and other means (7). In 11 cases respondents stated that they re-code texts into their own format. In nine cases this involves manual intervention. The output formats most frequently used are Post-Script (6), TEX, LATEX (6) and others (10).

Hardware configurations are given as IBM compatible PC (15), Apple Macintosh (5), workstation (8), mainframe (8). Operating systems are MS DOS (14), MACINTOSH (4), UNIX (4) and VMS (4). The most frequently used programming languages are C (7), PASCAL (3), SNOBOL (3), FORTRAN (3), and COBOL (2).

Corpora or texts are available to researchers (16), to all users (8), to libraries (3), but in no case explicitly to industry. The picture is very similar in terms of the availability of linguistic and statistical annotation to external users. Concordances and indexes are usually available to all users. Data is distributed to external users on diskette (24), magnetic tape (8), CD ROM (4), via network (6), and other (3). In ten instances there are restrictions on use imposed by the original copyright holders. Even so, the majority of respondents distribute their material without contracts (18), while 11 institutions give out licences. In 29 cases corpora or texts are not stored in archives or repositories; in three cases they are stored at the Oxford Text Archive, and in three other cases at other repositories.

2.4.3.6 *User needs*

Apparently all corpora, text collections and even most single texts are intended for an astonishingly wide range of uses, purposes and applications. Among the most frequently stated user needs we find:

- for lexicographic purposes (18)
- for research in linguistics (19)
- to extract statistical data (17)
- for literary research (13)
- to build up a multifunctional corpus (13)
- for stylistic research (11)
- for research on text types/sublanguages (10)
- for research in sociolinguistics (6)
- as test-beds for NLP-components/NLP-systems (6)
- for research in language learning/teaching (5)
- for the preparation of a scholarly edition (4)
- as test-beds for speech components/speech systems (4)
- to enrich a computational lexicon for NLP-systems (4)

- for commercial applications (3)
- for research in psycholinguistics (3)
- to extract statistical regularities for designing
- for probabilistic speech components/systems (3)
- for research on terminology (3).

In the sample, NLP-related purposes are specified only by owners of comparatively large corpora or collections. Some complex applications are specified. The *ATR Dialogue Database* is used as a test-bed for Japanese-English translations of spoken language, while the main use of the *Suzanne Corpus* is as a test-bed for the development of a comprehensive standard taxonomy of grammatical annotation for modern English. Some corpora are explicitly built up in order to develop NLP-related software for specific customers, among them Dutch publishers, a Swiss bank, and an aircraft manufacturer.

Generally speaking, corpora and text collections are seen to be useful for a wide range of applications. This is not self-evident, because the survey (particularly the list of addresses) is certainly biased in favour of academic institutions and individuals concerned with corpora or texts. So far, only very few respondents have a background of cooperation with commercial organizations or projects. As the survey is continued, it will include more commercial institutions dealing with corpora, so the full extent to which corpora are already used in NLP software development will become more apparent.

2.4.4 *Recommendations*

Even though the questionnaire developed for the NERC survey on textual data has a number of advantages compared with earlier surveys (this being due to its special focus on text [and speech] representation, corpus composition, recodifying software, accessibility, annotation, tools etc.), the overall structure is both too complex and at the same time too comprehensive in its aspirations. Institutions and individuals with just a few texts and little or no proprietary software become easily lost in the questionnaire, while those institutions holding large corpora and having a team for corpus technology development at their disposal find it extremely difficult to represent the history and growth of their projects within the framework of a strict questionnaire.

Nevertheless, in the course of the feasibility study the members of the NERC consortium became increasingly convinced of the necessity for the corpus user community as a whole to constantly have easy access to very recent and highly dependable information on all aspects of corpora, text collections, corpus technology, and developments in the field of corpus linguistics in general. A clearing centre having all the relevant information at its disposal would contribute to a cost-effective development of corpus resources by helping to avoid cumbersome sidetracks, by stimulating reusability of texts and techniques and by preventing the ineffective duplication of tasks which have already been successfully mastered.

The continuing necessity of information collection results in a web of short-term and medium-term recommendations:

A: Short-term recommendations

(1) The list of addresses of corpus holders will be further supplemented using the expertise of both the original and the new members of the NERC consortium, and questionnaires will continue to be distributed.

(2) Corpus and text holders who have received the questionnaire but have not yet responded will be (and are already being) encouraged, in various ways, to respond quickly.

(3) A follow-up evaluation of returned questionnaires will be carried out in May 1993 with the assistance of the new members of the NERC consortium.

B: Medium-term recommendations

(1) A Data Base containing all the information relevant to the community of corpus creators and corpus users will be designed and built up. This data base will document corpora, text collections and single electronic texts with specifications for text representation, accessibility, conditions of use etc. It will also provide information on acquisition and reusability software, on access systems and on annotation tools. Furthermore, it will contain a continuously up-dated bibliography on literature relevant to the areas of corpus linguistics, corpus technology, and applications of corpora for NLP development.

(2) The larger corpus centres will be visited by experts collecting the necessary information in the required form.

(3) Much shorter questionnaires will be developed for smaller institutions and individuals holding corpora and single electronic texts.

(4) Special questionnaires will be designed for non-linguistic institutions holding large amounts of useful machine-readable text material (e.g. newspapers, full text data banks, and texts on CD ROMs).

(5) Cooperation will be continued and new links will be established with partner institutions in the U.S., in Japan and other areas.

(6) The survey will be organized as an on-going process. All data collected will instantly be made available to the corpus community itself as well as to the NLP and speech communities.

The members of the NERC consortium agree that it is their responsibility to put these recommendations into effect.

2.5 *Interviews with corpus holders, corpus users and experts*

2.5.1 *A description of activities*

One of the central tasks of the work package *User Needs* was the collection of relevant data by conducting in-depth interviews with individual researchers and the representatives of institutions and organisations in the areas of:

- corpus building
- corpus technology development
- lexicography
- foreign language teaching
- speech
- NLP applications for corpora
- academic applications for corpora
- Artificial Intelligence and Machine Translation.

Our goal was to extract as much information as possible, not only on the actual uses to which corpora are put, to but also on potential or hypothetical uses which depend either on the availability of suitable corpora or on the development of suitable corpus technology, particularly annotation software that allows the identification and retrieval of linguistic phenomena for further analysis, be it automatic or intellectual.

From the interviews conducted it has become evident that the focus in corpus linguistics is gradually shifting from building corpora to developing software tools aimed at structuring mass data. The main bottleneck in NLP today is a lack of reliable linguistic data derived from and controlled by corpus analysis. This is true for the syntactic and semantic description of lexical items, and it is also true for grammatical data, for which much larger corpora are apparently needed than was earlier assumed; this is particularly the case for intratextual features beyond sentence boundaries, like anaphoric resolution. Even in areas of NLP where today there appears to be only a limited demand for corpora and corpus technology, like AI and MT, successful robust systems cannot be developed until there are lexicon and rule components based strictly on corpus analysis and not on the linguist's intuition or competence.

While concrete user needs (corpora, processed language data, acquisition and reusability software, access systems, and annotation tools) were at the centre of the interviews, it also became evident that one of the main problems in the design and preparation of corpus-based NLP systems today is the unavailability of relevant background information. In the case of smaller and medium-sized commercial enterprises, particularly, there is a strong demand for expert consultation on theoretical concepts in corpus linguistics, recent developments, new programs, funding facilities, competing approaches, access to information on corpus resources and access to the resources, evaluation of existing software, guidance on corpus design, market research, legal advice and even custom-tailored training programs for the academic and technical staff of the project.

Much money has been spent in vain and much working capacity has been wasted because of the lack of expert consultation, and also because project leaders do not always realize how necessary it is to obtain the relevant information in the preparatory stages of project design. But it also has to be admitted that up to now this kind of comprehensive information is not easily available at national corpus centres. Basic knowledge, ancillary information and corpus resources (both language data

and corpus technology) are scattered over many organisations still operating on their own instead of forming a strong network that can work as a clearing house.

Interviews with the following experts were transcribed in detail:

-**Bodil Nistrup Madsen**, Professor of Computational Linguistics at the Copenhagen Business School; Chairman of International Standard Organization TC 37, SC 3: Computational Aids Terminology (Topic: Corpus Linguistics and Terminology)
-**Paul Procter**, Senior Editor, International Dictionaries, Cambridge University Press; Chief Convenor of the Cambridge Language Survey (Topic: the Cambridge Language Survey)
-**Jeremy Clear**, then Project Manager of the British National Corpus, Oxford University Press (Topic: the British National Corpus)
-**Ramesh Krishnamurthy**, Development Manager, COBUILD project, Birmingham University School of English and Harper-Collins (Topic: the Bank of English)
-**Simon Sabbagh**, Eurolang project, SITE, Paris (Topic: Machine Aided Translation) [conducted by Pierre Lafon and Daniel Candel)
-**Kenneth Church**, Computer scientist, Bell Labs, Murray Hill, New Jersey (Topic: User Needs in U.S. Commercial NLP Research)
-**Susan Hockey**, Director, Center for Electronic Texts in the Humanities, Rutgers State University of New Jersey (Topic: Corpus Technology and Electronic Publishing)
-**Mark Liberman**, Director, Linguistic Data Consortium, University of Pennsylvania, Philadelphia (Topic: Serving the Corpus User Community)

Additional interviews were conducted, analyzed and evaluated, but for various reasons (confidentiality, limited relevance, repetition of arguments, etc.) they have not been fully documented. The interviewees were:

-**Robert Ilson**, University College, London, Chief Editor, International Journal of Lexicography, (Topic: Lexicography and the Use of Corpora)
-**Walter Grauberg**, Foreign language teaching methodology, Nottingham, U.K. (Topic: Foreign Language Teaching and Corpus Linguistics)
-**Antoinette Renouf**, Research and Development Unit for English Studies, Birmingham University, School of English (Topic: the Monitor Corpus)
-**Peter Mohler Zuma**, Zentrum für Umfragen, Meinungsforschung, Analysen [Center for Surveys, Market Research, Analyses], Mannheim (Topic: Content Analysis and Corpus Technology)
-**Gerhard Budin**, Infoterm, Wien (Topic: Terminology and Corpus Technology)
-**Khurshid Ahmad**, University of Surrey, Guildford, U.K. (Topic: The "Translators Workbench" and Multilingual Corpora) [conducted by Nicholas Ostler]
-**Dafydd Gibbon**, Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Germany (Topic: Corpus Technology and Speech Representation)
-**Uri Zernik**, General Electric Laboratories, Schenectady, New York (Topic: Commercial NLP Research in the U.S.)
-**Jonathan Cobb**, W.H. Freeman and Company Publishers, New York, N.Y. (Topic: Electronic Publishing and Legal Aspects)

-**Luciano Nebbia**, CSELT, Torino (Topic: Speech Research in Synthesis, Recognition and Understanding) [conducted by Vito Pirrelli]

2.5.2 *Evaluation*

The list of interviewees shows that information was sought from a diversity of experts who represent a wide range of subject areas. They are all interested in corpora, of course, and they all agree that better designed and larger corpora than those already existing have to be built up quickly in order to satisfy the needs of the corpus user community. For some purposes it may be sufficient to make the corpora available on CD-ROMs. But anyone who wants to use more sophisticated access software will either have to download the texts onto hard discs, or access them via physical electronic networks permitting interactive queries.

As for the subject areas represented in our interviews, the following aspects seem to be important:

(1)**Corpus providers**: Even in recent corpus projects, relatively little effort has been made to explore the needs of a wide user community. Corpora are still being built up to satisfy one basic need, e.g. application to lexicography. Moreover, the NLP research community has never specified the design and size of a general language corpus, although everyone agrees that these corpora are needed. The design question should be raised in connection with concrete NLP projects.

(2)**Lexicographers**: All lexicographers agree that corpora are essential for the quality of dictionaries (monolingual and bilingual), even if additional evidence is used. Most agree that for a general purpose dictionary a corpus should contain at least 100 million words, provided that the evaluation is basically done by hand. For context-based automatic word sense disambiguation, corpora would have to be much larger.

(3)**The NLP Research and Development community**: It seems that the large NLP research laboratories of the big transnational corporations like to see themselves as self-contained units, which do not depend on support from other corpus centres. Strong optimism still prevails that the linguistic knowledge (grammatical and lexical) available today is more than sufficient for the final perfection of NLP, including speech understanding systems. The next few years will demonstrate whether or not this attitude is justified.

However, everyone agrees that corpora and corpus-related resources have to be made available at a precompetitive level to smaller commercial (and academic) NLP developers, in order to help them compete with large corporations.

(4)**Related subject areas**: There is a growing number of related subject areas interested in the automatic analysis of texts, in processing raw textual data and in designing very specific software tools on the basis of more general corpus handling software. These are subject areas like *language teaching, psychotherapy, content analysis* (in the framework of market research), and the preparation of *scholarly editions*. The evolution of electronic publishing could also profit from close cooperation with the corpus community. For these and related subject areas it

is important to offer consulting facilities, access to information on the state of the art and access to data and software.

3 Recommendations

The work package *User Needs* has fed directly into the specific work packages dealing with text representation, corpus design, acquisition and reusability software, annotation schemes and annotation tools. As soon as results from the information sources described above became available, they were communicated to the other members of the NERC consortium. Therefore our recommendations have all been formulated against the background of the user needs identified in the execution of the feasibility study as a whole, regardless of which work package they are associated with.

The recommendations given below thus relate to topics of a more general nature, and deal with organisational, service-related and technical matters. Within the last few years practically all experts have come to share the understanding that the economic strength of the European language industry in general, and of NLP research and development in particular, depends on the availability of well-organized corpus service facilities. Only a strong network of the national corpus centres within the European Community will be able to provide the necessary expertise, linguistic data, and software which are the backbone of the development of commercially successful NLP systems.

Financial and other restrictions at the different stages of corpus technology in different European countries can easily slow down the creation of such a strong network. But it has to be remembered that the synergic effects of combining corpus resources depend on the financial means available at each national corpus centre. Additional efforts have to be made to adjust software and data to common standards; only when this is achieved will it be possible for each centre (and for each service client) to use the data and software available without specific adjustments. The members of the NERC consortium therefore urge the rapid implementation of the following proposals:

(1) The national corpus centres should form both a physical (electronic) network for data and software exchange and an organizational network for the coordination of their activities.

(2) Each national corpus centre should act as broker for all corpora, raw and processed linguistic data, corpus technology and other corpus-related resources available at any of the other corpus centres.

(3) The Network of National Corpus Centres (NNCC) should develop or commission the development of exchangeable, standardized basic corpus software for acquisition, reusability, exchange, access and annotation. This software must be language-independent; where this is not possible, it must be modular so that language-specific modules can be supplied by the national corpus centres. It is understood by the members of the NERC consortium that the operating system to be used for software development will be UNIX, but that software versions will also be produced under other operating systems as long as there is strong demand.

(4) The NNCC will contribute to European and international standardization initiatives. It strongly advocates strictly defined minimal standards for the representation of written and spoken texts (TEI-compatible).

(5) The NNCC will serve as a clearing house (or rather, a number of decentralized clearing houses) dealing with corpora, other raw linguistic data, processed linguistic data, corpus technology and other corpus-related resources. It will coordinate the world-wide survey on textual data as a continuous task, including non-linguistic text archives and repositories.

(6) The NNCC will be the European partner for comparable institutions outside Europe (e.g. the Linguistic Data Consortium and the Center for Electronic Texts in the Humanities).

(7) The NNCC will coordinate corpus activities in Europe by issuing recommendations and proposals and by commissioning smaller projects, to the advantage of all national corpus centers.

(8) The NNCC will offer extensive consulting facilities to commercial and not-for-profit language industry projects. The NNCC will also offer training activities to the NLP research and development community in general (e.g. by organizing workshops) and to specific projects (e.g. by providing in-house training).

(9) The NNCC will conduct market research on corpus-related NLP needs. It will publish a newsletter and brochures informing the corpus user community about the current state of the art, new projects and planned activities.

(10) The NNCC will develop guidelines dealing with copyright and other legal matters related to corpora and corpus-related material. It will provide a platform for a European settlement of the copyright issues involved. It will develop proposals for contracts between corpus providers and corpus users at a European level.

References

Teubert W. (1992): "Zur Behandlung von Präpositionalattributen im Wörterbuch". In: *Cahiers d'Etudes Germaniques,* No. 23.

Relevant NERC Papers

Endres B., Wagner F. (1992): "Synoptic Report on the Needs of Corpus Users", Interim Technical Report, IDS Mannheim, NERC-119.

Liebert W.A. (1992): "Textual Reference Corpora: User Needs. Indexed Bibliography of the Years 1985 to 1992", Working Paper, IDS Mannheim, NERC-126.

Liebert W.A. (1992): "Textual Reference Corpora: User Needs. Abstracts of main articles from the bibliography Textual Reference Corpora", IDS Mannheim, Working Paper, NERC-128.

Liebert W.A. (1992): "Textual Reference Corpora: User Needs. A report on the relevant literature in the years (1985-1992)", IDS Mannheim, Working Paper, NERC-125.

NERC Consortium, (1992): "Workshop on textual corpora", Report to EC, Pisa, NERC-82.

Rettig H. (1992): "Evaluative report on the Corpus Survey", IDS Mannheim, Technical Report, NERC-136.

Chapter 2

Corpus Design Criteria

1 The problem

The objective of the NERC project is to investigate the feasibility of a permanent network of European Centres responsible for providing corpus data, facilities, and services available in the public domain. This entails the definition of a common European strategy for corpus composition. The corpus design must be responsive to the needs and requirements of present and future corpus users and applications, among which are activitities in the field of language engineering and technology. The aim of workpackage 6, 'corpus design criteria', was to investigate which corpus design is the most feasible and multifunctional.

Different types of users need different types of corpora. Several alternative strategies for corpus composition can therefore be identified. For example, workers in NLP emphazise the importance of sublanguages; a corpus is conceived as a collection of sublanguages. In the USA particularly, the relevance of the mass of data is stressed (Pisa Workshop on Corpora, 1992, NERC-82); a corpus is conceived of as a collection of all the texts available in the public domain. Lexicographers and other language analysts stress the importance of large, carefully organised collections of texts; the corpus is conceived as a "balanced" corpus, i.e. it is tuned in such a way that it can be viewed as a small scale model of the linguistic material to be studied. During the project, the following options for a polyfunctional corpus have been considered:

(a)A corpus based on the principle of availability only. A collection of all the printed texts and transcriptions of spoken texts available in the public domain.

(b)A corpus based on design criteria.

1.Specialized task- or sublanguage-oriented corpus. A collection of various selected sublanguages.
2.A general purpose corpus. A collection of a broad variety of written and transcribed spoken material reflecting language variety.
3.A monitor corpus. A large and dynamic text corpus part of which is discarded and replaced by a new one after the textual material is automatically analyzed for specific linguistic or textual phenomena.

(c)A monolingual or multilingual corpus.

Which type of corpus is most appropriate in the present framework, was determined by the requirement of multifunctionality, and by what is feasible in the short, medium and longer term. Clearly, then, this Work Package is closely related to Work Packages 1 and 2, Survey and User Needs respectively.

2 Investigations

2.1 *The polyfunctionality aspect*

The polyfunctionality aspect was investigated by an evaluation of three reports on users and user needs which resulted from Work Packages 1 and 2: (Liebert, 1992, NERC-125), (Endres and Wagner, 1992, NERC-119), and (Rettig, 1992, NERC-136). (Liebert, 1992, NERC-125) distinguishes classes of users as they appear in the literature. (Endres and Wagner, 1992, NERC-119) is a synoptic report on the needs of corpus users based on the related NERC partners. (Rettig, 1992, NERC-136) is an evaluative report on the NERC Corpus Survey. In an earlier stage of the project, the Workshop on Textual Corpora 1992 in Pisa provided useful information on recent developments in this field.

2.2 *The feasibility aspect*

The feasibility aspect was investigated by an evaluation of the state of the art in corpus design, and by an exploration of the conditions on the actual collection of textual material for corpora.

2.2.1 *Corpus design: state of the art*

The state of the art was investigated by an inventory and evaluation of what is proposed or realized for corpus composition. Topics included design parameters in written and spoken language corpora as well as quantitative factors. Design parameters concerned different corpus types, contents, selection principles, text types, their hierarchical structure, and their definition in terms of external (functional) and internal (linguistic) parameters. Quantitative aspects concerned size, proportions of text types, and sampling techniques.

Some reports have been prepared in which the state of the art of corpus design (design parameters, quantitative aspects) is evaluated: (Alvar Ezquerra and Corpas Pastor, 1992, NERC-84), (Kruyt and Putter, 1992, NERC-129), and (Malaga Group, 1992, NERC-12). (Huizhong, 1986, NERC-95), (Nakamura, 1992a, NERC-53), and (Nakamura, 1992b, NERC-97) particularly focus on text typology based on internal, linguistic parameters. An overall evaluation of topic is provided by (Kruyt, 1992a, NERC-93).

2.2.2 *Collection*

Experimental evidence was obtained with respect to the availability and acquisition of written and spoken text material for corpus building. Topics concerned the potential and actual suppliers of language material, right holders and permissions, the characteristics of available written and spoken material, and the costs of acquisition and data processing. Another investigation concerned the availability and usefulness of tools for text classification on the basis of external parameters.

Reports on actual experiences with the availability and acquisition of textual material have

been prepared by (Krishnamurty, 1992, NERC-57), (Vercouteren and Meijer, 1992, NERC-86), and (Vercouteren et al., 1992, NERC-52). (Dutilh and Kruyt, 1992, NERC-94) discusses some text classification systems and includes a test on text classification. These issues have been evaluated and related to corpus design in (Kruyt, 1992b, NERC-115).

2.3 *Final report*

The results in terms of user needs, design criteria, and collection, mentioned in the previous sections, have been evaluated in the present framework by (Kruyt, 1993, NERC-168). Minimal design requirements for a polyfunctional corpus are formulated and some draft recommendations are presented. These are included in an abridged and revised version in the remaining sections of this chapter.

3 Main results

3.1 *Users and user needs*

Corpora are used by researchers and developers from many different disciplines and domains. Some disciplines are considered 'primary academic disciplines', other disciplines 'disciplines of both academic and commercial interest'. The latter include: lexicology and lexicography, computational linguistics and related fields, communication theory and practice, and language teaching and computer based training. Although the situation varies somewhat from one country to another, the overall picture is that more than half of corpus users are working in fields of actual or future commercial interest, running applications that might not be so different from that of future commercial users (Endres and Wagner, 1992, NERC-119). For the most part, these users cover user groups assumed to be the central users of a future network of European corpora (Liebert, 1992, NERC-125). In the present framework, the needs of these users are most interesting.

The following user needs with respect to corpus design were identified:

-There is a salient common need for (very) large corpora. There seems to be no common concept of minimum size. Neither does there seem to be a maximum size: corpora should be open-ended, or as large as possible.

-Assembling multifunctional general language corpora is most urgently needed. The corpus should cover many domains, registers, communicative situations etc., preferably organized in such a way that it is assumed to be representative or "balanced", and preferably with the possibility of separating specific subcorpora. Sophisticated corpus design should be guided by recent classification studies. Extensive documentation about the constituent texts is needed.

-One may expect a lot of demands for specialized corpora. There probably exist nearly as many needs for specialized corpora as there are different research interests.

-Open-ended monitor corpora are needed, especially for updating.

-There is an increasing request for speech corpora and corpora of spoken language. The needs and desiderata are still very heterogeneous.
-Multilingual corpora, containing -among others- parallel texts, are of increasing importance, especially for the European Market. Sets of harmonized monolingual corpora built up in different countries should be integrated into a multilingual corpus.

Meeting the needs of corpus users is complicated by several factors. The most general problem is copyright on the written and spoken texts to be included in a contemporary corpus. This is also a reason why requests from the language industry, the computer industry and publishing houses often cannot be granted. Especially relevant to multilingual components in corpora is a lack of knowledge of international law for contracts.

3.2 *Corpus design: state of the art*

The NERC Corpus Survey shows that "the 'typical' corpus is balanced and synchronic and that all textual material is stored. A probably relevant issue of discussion among corpus-holders could be about possible forms and modes of corpus organization." (Rettig, 1992, NERC-136). 'Topic' or 'subject matter' is the most frequent selection criterion, for both written and spoken corpora.

(Alvar Ezquerra and Pastor, 1992, NERC-84) and (Kruyt and Putter, 1992, NERC-129) report on studies of corpus designs in the literature. In corpus design, diversity is dominant, at various levels: size, selection principles, text types, structure, sampling techniques, etc. Many decisions on corpus composition are described but not accounted for. An evaluation of the common features in the variety of corpus designs resulted in a general text typology, a separate typology for spoken language corpora, and a separate subject ('topic') typology, 'topic' being a dominant selection criterion (Kruyt and Putter, 1992, NERC-129); see appendix A). The proposal for the 'core corpus' design presented in section 6 is based on these typologies.

Different major approaches in corpus composition have been evaluated in a polyfunctional framework (Kruyt, 1992a, NERC-93). Methods applied at the various levels of corpus composition (design, selection principles, text types, their structure and proportions, sampling techniques) are presented in a more or less contrastive way. From the options discussed, the following are thought to apply to a polyfunctional corpus (Kruyt, 1992a, NERC-93: section 3):
-The corpus should cover a broad variety of textual and language phenomena.
-It should be very large.
-It should include written and transcribed spoken text.
-It should contain full texts rather than samples.
A well-coordinated corpus was considered best to meet these conditions. "Design principles should account for typical patterns of use (production/reception) and a wide range of registers or sublanguages, text types be well-defined and based on both external and internal parameters. It should be well-documented. Parts of the corpus could be selected and expanded for particular user needs and applications" (Kruyt, 1992a, NERC-93: section 3). A monitor corpus is considered a second alternative.

Some of these requirements, for example text typologies based on internal parameters, are not

feasible in the short term. With respect to external parameters, classification systems available for library purposes are useful to some extent only, as many texts get no classification code at all (Dutilh and Kruyt, 1992, NERC-94).

3.3 *Collection*

The investigations into the availability and acquisition of language material for corpus building (Krishnamurty, 1992, NERC-57), (Vercouteren and Meijer, 1992, NERC-86), (Vercouteren et al. 1992, NERC-52), have been evaluated in relation to corpus design in (Kruyt, 1992b, NERC-115). The main conclusions are summarized here.

-Corpus practice shows that a broad variety of language material for corpus building is available. Appropriate spoken material, however, is scarce. The availability of language material varies in different countries.

-Most textual material is still in paper form only. Machine-readable texts mainly concern newspapers and books. A large electronic corpus covering a broad variety of language uses can therefore not yet be obtained by acquisition of machine-readable material only. The increasing availability of electronic textual material supports the tendencies towards very large full-text corpora and monitor corpora.

-Legal issues, a.o. copyright, complicate acquisition and use of corpus materials.

4 Conclusions

The various investigations concerning polyfunctionality and feasibility have provided results that do not essentially diverge as far as corpus design is concerned. We can therefore be rather sure that the following conclusions apply.

A multifunctional corpus should preferably meet the following conditions:

-it should be large and open-ended,

-it should contain written and transcribed spoken language,

-it should cover a broad variety of language types, 'topic' being taken into account as a selection criterion,

-it should contain full texts rather than samples,

-it should be extensively documented, so as to facilitate the selection of specific subcorpora, which can be expanded in order to construct large specialized or task-oriented corpora,

-in view of the need for multilingual data, part of the corpus should be designed according to the same specifications as other national (sub)corpora, so as to be able to integrate them into a multilingual corpus; parallel texts should be provided for as well.

As a further development, monitor corpora could be constructed, provided with a set of new types of analytical linguistic software. A monitor corpus allows new kinds of patterns of language to be continuously detected.

The intended multifunctional corpus comes very close to the option B2-corpus mentioned in

section 1, combined with characteristics of the options B1, B3, and C.
A corpus based on the principle of availability only (option A) does not cover the required broad variety of language uses. If 'sublanguage' is conceived of as the restricted task-oriented or specific-domain languages usual in NLP-applications, a collection of selected sublanguages (option B1) can hardly meet the condition of covering a broad variety of language uses. The intended multifunctional corpus can function as a frame of reference for the specific characteristics of such specialized, task-oriented or specific-domain languages.

If these requirements are related to what is practically feasible and ready for implementation, some factors have to be taken into account.

-The general outlines for a prototype multifunctional corpus are clear. For implementation, further specifications are needed at various levels. These are presented in the NERC-proposal for the 'core corpus' design presented in section 6.
-The amount of data per language type to be included in the corpus is influenced by the different availability of the various types of language material (e.g. spoken language, material in machine-readable form).
-Legal constraints on use may influence the selection of corpus data.
-Different situations in different countries have to be taken into account in the implementation schedule.

5 Recommendations

It is useful to make a distinction between a core component and a peripheral component of a national language corpus. The core component is defined as the minimal multifunctional corpus sketched above and specified below. The peripheral component contains all other electronic textual materials stored at the national node in the European network. This component can contain material obtained by the availability principle, or acquired for a particular need or project. The peripheral component has relevance in view of the obvious usefulness of very large collections of texts (cf. option A). The two components are to be clearly distinguished by documentation.

The core component meets the conditions outlined in section 4. Its composition should be as equal as possible for the various EC-languages, so as to be able to integrate national language core corpora into polyfunctional multilingual corpora. Extensive documentation per text of all features that could have relevance for specific selections is strongly advised. This ensures flexibility of corpus use. Specific text categories can be selected, expanded or systematically re-categorized in terms of changing user needs criteria.

In order to guarantee consistency among the various EC-language core corpora, it is advised that, during implementation, a board of representatives of the national centres, supported by EAGLES, evaluates the practical feasibility of the design in the various countries.

For the implementation schedule of a European network of national corpora, it is useful to distinguish between the languages for which large corpora are available already, and the languages for which a corpus has not yet been constructed.

In the short term, a network of national centres contains at least a multi-million corpus of parallel EC-documents. Other materials available can be added on a voluntary basis in each node.

In the medium term, core corpora should be constructed in the EC-languages. Other parallel texts should be included.
In the longer term, monitor corpora should be constructed as a complement to the continuously updated core corpora.

6 Core corpus design

For the core corpus we suggest a design with the following characteristics:

* Size: 60 million words
* Language variant: standard
* Time limit: contemporary language: 1980 ->
* Contents:

Written component

59 million words -> 10 million parallel texts
49 million comparable texts

Selection principle: topic
Text types:
'science & technology': 35 %
'society/daily life': 45 %
'belief, thought, arts': 20 %
Text media:
newspapers & magazines 45 %
books 45 %
ephemera & correspondance 10 %

Spoken component

1 million words

Selection principle: topic
Text types:
'science & technology': 35 %
'society/daily life': 45 %
'belief, thought, arts': 20 %
Text medium:
monologue: public/private 30 %
dialogue: public/private 70 %

The following remarks may clarify some of the selected specifications.

The proportions of written vs. spoken language is determined by feasibility.

'Topic' or 'subject matter' appeared to be a very important selection principle for both written and spoken language. This was confirmed by the investigations on user needs. As this selection principle has relevance for specialized languages as well, it was selected as the main selection principle.

The text types 'science & technology', 'society/daily life', and 'belief, thought, arts' have been established by grouping the items in the subject typology in appendix A (table 2). Global groups are preferred over detailed specifications in view of feasibility and consistency. The percentages are based on the frequency data presented in the subject typology.

'Science & Technology' includes: technics (12), medicine (8), history (8), science (7), physics (6), biology (6), mathematics (5), anthropology (5), language (5), architecture (5), computing (4), agriculture (4) geography (4), chemistry (3).

'Society/daily life' includes: law (11), politics (8), economy (7), education (7), sociology (7) civilisation (6), military (5), media/communication (5), traffic/transport (3), finance (3), ecology (3), and sports (11), leisure (7), household (5), travel (5), fashion (4).

'Belief, thought, arts' includes: religion (13), arts (9), philosophy (7), psychology (7), literature (4).

The topics are treated in texts in various media. The selected specifications are based on the text typologies in appendix A (tables 1 and 3), the percentages on the assumed feasibility, in particular the availability of machine-readable material.

With respect to the spoken language media, 'monologue' includes lecture/speech/sermon, commentary, and narration; 'dialogue' includes conversation, discussion/debate, and interview.

The core corpus design meets the conditions outlined in section 4. The specifications are based on the results concerning the state of the art in corpus design (cf. the typologies in appendix A). Considerations of feasibility (section 4) are taken into account as well. In general, potential users can only indicate but not specify what corpus they exactly need. This specific elaboration of the global minimal requirements is therefore to be considered as a concrete proposal open to discussion by corpus users and the new NERC partners.

Relevant NERC Papers

Alvar Ezquerra M., Corpas Pastor G. (1992): "Design criteria". Working Paper, University of Malaga, NERC-84.

Dutilh M.W.F., Kruyt J.G. (1992): "Feasibility experiment design criteria. Investigation into text typological classification tools", Working Paper, INL Leiden, NERC-94.

Endres B., Wagner F. (1992): "Synoptic report on the needs of corpus users", Interim Technical Report, IDS Mannheim, NERC-119.

Huizhong Y. (1986): "A new technique for identifying scientific/technical terms and describing science texts". In: *Literary and Linguistic Computing*, 1(2): 93-103, 1986. NERC-95.

Krishnamurthy R. (1992): "Data collection", Working Paper, COBUILD, NERC-57.

Kruyt J.G. (1992): "Evaluative report on design criteria for corpora construction I: selection principles", Interim Technical Report, INL Leiden, NERC-93.

Kruyt J.G. (1992): "Evaluative report on design criteria for corpora construction II: availability of texts for corpus building", Interim Technical Report, INL Leiden, NERC-115.

Kruyt J.G. (1993): "Design criteria for corpora construction in the framework of a European Corpora Network", Technical Report, INL Leiden, NERC-168.

Kruyt J.G., Putter E. (1992): "Corpus design criteria", Working Paper, INL Leiden, NERC-129.

Liebert W.A. (1992): "Textual Reference Corpora: User Needs. A report on the relevant literature in the years 1985 to 1992", Working Paper, IDS Mannheim, NERC-125.

Malaga (1991): "Design of a Spanish corpus within the framework of a European corpus", Working Paper, University of Malaga, NERC-12.

Nakamura J. (1992a): "Determining text typology by means of Hayashi's quantification method type III", Working Paper, University of Tokushima and COBUILD, NERC-53.

Nakamura J. (1992b): "On the structure of the Bank of English based on the distribution of pronominal forms", Working Paper, University of Tokushima and COBUILD, May 1992. NERC-97.

NERC Consortium (1992): "Workshop on Textual Corpora", Report to EC, Pisa, NERC-82.

Rettig H. (1992): "Evaluative report on the corpus survey", Technical Report, IDS Mannheim, NERC-136.

Vercouteren W., Meijer M., Grinwis J. (1992): "Supply and demand on the linguistic market", Working Paper, INL Leiden, NERC-52.

Vercouteren W., Meijer M. (1992): "Facts and figures on data collection", Working Paper, INL Leiden, NERC-86.

APPENDIX A

Schematic Outlines of Text Typologies

These tables present the text typologies and the distribution of the distinguished categories over the literature items. They present schematic overviews that facilitate the observation of common features and differences.

Terminology has been interpreted in order to find common notions in the diversity of corpus compositions. On the basis of the results of the interpretation process, generally applied rather than corpus-specific text have been distinguished, and only these are presented in the tables. As a consequence of the evaluation method, original hierachical relationships between text types have been replaced by new hierarchies based on grouping of related text types.

The meaning of "+" in the tables is 'explicitly included as a text type', "-" means 'explicitly rejected as a text type'. "S" refers to a spoken language corpus or corpus component, "W" refers to a written corpus (component). "A" refers to the feature 'administrative data' (rather than text type). A question mark has been placed in doubtful cases.

The subject typology is ordered according to frequency.

Table (1) Text typology.

	Bou W+S	Hoz. unspec.	Svartvik c.a. W+S	Juilland c.a. W	Kuccra c.a. W	de Vriendt-de Man S	Uit den Bogaart W+S	de Jong S	Lara W+S	Allén W	Altenberg S	Renouf W	Renouf S	Gonzalez c.a W	Stapborsius W	Feldweg S	Morales W	Collins c.a. W+S	Summers (selective component) W	Crowdy S	Bindi c.a. W	Martin c.a. W+S	Werkgroep Taalbank W+S	Atkins c.a. W+S	Biber W+S	Malaga W	Malaga S
LITERARY GENRE																											
poetry (3+, 4-)			−							+		−							+								−
narrative (10)				+								+	+	+		+						−		+	+	−	+
(auto)biography (4)					+							+	−	+										+	+	+	+
novel/short story (7)							+					+	+	+							+	+		+		+	
historical (4)												+	+	+								+				+	
sciencefiction (3+, 1-)												+	−						+			+				+	
humour (7)					+							+	−						+		+	+				+	+
theatre/drama (7+, 3-)			+	+	−							−		+			+	+	+					+			−
TOPIC																											
topic (11+, 2-)			−				+					+	+		+				+	+		+	+	+	+	+	−
MEDIUM																											
books (7)										+		+	−						+								
letters/correspondance (5)			+									+	−		+			+	+		+			+	+		
newspapers (17)	+		+	+	+		+	+	+	+		+	−		+		+		+		+	+		+	+	+	
brochures/leaflets (6)												+	−		+				+			+					
FICTION/NON-FICTION																											
fiction (11)			+		+					+		+	−		+				+			+	+	+	+		
non-fiction (3)												+	−		+											+	
STYLE																											
distance (9)			+					+	+			+	−							+		+					+
popular/solemn (7)			+		+				+			−	−						+		+	+	+				
specialised/lay (10)			+	+	+				+	+		−	−	+		+					+			+			
(= technical)																											
OTHERS																											
handbooks/textbooks (7)	+		+									+	−	+	+				−		+				+		
translations (3+, 1-)		+														+	−		−					+			−

Table (2) Subject typology.

	Bou	Svartvik c.a.	Juilland c.a.	Kucera c.a.	Uit den Bogaart	Lara	Altenberg	Renouf	Gonzalez c.a	Slaphoutsius	Moralcs	Summers (selective component)	Crowdy	Bindi c.a.	Martin c.a.	Werkgroep Taalbank	Biber	Malaga
Religion (13)	+			+	+		+	+		+		+	+	+	+	+	+	+
Technics/-ology (12)			+		+	+		+	+	+	+		+	+	+			+
Law (11)		+			+			+	+	+	+	+	+	+	+	+	+	+
Sports (11)		+			+			+		+	+	+	+	+	+	+	+	+
Arts (9)		+		+				+		+	+		+				+	+
Politics (8)							+	+			+	+	+	+		+		+
History (8)					+		+	+		+	+		+					+
Medicine (8)				+				+	+		+		+	+	+			+
Philosophy (7)					+			+	+		+		+	+	+			+
Economy (7)					+			+	+		+		+	+	+			
Education (7)					+			+			+		+	+	+			+
Psychology (7)					+			+	+		+		+	+				+
Science (7)					+			+	+		+		+	+				+
Sociology (7)		+	+	+		+						+	+			+		
Leisure (7)					+			+	+		+		+	+	+			
Civilisation (6)								+	+		+	+	+	+	+			
Physics (6)					+			+	+		+					+	+	+
Biology (6)					+			+	+		+		+	+				
Mathematics (5)					+			+			+		+	+				+
Household (5)								+	+		+		+	+				
Travels (5)								+	+		+				+		+	+
Anthropology (5)					+			+			+				+		+	+
Military (5)								+			+		+				+	+
Media/communication (5)								+	+		+	+					+	+
Language (5)								+			+	+	+				+	+
Literature (4)								+			+	+	+	+				
Architecture (4)									+		+	+	+					
Fashion/clothes (4)								+			+		+				+	+
Computing (4)								+			+	+					+	+
Agriculture (4)								+			+			+			+	+
Geography (4)									+			+	+				+	+
Ecology/Environment (3)									+				+	+			+	+
Traffic/Transport (3)							+					+						+
Chemistry (3)										+		+		+				
Finance (3)										+		+	+					

Table (3) Spoken typology.

	Svartvik e.a.	de Vriendt - de Man	de Jong	Altenberg	Renouf S	Feldweg	Collins e.a.	Crowdy demographic	Crowdy context governed	Martin e.a.	Aikins e.a.	Biber	Malaga
COMMUNICATIVE SITUATION													
face to face	+			+	+			+		+			+
telephone	+			+	+			+		+			+
NUMBER OF PARTICIPANTS							A			+			
monologue	+							+			+		
dialogue	+							+			+		
GENRE MONOLOGUE													
lecture				+	+			+		+			+
commentary	+		+		+		+	+					+
speech				+			+	+		+			
sermon				+				+					
demonstration					+			+					
narration					+	+							+
GENRE DIALOGUE													
conversation				+	+	+	+	+					+
interview		+	+		+			+					+
discussion/debate				+	+			+					+
consultation								+					
proceedings								+					
chat shows								+					
meeting/gathering								+					+
SETTING							A						
education				+	+			+		+			+
commerce/business		+						+					+
radio		+		+	+			+	+	+			+
tv					+			+	+	+			+
political and social organs				+				+					+
private - personal		+						+		+	+		
work										+			
institutional								+			+		

Spoken typology (2)

	Svartvik e.a.	de Vriendt - de Man	de Jong	Altenberg	Renouf S	Feldweg	Collins e.a.	Crowdy demographic	Crowdy context governed	Martin e.a.	Atkins e.a.	Biber	Malaga
SPONTANEITY													
scripted/planned							A				+	+	
unscripted/preplanned	+						A	+				+	
spontaneous/unprepared	+	+					A			+		+	+
TOPIC	+			+	A	+	A	+			+	+	?
STYLE													
distance	+		+		+		A					+	
specialised/lay (technicality)								+				+	
PARTICIPANT DETAILS													
sex			+		?		+	+			?		
age			+		?		+				?		
ethnic group					?	A	A						
region			+		?	A	+	+			?		

Table (4) Regional and temporal aspects.

	Malaga S	Malaga W	Biber W+S	Atkins e.a. W+S	Werkgroep Taalbank W+S	Martin e.a. W+S	Bindi e.a. W	Crowdy S	Summers W	Collins e.a. W+S	Morales W	Feldweg S	Staphorsius W	Gonzalez e.a W	Renouf S	Renouf W	Altenberg S	Allén W	Lara W+S	de Jong S	Uit den Bogaart W+S	de Vriendt-de Man S	Kucera e.a. W	Juilland e.a. W	Svartvik e.a. W+S	Hoz. unspec.	Bou W+S
Regional aspects (22)			+		+		+	+	+	+	+	+	+	+	+	+	+		+	+		+	+	+	+	+	+
Temporal aspects (16)			+	+			+	+	+	+	+			+	+	+				+	+		+	+	+		

Chapter 3

Text Representation:
Written Language/Spoken Language

A. Written Language

1 Introduction

This report is the starting point for research which should, in the framework of the NERC project, lead to the definition of a minimal level of textual representation for European corpora. As in all projects that aim to win wide support, the legitimacy of the enterprise implies taking into account work on standardisation in the field concerned. Consequently, this study is based on a critical examination of the Text Encoding Initiative (TEI). This project - the history of which is sketched out below - aims to define a certain number of guiding principles for the coding and exchange of texts in electronic form, and to get these principles adopted on the widest scale possible.

The principles outlined in this project do not constitute a standard in the way that the official ISO does. However, the research has been carried out with standardisation in mind, that is, with the objective of being accepted and used by the largest possible number of people. The principles derive from an ISO standard, namely the SGML standard, i.e. Standard Generalized Markup Language (ISO 8879), and are to be seen as an application of that standard. It should be noted that the SGML standard is a metalanguage that allows the generic structure of documents to be described and their encoding to be carried out in conformity with the model thus defined (the Definition of Type of Document, or DTD). Using the SGML standard thus calls in the first instance for the definition of an application of the standard, which involves, amongst other things, the conception and writing of one or several DTDs.

Could standards other than those proposed by the SGML have been chosen to define a text representation scheme in the framework of the NERC project? The answer is, categorically, no, and the reasons will be explained in the first part of this report. Assuming this to be the case, is the TEI the appropriate application?

The TEI is not an application in the usual meaning of the SGML standard. Within the TEI framerwork, one does indeed define a certain number of DTDs on a mode that TEI editors have humorously termed a "pizza" mode (a common base and toppings to be chosen from the selection proposed) as opposed to the "a la carte" mode which would consist in proposing a choice from a group of DTDs. But one also defines the rules of extending and modifying these DTDs. In this respect, the TEI constitutes a meta-application of the SGML standard. Its use implies a preliminary adaptation. The TEI is thus a stage between the total generics of the SGML standard and the specificities of a given particular application.

It is possible to ignore this standardisation stage and to define a "corpus" application directly from the SGML standard.

At the time of writing this chapter, phase 1 of the TEI work is fully published and phase 2

partly so, this having been due for completion at the end of 1992, give or take a few months. Ignoring the TEI today amounts to placing oneself outside the standardisation movement. We do not think that the TEI proposals are restricting enough to justify such an attitude. Positioning oneself outside the TEI would mean redoing the work of consensus-seeking that has already been undertaken within the TEI framework. One may, of course, question this or that choice, and prefer a solution that differs from the one recommended by the TEI. But we must bear in mind that these choices will have already been discussed and that the adoption of another solution would not necessarily mean achieving a larger consensus. Moreover, the solutions recommended by the TEI are, in a large number of cases, multiple and skeletal, which makes it possible for these solutions to meet particular demands satisfactorily.

The TEI thus offers an interesting research framework for defining a minimal level of textual representation.

-First, it saves time: we can discuss and adapt the TEI proposals instead of starting out from scratch.
-Second, it gives legitimacy to the recommended solutions, as these will be part of an initiative which, although it does not have the official status of a standard, does nonetheless have a de-facto character of standardisation (information available currently suggests that the TEI could be proposed as an ISO technical report).

In the second part of this report we shall define what the TEI is and what it means to work within the TEI framework. The final sections of our account outline our proposals for defining a minimal level of textual representation. These proposals are grouped into four sections. The first section looks at internal variables. It covers the various problems of defining a model of the minimal structure of the document. The second section discusses peri-textual variables. It covers problems of document identification. In TEI terms, this means the definition of a header. Section three deals with the localisation of variables, that is the method adopted to encode references within the corpus. The final section considers a number of issues which are peculiar to the French language.

2 Why the SGML standard?

There is only one method available today for representing documents with a view to obtaining electronic versions that can be exchanged widely and can be used in various contexts: this method is the application of the SGML standard. SGML is the only language that is sufficiently neutral to fulfil these functions with regard to the types of document and to the functionalities that one wishes to mark them with. It is the only one that, thanks to its status as a standard, guarantees the durability of the codings. The answer to the question "why SGML?" is therefore primarily pragmatic: no other solution is available.

In the period immediately following the adoption of SGML as the international standard (1986), there was a certain degree of hesitation in applying it. This was due, in particular, to the complexity of the language and the absence or the immaturity of the associated tools. Today, this hesitation has been completely swept away, and SGML is at the heart of all large-scale editing and

projects which involve extensive handling of documents. It is now generally agreed that SGML is an extremely powerful language that allows for innovation in creating, managing and manipulating electronic documents, and that the associated tools enable the development of SGML-based applications which are perfectly user-friendly and likely to offer substantial help to users. For a project like NERC, which involves a standardisation dimension, SGML thus indisputably constitutes the framework for developing a model of text representation.

We summarize below an essential contribution of the SGML standard which, whatever the structural model retained may be, fully justifies its adoption for tagging corpora in the framework of the NERC project.

The SGML standard allows the structure of documents to be defined, however elaborate this structure may be. It also allows, at the lowest level, the problem of the coding of characters to be resolved. This problem is not minor. We are aware of how difficult it can be to decode a photocomposition tape simply in order to read it. The problem of character coding still limits the exchange of electronic documents to this day (multiplicity of the sets of standardized characters, different conventions between the PC and Macintosh environments, etc.).

The SGML standard allows all the characters in a document to be encoded by working from the ISO 646 (ASCII) set of characters, which is the smallest set of characters common to all systems, and which, amongst other things, allows the various international networks to be traversed. Considerable standardisation work has been carried out within the SGML framework in order to compile an exhaustive register of sets of characters and to define groups of standardised entities. Today we must use this work to ensure that the texts coded by the NERC project are durable, at least at the level of character encoding[8].

3 The Text Encoding Initiative

The Text Encoding Initiative (TEI) is an international research project which aims to perfect and to make known the key principles for coding and exchanging documents in electronic form. Launched at the end of 1987, the project published and circulated an initial version of its results in 1990 (Sperberg and Burnard, 1991). These proposals were to be tested in fifteen major research projects. Publication of the final version is currently underway. The NERC teams have accepted, in principle, the feasibility of a project consonant with the TEI. We refer to (Belica, 1992, NERC-36), (Lopez Guzman, 1992, NERC-83), (Putter and Kruyt, 1992, NERC-91).

3.1 *General Characteristics of the TEI*

The project's main objective is to search for maximum consensus rather than to propose narrow viewpoints which are liable to lead to controversy. The solutions recommended by the TEI have the following characteristics in general:

8 It is to be noted that SGML word-processing programmes such as Author/Editor (SoftQuad) suggest in each of the Unix, Macintosh and PC environments, * SGML documents coded either with standardised entities, or with the set of characters used in that environment.

-a modular character: in most cases, one finds a minimal necessary kernel, to which one may add optional propositions; some of these are in fact strongly recommended.
-a flexible character which allows for alternatives: for all delicate questions, there is a (highly recommended) solution which is defined in terms of the possibilities available to implement it. The TEI has thus avoided the problem of getting everyone to agree without ever taking sides and, at the same time, has avoided constructing too rigid a framework.
-an extensible and modifiable character: given a certain number of rules, it is possible to stretch or modify the defined solutions. This controlled freedom is essential for the success of a project of such scope.

The direct consequence of such an approach is the need for an intermediary stage before the principles are put into effect. An application which conforms to the TEI will only retain some of the TEI proposals and will eventually develop those according to an axis that corresponds to its specific needs. A TEI-conformant application may therefore be extremely simple.

3.2 *Conformity with the TEI: Definition*

A document is said to conform to the TEI if it adheres to the following principles:

-It includes the SGML declaration and TEI declarations of the type of SGML document modified, if need be, in accordance with the rules given by the TEI.
-It includes relevant documentation for any non-TEI elements used.
-It contains the relevant elements of the header, as defined by the TEI.
-It uses the ISO 646 set of characters as its exchange format.

3.3 *Conformity with the TEI in the NERC project framework*

It is difficult to make recommendations which conform to the TEI given that this study was carried out before the publication of the TEI phase 2 chapters. In particular, since the chapters concerning the corpus have not yet appeared, we have been obliged to base our work on the provisional conclusions of phase 1 and on information given orally.

However, the fact that the rules of the TEI are not yet fixed has enabled us to engage in a most useful dialogue. We thank in particular Lou Burnard, co-editor of the TEI recommendations, for the interest he has shown in our comments. Likewise, it is important that the questions raised in this study be brought to the knowledge of the TEI working group before the final phase is completed.

In practice, our procedure for arriving at recommendations has been as follows:

-analysis of needs as regards the minimum coding of texts;
-comparison with the TEI proposals;
-definition of the elements constituting a minimal level of representation in terms that are, as far as possible, compatible with those of the TEI ;
-adaptation for our analysis of an experimental TEI DTD. This DTD, which was given to us by Lou Burnard, cannot under any circumstances be considered "a TEI DTD"; it merely adheres to its general principles and spirit. The TEI DTD had not been published at the time we carried out our study, but this experimental DTD enabled us to anticipate its publication in a constructive way;
-provision of an example of the use of this DTD, thus making it possible to illustrate how the TEI principles would be applied in practice.

This procedure calls for the following two observations. First, a revision stage for the DTD suggested in our final report will be necessary when the TEI and TEI DTD conclusions are finally published. Second, the example given in the final report does not have any particular value as a representative sample. It is an excerpt from theTreaty on European Unity (the Maastricht Treaty)[9], written in French, and is a very technical text. Its main function is to illustrate the principles defined by the TEI and to provide a solid base for the discussion of these principles.

4 Definition of a minimal level of text representation for European Corpora

The following proposals offer a working basis for the development of a minimal level of text representation for European corpora.

4.1 *Exchange and Processing*

The notion of a minimal level implies that it is possible to define a set of encoding rules that are recognised and accepted by everyone as the basis for encoding each of the corpora. For some applications, these rules certainly do not form an adequate coding level and they must consequently be supplemented by other rules.

Indeed, an essential distinction needs to be introduced when the problem of text encoding is considered. Is the objective of this encoding to allow for the exchange of texts or to ensure their processing? By processing we mean applications as varied as linguistic analysis, use for automatic translation, editing, etc. Clearly, the requirements and the level of complexity are not the same in each case.

The minimal level of representation that we are aiming for in the NERC project should enable the exchange of documents, that is:

-on the one hand retaining the capacity to reread and reprocess documents (non- proprietary encoding);

9 The electronic version was kindly provided to NERC by the EC Publication Office.

-and on the other hand transmitting a minimal level of information (other than the characters of the text themselves) to meet the requirements shared by everyone concerning the encoding of texts. It is assumed that this information is sufficiently important in most applications to warrant the effort needed to enter this information on the first occasion and to ensure that it is not lost when a text is transmitted.

Having defined the exchange format, it may be necessary to supplement the coding for particular applications.

4.2 *Internal Structure*

An electronic corpus is made up of a header which describes it as an electronic object, followed by the corpus itself. We will refer to the electronic corpus as the NERC corpus. The corpus itself is made up of a series of units: these are the electronic equivalents of either written elements or of transcriptions of spoken elements. In the discussion which follows, we are only interested in written elements. The notion of the header is examined below.

At the heart of the constitution of a corpus is the notion of "text": a sequence of words that go together, is clearly delimited and is identifiable. A corpus is the aggregation of texts and of groups of texts, where "group of texts" designates a collection of texts that have some feature in common, for example a collection of articles from an edition of a newspaper. Each text and each group of texts is preceded by a header that describes it as an electronic object. These headers are defined later in this document. A text and its header, or a group of texts and its header, constitute a corpus unit. More generally, we consider that a group of texts is made up of a collection of texts (with their headers) and/or of groups of texts (with their headers).

4.2.1 *Global Structure of a Text*

A text is made up of zero, one or two titles followed by the body of the text. If there is a hierarchical structure evident within the text (chapters, sections, etc.), this structure will be noted. We propose calling the different hierarchical levels "level i division". It seems reasonable to us to envisage five possible levels. If this proves insufficient, we propose a "division" element which may contain, in a recursive way, the same "division" element.

If there is no obvious hierarchical structure, we consider that the document is made up of standard text content as defined below in 3.2.2.

A level i division is made up of:
-an optional title,
-possibly followed by standard text content,
-possibly followed by a series of divisions at a lower level.

4.2.2 *Standard Text Content*

Within a division level, text content is very often organized in paragraphs or, more generally, in "textual units of the paragraph type". We suggest a minimal list of such units. Whenever they exist and are obviously identifiable, these elements should be marked.

-paragraphs;
-lists, that is any enumeration of items, ordered or not; more precisely, a list may be described in the following way:
  - an optional title,
- followed by a series of items, each item possibly preceded by an introducer (of type a), b), or 1.2...);
-notes; by this term we mean the insertion at a given point in the standard text of a portion of text of a different nature, whatever the position of this text may be on the printed page (footnote, in-text note, margin note, etc.);
-subtitles or legends for figures, diagrams, etc. This category includes all text fragments which function as titles, with the exception of division titles of the hierarchical structure and clearly identified SGML elements;
-lines in poetry;
-quotations or, more generally, all imported texts (including texts in foreign languages);
-numerical data or mathematical formulae if they can be captured.

Within these elements, some textual units at the lowest level should be marked again. We call these "textual units of the character type". They are:

-highlighting; one can append an attribute to this element to account for the way in which this highlighting is translated (italics, bold,...)
-abbreviations and acronyms;
-proper names, place names, book titles;
-elements of the `list-introducing' type present within the text.

4.3 *List of the Names of Elements*

In summary, the list of the names of elements retained is as follows (we refer to the DTD in the Final Report of Work Package 3 (Lafon and Vignaud, 1992, NERC-150)).

abr abbreviation or acronym

anynamename, whether this be a proper name, a toponym, or the title of a book; see the comment below on the use of this element.

corpuscorpus; the highest level of the DTD

div division allowing for as much interleaving of levels of divisions as required.

div1,5 divisions from level 1 to 5

flheadfloating headings: floating titles or subtitles; this element serves to mark all the portions in a text that have the role of titles without being titles of a hierarchical level or of a clearly identified element

group group of texts

head title of text, of group of texts, of division, of poem, of list

header header (not described in the DTD)

hihighlighting; an attribute allows for the type of highlighting (bold, italics, etc.) to be specified

item list item

l poem line

labelinitiator of list item; see below in 5.3 for comments on the use of this element

listlist or enumeration; see below in 5.3 for comments on the use of this element

nerccorpus as an electronic object (i.e. the corpus plus its header)

noteany kind of note (footnote, margin note, etc.)

numnumerical data; this element should not be used to tag a number employed as reference for a logical element (article 2, chapter 4, etc)

p paragraph (on the use of this element, see Lafon and Vignaud, 1992, NERC-150)

ptrpointer; this element allows for another location within the same document to be pointed to. The target of the pointer is marked with the help of identifiers appended to all the SGML elements of structure

quote any kind of imported text

subsubscript; this element serves to mark characters that are take the form of suffixes

supsuperscript; this element serves to mark characters which take the form of exponents

texta text from the corpus

unita unit of the corpus as an electronic object, i.e. a text or a group of texts preceded by its header

4.4 *The Headers*

The header is an obligatory element of a document coded in SGML in conformity with the TEI. The aim of this element is to supply the necessary information about the electronic document, its source text, its coding and its revisions. The header is to the electronic document what the title page is to the printed book. The TEI defines a basic header plus sets of additional elements which allow the creation of more complex headers.

In the case of corpora, we need to define a header for each corpus. This header has to bring together all the information valid for all the texts of the corpus, a header for each text that brings together particular information relative to this text and a header for each group of texts that identifies this group and brings together the information valid for all the texts in the group.

According to the TEI, a (basic) header is made up of four parts:
-Description of the file: this contains a full bibliographic description of the electronic document (obligatory)
-Description of the coding: this describes the conversion of the source document to the electronic document (optional)
-Document profile: this contains information concerning the context and the classification of the document (optional)
-History of revisions: this allows the tracing of revisions carried out on the electronic document (optional)

4.4.1 *The Corpus Header*

At a minimal level of representation, the header of a corpus must contain the following obligatory elements:
-The description of the electronic corpus; within this element the following sub-elements should be included:
- the title of the electronic corpus
- mention of those in charge (names of the persons in charge of compiling the corpus)
- the size of the electronic corpus.
-The description of the coding; within this element the following sub-elements should be included:
- a description of the principles and practices guiding the coding of the corpus
- and, optionally, a "description of the project" element, mentioning the aim of the creation of the corpus.
-The profile description; within this element the following sub-element should be included:
- creation: this element contains all information (of a non-bibliographic nature) relative to the constitution of the corpus.

4.4.2 *The Header of a Group of Texts*

The header of a group of texts must contain the following obligatory elements:
-A description of the group of texts: within this element the following sub-elements should be included:
  - identification
- description of the source: this element brings together all bibliographic information common to all the texts in the group.
-Optionally, a "description of the encoding" element, which contains the following sub-element:
- description of the project mentioning the aim of the constitution of the group of texts.
-The profile description; within this element the following sub-element should be included:
- creation: this element contains all information (of a non-bibliographical nature) relative to the constitution of the corpus.

4.4.3 *The Header of a Text*

The header of a text should contain the following obligatory elements:
-The description of the text: within this element the following sub-elements should be included:
  - identification
- description of the source: this element brings together all bibliographic information about the source text, and more precisely the following obligatory elements: mention of the title, the author, the title of the publication series if there is one, date of publication and publisher.
-The profile description; within this element the following sub-element should be included:
- creation: this element contains all information (of a non-bibliographic nature) relative to the text and in particular a sub-element "date of creation" which contains, in a standardised form, the date of the creation of the text; this last element is used only if the information available is reliable.

4.5 ***Disambiguation of Punctuation; Accented Capitals***

Here we deal with problems that are specific to the French language, but can be easily generalized to other languages.

4.5.1 *Disambiguation of Punctuation*

A number of problems related to punctuation become evident when the corpora are exploited. They stem from the fact that the same punctuation mark has different meanings depending on the context in which it is used. For example, a full stop may mark the end of a sentence or indicate an abbreviation. A comma may be a punctuation mark or may, in numerical data, indicate a decimal. The dash and apostrophe may or may not group words together.

One way of disambiguating these uses is to code punctuation marks in several different ways; onecould keep the use of the apostrophe to key in "il l'a dit" (he said so) for instance, and define an

entity `&apostophe;' to key in the apostrophe grouping "aujourd'hui" (today) which would be coded:

aujourd&apostrophe;hui

A second possible solution is to mark the contexts in which the use of the punctuation mark varies. Thus, in marking abbreviations, one would distinguish the full stop at the end of the sentence from the full stop after "Mr." which is keyed in as:

<abbreviation>Mr.</abbreviation>

We recommend marking contexts when the context marker might be interesting for other areas of study, that is for numerical data and abbreviations (disambiguation of the comma and the full stop). For dashes and apostrophes we recommend the use of entities to distinguish the two types of use.

The following question remains to be answered: does this type of marking belong to a minimal level of text representation or should it be carried out at a later stage of file processing? In the case of texts retrieved in an electronic form (texts coming from networks or photocomposition tapes), this question is formal. Indeed, a certain number of retroconversion programmes will bring these texts to the desired level of representation. The modules serving to recognise abbreviations, numerical data, dashes and non-delimiting apostrophes may be included in these programmes or, alternatively, restricted to the data processing sequence.

In the case of texts keyed in by hand, it may be better, from the point of view of quality, to mark at the keying-in stage the numerical contexts and the abbreviations, considering them thus as a minimal level of text representation.

The recognition of dashes and non-delimiting apostrophes must be excluded from this minimal level and possibly become part of subsequent file processing.

To eliminate the extra expense linked to the development of several programmes for the disambiguation of punctuation in different research departments, one could envisage the concerted development of a single programme (for each language) for disambiguating punctuation; this programme would be applied to SGML texts coded in conformity with the minimal level of text representation defined by the NERC project.

4.5.2 *Accented Capitals*

The increasingly rare use of accented capitals poses problems for the exploitation of corpora. As with punctuation, it would be interesting to develop a single reference programme allowing these accented capitals to be restored. This programme would be applied to SGML texts coded in line with the NERC recommendations.

It should be remembered that standardised entities exist for designating accented capitals; this is the set of public entities ISO1at1 summoned by the SGML declaration:

<ENTITY % ISO1at1 PUBLIC
"ISO 8879-1986//ENTITIES Added Latin 1//EN">

5 Conclusion and Recommendations

Our study does not claim to provide a definitive answer to the problem of encoding texts in the context of the exchange of corpora. It may however offer an important step forward in the effort that must be made towards a true standardisation of recordings and towards putting an end to the heterogeneity currently observed in this field. This heterogeneity makes the exchange and shared use of textual corpora almost impossible.

The various names of elements we have defined do not particularly allow for a text to be encoded in a one-to-one way. They do however guarantee the encoding of textual characteristics that we consider essential. When it is a question of ensuring the exchange of texts, this level may be considered sufficient. A one-to-one relation remains an ideal to be aimed for, but it is not certain that this is necesary in an exchange process, the essential thing being to ensure good inter-comprehension in the analysis of recordings. It is to be noted that a concern for one-to-one mappings is not apparent in the TEI recommendations.

This study highlights a research area and sets out a number of objectives which remain to be achieved. On the one hand, it will be necessary to go back to the formalisations suggested when the TEI phase 2 research results appeared, and in particular the section concerning corpora. On the other hand, if a corpus is to be constructed on the basis of these recommendations, it will almost certainly be necessary to go back over the elements to be marked on account of a better specification of the envisaged usage.

Adopt the SGML standard.

Conform with the TEI, but choose a restricted marking sub-set, which seems to be indispensable for the majority of applications and therefore likely to suit the greatest number of people in the community of users.

Choose markers that leave very little open to interpretation and that are based, as much as possible, on indisputable formal marks in the text, to enable the greater part of the tagging to be automated.

Write a manual for electronic text-records in all European languages.

Relevant NERC Papers

Belica C. (1992): "The TEI proposal and its feasibility. An evaluation in respect to the IDS Corpus philosophy", IDS Mannheim, Working Paper, NERC-36.

Lafon P., Vignaud D. (1992): "Répresentation des textes écrites", INALF Paris and AIS Berger-Levrault, Technical Report, NERC-150.

Lopez-Guzman J.M. (1992): "Acquisition and reusability of material for corpus generation", University of Malaga, Working Paper, NERC-83.

Putter E., Kruyt J.G. (1992): "Evaluation TEI-Guidelines draft version 1.1", INL Leiden, Working Paper, NERC-91.

B. Spoken Language

1 Introduction

The conventions for putting representations of the spoken language into machine readable form are not as advanced as those for the written language, and they are inherently less stable. To print or type a text is a process which, among other things, causes the language expression to conform to a set of rigorous conventions, tending towards standardisation. Writing was used from its introduction as a means of making a lasting record of what would otherwise have to be remembered from hearing its spoken form. But a tape recording of a conversation is just a physical record of the sound wave, lacking analysis, and no conventions have been introduced. Whereas it is an everyday activity for people to type or print, or to be aware of others doing it, it is relatively unusual to encounter people engaged in transcription. Hence there is no universal standard that people can refer to on occasions when transcribing is necessary.

The lack of similarity between speech and writing as physical events makes it extremely difficult to transfer information from one medium to another, except for the alphabetic code. Anything not coded is a problem. Intonation and pausing, pace, voice quality and stressing all need special attention, and speaker changes, interruptions, overlaps, hesitations and the like are not features of the written form. The conventions of the playscript make a few concessions of layout in order to separate language to be spoken from the rest, and to indicate speaker turns.

Transcription has been found necessary for a large number of reasons, but in almost all cases the motivation for making the transcription has a strong effect on the transcript. The selection of what to transcribe is one major variable, the sound wave being so rich and varied. Even in the small and specialised field of academic research there is very little common ground between one transcription venture and another, and as a result transcribed material tends not to be reusable.

For corpus inclusion there are no standards and very little experience. A number of transcription systems exist, and the IPA (International Phonetics Association) alphabet is a world standard for an unambiguous rendering of the sound wave. The Speech Community (see section 3 below) uses IPA as a rough guide to what is on the sound wave, but other users find it very difficult to read, and prefer an orthographic transcription, which is highly conventionalised and is very close to the ordinary shape of the written language.

The questions posed in this chapter are therefore:

- What features of the sound wave apart from the alphabetic codes should be recommended for documents which are destined to be included in a corpus?

- What are the best conventions for representing these features?

The Text Encoding Initiative has been studying possible answers to these questions, lagging a little behind its parallel work with the written medium. A draft set of recommendations (Sperberg and Burnard, 1991) was made available to NERC, and has stimulated a detailed critical response. TEI adds a further and very important question to the two above:

•What features of a speech event other than the sound wave is it necessary to encode?

Spoken language transcription is an area which is likely to benefit substantially from technological advances in the new decade. The digitisation of the recordings makes the soundwave tractable to a digital computer. Work on speech recognition is already contributing to the alignment of the transcription with the sound wave, and may eventually automate the process of transcription (though most commentators think that the flexibility needed for the final step is still a long way off). Routine transcription for legal etc. reasons expands with the expansion of spoken media, and the data retrieval power of modest computers.

The proposals in this chapter and its annexes (i.e. the NERC Working Papers and Technical Annexes listed at the end of this chapter) should therefore be seen as sensitive to the time of writing, and if adopted should be kept up to date.

2 Organisation of the Chapter

A collection has been made in the NERC Working Papers listed below (references at the end of this chapter) of leading descriptions and manuals for the transcription of spoken language:

Scheiter, 1992, NERC-43
J.P. French, 1992, NERC-50
J.P. French, 1992, NERC-47
Anderson, 1992, NERC-163
Candel, 1992, NERC-157
Kirk, 1992, NERC-158
Villena Ponsoda, 1992, NERC-141
de Jong, 1992, NERC-159
Edwards, 1992, NERC-164

It is clear that their purposes are varied and not always specifically to do with corpora. While the surface conventions differ a great deal, and the emphasis of each is slanted according to the job for which it is designed, there is not ultimately a lot of disagreement about the answers to the first question - what features should be encoded.
The above Working Papers and Technical Annexes report current practice.

The descriptions range from fairly idealistic treatments of what would qualify for inclusion if resources were not to be taken into account; through some reports guided by experience; to suggestions made on the basis of massive experience. It should be stressed that resources are a key issue in this discussion. As the costings show, the capture of informal spoken language is extremely labour intensive, requiring anyone recommending procedures to ensure that they are cost effective.

The existing manuals etc. are agreed in demanding a high standard of accuracy and detail in the transcriptions. The linguistic study of speech may be fairly superficial as compared with the work of the Speech Community, but it still requires transcription conventions that are prohibitively

expensive unless applied only to small sample texts. Users have in the past been less than complimentary about transcribed material which is limited in its detail, and those who have corpus-sized needs will reject transcriptions which confine the researcher within a specific and narrow view of the spoken language. It is obviously a difficult task to reconcile the expectations of a wide variety of users with the practical limitations of budgets.
Section 3 then considers relations with the Speech Community (Payne, 1992, NERC-132). A few years ago the views and needs of phoneticians were seen to be of a different order from those of descriptive linguists. Tiny samples, production of speech under experimental conditions, exacting standards of recording and the full scientific paraphernalia for scientific enquiry were the norms of speech research. A visiting corpus linguist would find this environment very different from his or her own. Corpus linguists would be trying to maximise the amount of data that could be acquired within a budget and a timescale, and had no need for high technology for acoustic analysis.

Although the speech research tradition continues to study the sound wave in great detail, the differences in style between them and the corpus linguists are a lot less marked now. There is considerable convergence of interest to report. The priorities remain, but recognition of the value of extended and authentic data is increasing among the Speech Community, and the results of technical breakthroughs in acoustic research are potentially of great value to the corpus linguist. The two groups are now engaged in joint ventures (e.g. the SALT Club in UK - Speech and Language Technology), and discussions are expected to continue in other parts of Europe after NERC in order to maintain the convergence.

Section 4 deals with TEI, introducing the relevant annexes and making some general points about the nature of standardisation at the present time. Finally, in section 5, we present our own proposals arising from this survey and study.

3 Speech Community

This section incorporates discussions with Roger Moore, of DRA Malvern, a leading expert on the study of the spoken language. Although Moore does not necessarily agree with the recommendations of this report, his meetings with NERC representatives have been helpful in charting the old and new relationships between the two groups of scholars.

RECORDING CONVENTIONS. The cassette recorder is ubiquitous and is used even in adverse conditions. Moore reports that he is about to publish a Guide to recording conventions, and it is recommended that corpus creators follow the recommendations as far as practicability allows. If possible a detailed response by NERC to the Guide will be included in the final version of this paper. For any level of transcription, a high quality recording improves the efficiency of the transcription process: for anything beyond level Two (see section 5 below) the quality must be well above domestic.

SOUND AND TRANSCRIPTION ALIGNMENT. Dr. Moore reports that there exist automatic procedures for aligning orthographic transcriptions of English with graphic representations of the

sound wave. This facility is not part of the SAM workstation specification (ESPRIT Project 2589), but is provided in research laboratories, probably for several languages. It is strongly recommended that such a facility be made available as part of NERC specifications, either by asking research laboratories to provide a service, or by funding the development of a portable software package to partners in a corpus network.

CLASSIFICATION OF SPOKEN LANGUAGE. This chapter concerns only one sector of the language delivered in the spoken medium; impromptu, unrehearsed, unscripted, informal conversations in natural settings, not intended to become part of a permanent record. Because of the linguistic, sociolinguistic and psychological interest in this data, it is often confused with spoken language as a whole. The issue of classification is dealt with in Chapter 2 (Table 3 in the Appendix), where it is shown that there is a mass of public spoken language, formal spoken language, radio, television, film etc, spoken language, and many other varieties. The baseline conventions (below) apply throughout, but in many cases it may not be possible to assemble all the necessary data. A local unscripted radio programme, for example, may exist only in orthographic transcription of around level Two. Descriptions of corpora should be encouraged to be explicit about the forms in which spoken data is held. For many types of research, particularly in the Speech Community, spoken language is elicited by a variety of techniques. The discoursal quality of the language is unimportant, because the research may concern principally some feature of articulation. Usually the language elicited is in very short stretches, and although a large quantity of these is regarded as constituting a corpus, there is little risk of confusion here between a corpus of unrestricted naturally occurring data and the unashamedly artificial recordings for speech research.

Recently, however, speech research has moved into areas where larger quantities of elicited spoken language are recorded, and collections of these recordings are being circulated for research purposes. Again it is important for this kind of spoken language to be kept distinct from unelicited speech, and conventions of classification should be instituted and recognised.

The issue of authenticity, or naturalness, of spoken language has been controversial in linguistics and applied linguistics for many years, and data collectors range from purist to interventionist. Little is known about the characteristics of natural spoken discourse (Warren, 1993) but native speakers are quick to notice the invented example or the scripted dialogue presented as if spontaneous. At present the policy we recommend is to keep apart any spoken data which does not arise in the ordinary process of communication.

NERC recommendations will be fed into EAGLES, which will give further consideration to the establishment of common practice in language research involving computers. There are clearly some matters of terminology and definition to be sorted out, notably "corpus", "spoken language", "speech", and "authentic". There are priorities expressed by the community of discourse analysts, requiring spoken data that has been collected with the minimum of intervention, which should be respected without restraining the professional needs of acoustic phonetics. The establishment of a clear and detailed vocabulary for the description of spoken data is a target for EAGLES.

4 The Text Encoding Initiative

The conventions set out in the TEI paper (Johansson et al., 1992, Burnard, 1992a, Burnard, 1992b) are broadly acceptable provided that compatibility is the goal and not conformity (see below). This paper arrived when the Work Packages deliberations were at an advanced stage, and the second Report of TEI (called P2) was circulated too late to be taken into account in detail.

From the perspective of TEI, a document such as a transcription consists of two parts -a header, which gives details of the origin, provenance classification and circumstances of the document, and the text of the document with approved annotations. In corpus technology there is an established practice of separating the text from any header-type material, so that the integrity of the stream of natural language is not corrupted. It is therefore important that TEI headers can be kept to minimum identifiers, usually cross-references to databases of header material.

The question of compatibility of such databases with TEI remains, and this report does not attempt to address the issue. In practice the corpus providers await the availability of software which will enable TEI conventions to be established and maintained. There is likely to be very little disagreement about what information should be retained.

The question of annotations to alphabetically transcribed text is one of major current significance, and (Payne, 1992, NERC-122, Cauldwell, 1992, NERC-106, Belica, 1992, NERC-36) contain a critique of TEI proposals. Corpus work brings in considerations of the cost of large scale application of any conventions, and both the features to be encoded and the codes for the features must come under the closest scrutiny. TEI is more concerned with the codes to be used rather than the features to be selected, and this is where the distinction between compatibility and conformity is raised.

TEI conventions are concerned with efficient machine representation of a document, and less with the labour and needs of the human beings who are employed to put the documents into machine readable form. There are many cases of clashes between what a human can cope with cost-effectively, and what is most convenient for the computer. In most of these, the human and machine preferences are totally compatible, and are just variations in style; simple software can convert one to another when necessary.

A TEI-coded document is unreadable by human beings unless specially trained, and it is not intended to be read by humans. The codes disappear when they are converted to typographical or layout settings, and the document appears normal. In the same way, software is needed urgently to allow a transcriber or editor to use human-style conventions, converting them automatically into TEI representations. Only with this is it at all likely that corpus providers will adopt TEI conventions, on the simple grounds of unrealistic costs.

This is the meaning of compatibility. Those who wish to conform should be aided by software which converts compatibility to conformity.

To this end, (Payne, 1992, NERC-122) compares the TEI proposals with the coding conventions set out in the paper (J.P. French, 1991, NERC-47). They have been developed for English over many years and are basically compatible with TEI proposals. A summary of the papers dealing with the conventions established by J P French and NERC proposals drawn from them, follows in the next section. The papers are quoted in full at the end of the section for convenience.

5 Transcription Conventions

TRANSCRIPTION LEVELS The paper by J P French (J.P. French, 1992, NERC-50), see full text in Appendix A below), commissioned for this study, sets out four levels of detail of transcription.

Level One is an orthographic representation with minimal punctuation and no interactional information. It is the quickest and cheapest and is useful for basic word-frequency information, concordancing and collocation. The earliest known computer-held corpus of spoken English (1963) was transcribed in this way (Sinclair et al., 1970; Jones and Sinclair, 1972).

Level Two is a much enhanced orthographic transcription, as set out in (J.P. French, 1992, NERC-47). It contains basic information about speaker identity, speaker change, overlaps, laughs, etc. It is intended to be suitable for most linguistic studies that do not require intonational information: it is still cheap enough to be provided for corpora of several millions of words. This level would achieve reasonable compatibility with TEI conventions.

Level Three contains identification of tone boundaries and tonic syllables; also the precise analysis of overlap onset and resolution is provided. These enhancements make it suitable for more in-depth study of sociolinguistic issues, and for the description of intonation in discourse. The cost, however, goes up sharply compared with level Two. Both transcribing and checking need to be done by trained phoneticians, and short passages need to be replayed many times over, increasing the time taken. At this point also the recording quality becomes a significant factor, and many recordings made in natural settings cannot be transcribed at this level.

Level Four offers further detail on intonation, including the identification of head syllables and of tone; for English five basic tones. It includes a phonemic as well as an orthographic transcription. The orthographic version is aligned with a graph showing the waveform and pitch patterning and the phonemic version is aligned with the output of a sound spectrograph. This level of analysis is suitable for many studies of speech of discourse, but is impracticably expensive to provide for the user community in general. It is presented as a possible target for particular studies and applications, which could gradually accumulate a small set of text samples to this high professional level. Experience suggests, however, that the needs of researchers in this area vary so much that level Four may be restrictive as a standard.

THE SOUNDWAVE It is now a matter of routine to make a digitised version of a recording of speech, and this has many advantages over the analogue version. In particular, it is much less prone to deterioration through time or in the act of copying, and it is immediately capable of being processed by a digital computer. It is a fundamental recommendation of this Work Package that a digitised version of every sample of recorded speech is included as a component of this corpus. See (Cauldwell, 1992, NERC-48) for an example of the flexibility that is made possible.

BASELINE CONVENTIONS Balancing the costs involved with the results achievable, the variety of needs across the spectrum of users and the general needs of the user community with the research

opportunities available to a small but important group of scholars, NERC recommends the following set of baseline conventions for the computer representation of spoken language corpora:

(a)orthographic transcription at Level Two (see full text of J.P. French, 1992, NERC-47), achieving reasonable compatability with TEI conventions;
(b)an accompanying digitised representation of the sound wave;
(c)an automatic alignment of (a) and (b).

These conventions are cheap enough to apply to large corpora, and have all the information necessary for sensitive and detailed study of almost any aspect of the spoken word. With a synthesiser driven by (b), the sound recording can be heard while the transcription and waveform are on the computer screen.

The means of presentation of the transcription to the user should include opportunities for further analysis and notes, kept separate from the transcription itself and where possible following the conventions of levels Three and Four.

APPENDIX A

1 Introduction: A Multi-Level System

The present system of transcription involves a series of levels. Transcripts produced using the conventions of the lowest level, Level One, are the least detailed, consisting only of an orthographic representation of the words spoken, together with the barest minimum of punctuation. Level One conventions might be seen as appropriate to a research project where the main goals were ones of concordance, word-frequency documentation, and so on.

At Level Four, the highest level, data is not only represented orthographically, but carries with it information about the phonetic and acoustic properties of the speech signal. Because Level Four also contains a variety of phonological and other linguistic codings, a transcript produced according to the conventions of this level makes visible the relationship between the 'raw' acoustic/phonetic signal and the transcriber's linguistic interpretations.

Level Four transcription might be seen as appropriate to a project concerned with relatively fine-grained analysis of small amounts of conversational data. Levels intermediate between One and Four ascend in order of detail and complexity. Each Level is further explained and illustrated below.

2 Conventions at Each of the Levels

Level One

A Level One transcript is made in accordance with the following conventions:

(i)*Orthographic representations*

The words spoken are represented in accordance with standard orthographic conventions.

The only contractions used are those accepted as standard in the Oxford English Dictionary (*it's, isn't,* and so on).

(ii)*Punctuation*

Sentence boundaries are marked by a full stop and capital letter. Commas are not used within sentences. It is recognised that in the delimitation of sentences will in some cases be problematic and decisions on this may be somewhat arbitrary.

Direct quoted speech or quotations from written texts are placed in single quotation marks.

Apostrophes are used in accordance with standard conventions in possesives and in contractions (*John's, can't*).

(iii)*Interactional information*

A transcript at this level contains no interactional information. It is set out as continuous text; change of speaker is not marked.

Uses of Level One transcription: concordance study, establishment of word-frequencies.

Level Two

The transcription conventions at this level are set out in JP French, 1991, NERC-47. It is simple enough to be done at acceptable speeds, and allows spoken and written texts to be examined on a similar basis. Major structural features are indicated, but not intonation segments.

Uses of Level Two transcription. This is established and recommended for large volume routine transcription of conversation for corpora.

Level Three

A Level Three transcript contains all the information included at Level Two but with the following extra intonational and interactional information:

(i)*Intonational information*

Tone unit boundaries are marked (/).

Tonic syllables are marked (capitals).

Eg.:

/I don't suppose you've SEEN one yet./ Although it's an old IDEA/it hasn't been on the MARket very long./

(ii)*Interactional information*

Where talk occurs in overlap, the precise points of overlap onset (*) and resolution ($) are identified.

Eg.:

<MO1> /The thing IS/ we're not going to let them get aWAY with it *this time$./

<MO2> /*I can't recall$ their having tried it on beFORE/.

Uses of Level Three transcription: as with Level Two but may also be used for study of intonation in discourse and of various sociolinguistic issues (eg., male-female differences in interrupting/yielding to interruptions).

Level Four

Transcripts produced according to Level Four conventions include all information encoded at Level Three. However, they also contain additional intonational codings as well as acoustic and phonetic information.

(i)*Intonational information*

In addition to marking tone unit boundaries and tonicity in the manner of a Level Three transcript, Level Four transcripts also mark head syllables (underlined) and record tonality.

Five tones are recognised:

Fall `
Rise'
Fall-rise∨
Rise-fall∧
Level.

Eg.:

<MO1> /The thing ˬIS/ we're not going to let them get a·WAY with it this time./

<MO2> /I can't recall their having tried it on be^FORE/.

A computer generated graph showing the waveform and pitch patterning of the utterance is also included.

(ii)*Segmental-phonological information*

Level Four Transcripts also include a computer-generated 5 Khz sound spectrogram (156 Hz Kaiser window) together with a phonemic representation of the sound segments.

A further possibility at Level Four, not developed here, would involve grammatical classification and tagging of words and constructions.

Uses of Level Four Conventions: As well as providing a basis for the study of various phonological and linguistic issues in their own right, transcripts of this type allow one an understanding of the relationship of the essentially phonological categorisations of tones and segments to the raw acoustic data.

APPENDIX B

List of Codes and Macros

<M01>Alt MFirst male speaker
<M02>Alt BSecond male speaker
<M03> <M10> etc (change Alt B)Third, tenth, etc male speaker
<F01>Alt FFirst female speaker
<F02>Alt RSecond female speaker
<F03> <F10> etc(change Alt R)Third, tenth, etc female speaker

<M0X> Alt JUnidentified male speaker
<F0X> Alt TUnidentified female speaker
<X0X> Alt XNo individual speaker or gender identification
<ZG1> Alt GStart of guess at unclear word or utterance
<ZG0>Alt HEnd of guess at unclear word or utterance
<PN1> Alt QStart of guess at unclear proper noun
<PN0>Alt W End of guess at unclear proper noun
<ZGY>Alt UWhole unintelligible word or utterance
<ZF1>Alt ZStart of repetition
<ZF0>Alt SEnd of repetition

<ZZ1>Alt IStart of comment from transcriber
<ZZ0>Alt OEnd of comment from transcriber

“ Alt PStart of quote from written source
” End of quote from written source

MX (plural **MXs**)Replaces male name
FX (plural **FXs**)Replaces female name
XX (plural **XXs**)Replaces name where sex is uncertain, or surname used to refer to a group of people
[pause]Alt VUnexpected pause
[laughs]Alt COne person laughs
[laughter]More than one person laughs simultaneously
[coughs]One person coughs
[coughing]More than one person coughs simultaneously

[claps]One person applauds
[applause]Applause from more than one person
[jingle]Replaces advertising jingle

\+ replaces missing portion of broken-off word

Use of Codes

In general

All codes (except the punctuation-type codes) either take a space on either side or are placed on a new line. If a code is immediately followed or preceded by a punctuation mark, there must still be a space:

<FOX> But didn't you <ZGY>?
<FOX> Oh God that's right. [pause]

All codes in <diamond brackets> must have exactly five characters, including the brackets. Everything inside <diamond brackets> must be upper-case.

Speaker codes

Speaker codes always start a new line and are followed by one tab:

<M01>(text)
<XOX>(text)
<F15>(text)

Guess codes

Anything you're not sure about must be marked using guess codes.

Use **<PN1> <PN0>** for guesses at proper nouns (mainly names and places), including guesses at the spelling.

Use **<ZG1> <ZG0>** for guesses about any other word or phrase (including guesses at the spelling).

Use **<ZGY>** where anything is so unclear as to be unintelligible.

If a name, word or phrase of which you're unsure occurs frequently in a job, mark the first occurrence using the above guess codes, and immediately afterwards place a transcriber's comment **<ZZ1> spelt thus throughout <ZZ0>**. Thereafter, do not mark the name/word/phrase, but make sure you spell it consistently; we'll decide what to do with it in the office.

False start codes

"False start" is a slightly misleading description, because changes of tack and self-corrections aren't included. **<ZF1>** and **<ZF0>** are only used to enclose exact repetitions of whole or part words or phrases.

<ZF1> is placed before the first word of the sequence, and **<ZF0>** is placed before the last repetition (the "real start").

Er, erm, and anything inside [square brackets] do not in themselves take false start codes or affect false starts among which they occur. If one of these items occurs before the false start begins, place it before the <ZF1> code. If it occurs before the "real start", place it inside the codes (i.e. before the <ZF0> code). These are false starts:

<ZF1> I I <ZF0> I said
<ZF1> I <ZF0> I'm leaving
He said <ZF1> th the+ <ZF0> these would be okay
<ZF1> I don't I I [laughs] <ZF0> I don't think so
We {pause] <ZF1> did+ er <ZF0> didn't know
These aren't false starts:
I don't I can't do that
He's he is on his way
She is she's nearly here
This pra this product has been on the market
So she came er er erm [pause] and er saw me
Long and complicated sequences - if the whole sequence consists of repetitions or parts of the final "real start", then lump it altogether under one false start code. Otherwise, break it up into as many separate false starts as are necessary. (False starts can be placed immediately side-by-side):
<ZF1> It w+ it it it wasn't [pause] <ZF0> it wasn't me
<ZF1> It w+ it <ZF0> it d+ <ZF1> it <ZF0> it wasn't me
<ZF1> It <ZF0> it <ZF1> wasn't <ZF0> wasn't me

Incomplete words
Only type the existing portion if a speaker doesn't say the whole of a word. If the resulting broken-off portion could be mistaken for a complete word in its own right, or if it is only a single letter, then replace the missing portion with a + sign. Usually the broken-off portion is the last part of the word, and so almost all of the pluses will be on the end of the word; but watch out for the few cases where a speaker misses the beginning of a word, in which case the plus is on the beginning of the remaining portion.
When a contracted word like **I'm he'll they're** is broken off at the apostrophe, don't put a plus unless the broken-off portion is pronounced differently as a result of being part of a contraction (as in the case of **can/can't/don/don't**).
NB. Incomplete words are included in false starts.
It's not very unim important(no plus because "unim" isn't a word)
<ZF1> the+ <ZF0> these are(a plus because "the" is a word **mine** in its own right)
I c+ w+ I don't know(pluses because "c" and "w" are isolated letters)
<ZF1> I I <ZF0> I'm here(no pluses because "I" is counted as a whole word)
I can+ don't want to do it(a plus because the "can" is really the broken-off word can't")
It's ecolo [pause] +logically("ecolo" isn't a word but a sound "logically" is)

Transcriber comment codes
Use transcriber comment codes **<ZZ1> <ZZ0>** to enclose any messages and information that you feel are necessary arising from your transcription:
<ZZ1> background noise - tape count 332-463 not transcribed
<ZZ0>
<ZZ1> impossible to tell speakers apart <ZZ0>

Non-verbal aspects:
Non-verbal occurrences which might have a bearing on speech, such as coughs, other external noises and so on, are all enclosed in [square brackets]. If the [laughs], [coughs] etc applies to a different speaker, put it on a new line with a speaker code.
<M0X> I think [pause] I'm going to
<X0X> [coughs]
<M0X> shut up now.
<FOX> As you can [beep] hear, the geiger counter's recording [series of beeps] quite a high level of background radiation. [very fast beeps] Hm. Which seems to be getting higher all the time. Hang on.
Put in [pause]s wherever they stand out - wherever you wouldn't expect to hear one.

Pronunciation:
What to do with variant pronunciations:
Where a distinction is made between variant pronunciations (for instance, where people are talking about regional accents or discussing foreign words) instead of trying to make up a spelling for the word(s) in question type them with the normal spelling and put a transcriber's comment immediately after to indicate the pronunciation.
<M0X> I always say grass <ZZ1> short A <ZZ1> and laugh <ZZ1> short A <ZZ0> but my wife says grass <ZZ1> long A <ZZ0> and pass <ZZ1> long A <ZZ0>.

Scripts
Where a tape includes scripted material - such as the adverts or news on a radio show - we only need to type it out once. The first time a script occurs, put a transcriber's comment
<ZZ1> script starts <ZZ0>
on a new line, then type the script beginning on a fresh line. After the script, put another transcriber's comment
<ZZ1> script ends <ZZ0>
again on a fresh line. (Where several adverts or news items run one after another, just put the comments around the whole block rather than putting separate ones around each item.)

Text Layout

In general
Don't type anything that isn't there.
Do include everything that is there.
Follow the speakers' grammar, even if they don't use well-formed sentences or "good" grammar.

File identification
Start each file with the job number in {curly brackets}. The job number will always be a ten digits preceded by a capital S
{S0000000014}

For excessively long jobs, where the whole file will not fit onto one disk, split the file at the beginning of each new tape. Number the portions A, B, C and so on using transcriber's comment codes thus:
{S0000000014}
<ZZ1> section A - tape 11 <ZZ0>
Where a job covers more than one side of a tape, start the new side of the tape on a fresh line.

Speakers
Give each new speaker a new line, whether or not they can be positively identified or their speech heard.
Numbering begins at <M01> and <F01>, with each successive speaker of each gender receiving a new number in sequence.
(The only case in which the above numbering system does not apply is in jobs where one speaker is present as the centre of attention throughout (such as the DJ in a radio phone-in) - in which case, that speaker is <M01> or <F01> whether or not they speak first.)

Punctuation

Punctuation marks in use:
Only full stops, hyphens, open/close quotes, apostrophes, and question marks are allowed. (No commas, semi-colons, colons, speech marks, exclamation marks, rows of dots, accents etc.)
Full stops - only used to mark sentence boundaries.
Hyphens - only used for hyphenation between connected words.
Quotes - only used to open and close quotes from written sources.
Question markers - only used to mark functional questions.
Apostrophes - only to be used for possessives and contractions.
(See individual sections on sentences, hyphenation, quotes, contractions and questions)

Sentences
Full stop, one space and capital letter to mark a new sentence.
People do not always speak in complete or correct grammatical sentences. Try therefore to be guided by intonation - the rises and falls in the voice - as well as by the words themselves. If it sounds as though someone has finished a sentence and gone on to another (their voice drops, they take a breath and start on a higher note) then it's probably fairly safe to start a new sentence.
Often people change tack during speech, sometimes without pausing or apparently starting a new sentence. Generally, if this occurs soon after the beginning of the sentence, just type on through the change of tack without marking it in any way. If it occurs later on, give the change of tack a new sentence if you feel it's along radically different lines than what came before. Don't worry if this causes the first sentence (the one the speaker was on when they changed tack) to be grammatically incorrect or incomplete.
It is not necessary always to have a full stop at the end of every utterance. When a speaker is interrupted and doesn't finish what they're saying, don't put a full stop. When a speaker is interrupted

and carries on with the same sentence after the interruption, don't put a full stop or capital letter. When a speaker is interrupted and then goes onto a new thought afterwards, don't put a full stop but do put a capital letter.

<M0X> All I was trying
<F0X> I know exactly what you were
<M0X> No you don't.
<F0X> trying to say and I don't like it one bit.

Questions

Only use question marks to indicate things that function as questions. A phrase may be constructed like a question but not function as one:

<X0X> Locked myself out of the house today didn't I.

Or a phrase may function as a question without apparently being constructed like one:

<M0X> But you found them in the end?
<F0X> Oh yes. It turned out to be easier than I thought.

Be wary of rhetorical questions, often marked with things like

Okay right didn't I etc

- these shouldn't normally take question marks, unless a response is obviously expected eg

<M0X> Didn't I. [pause] Well didn't I?

Hyphenation

In general, hyphenate where you'd normally expect to:
in large compound words
great-grandchild
in phrases used as adjectives
devil-may-care silver-plated high-rise
in compound numbers under 100
twenty-two eighty-seven one hundred and thirty-six

But also hyphenate all dates (in order to connect the words so that concordance programs will treat them as a whole):

nineteen-ninety-one the eighteen-sixties two-thousand-and-ten sixteen-0-six fourteen-hundred-and-sixty-two (but NB: the nineteenth century, the twelfth century B C - no hyphens)

Beware - many very long words are no longer commonly hyphenated. When in doubt, do not hyphenate. (The spellchecker will hiccup on words where a hyphen should be included but isn't, but won't notice cases where a hyphen shouldn't be there but is.)

Beware the effect hyphenation can have upon meaning:
twenty-four-ounce cartons = cartons that weigh 24 ounces apiece
twenty four-ounce cartons = 20 cartons that weigh 4 ounces apiece

Quotes:

Do not mark quoted speech or thoughts except to start it with a capital letter (even if it begins in the middle of a sentence):

<F0X> All he said was We don't have to go to school today. I don't know the reason for it.
<M0X> They might be sitting there thinking We can do anything we're in power but what I'm telling you is they've got another think coming.
For quotes from a written source - reading aloud, quoting Shakespeare etc - open the quote with two single left-hand inverted commas (on Alt P) close it with two single right-hand inverted commas. Don't use quotation marks:
<M0X> Which play does that bit about ``the slings and arrows of outrageous fortune'' come from?
Where the quote is a newspaper headline, put lead caps on the important words:
<M01> The Guardian leads with ``Major Proposes Safe Haven Plan for Kurdish Refugees''. That's the main story there. ``The Prime Minister John Major yesterday announced in his address to the United Nations a proposal for the setting up of army-controlled camps for Kurdish refugees fleeing from Iran.'' Er there aren't many details here erm but I gather that it er didn't go down too well.

Words and Spelling

Abbreviations and letter-names

Do not put full stops after abbreviations. Titles are given capital letters as usual. For letter-names, used for example when people spell out words or use abbreviated phrases, use single capital letters. For plurals, don't use an apostrophe:
E G that would mean that M Ps should mind their Ps and Qs.
I bought a two hundred C C bike at ten A M today.
That'll be fifty P for the photocopy and fifty P for my time. I E a pound altogether.
When a word is spelled out, put a space between the letters:
S P E L L E D out
Where an abbreviation consists of letter-names, there should similarly be a space between the capitals when they're spelled out, but no spaces if they're run together as a word (acronym).
V A T (individual letters) **B B C R S P C A**
VAT (pronounced "vat") **AIDS NALGO**
Don't abbreviate **okay**.

Numbers

Type out all numbers in full except the one pronounced "oh", for which we're using the figure **0**
sixty-six four hundred and twelve three 0 seven nine point zero nought point eight
(Note that we therefore distinguish between the number **0** the letter **O** and the exclamation **oh**.)
NB: hyphenation of compound numbers below 100.

Contractions

Use standard contractions

's for "is", "has"**he's/it's/MX's/etc here**

've for "have" **I've/they've/etc been here**
'd for "had"**we'd/you'd/etc already gone**
n't for "not"**can't/don't/haven't/etc**
'll for "will", "shall"**I'll/FX'll/we'll/etc be here**
're for "are"**we're/you're/you're/etc here**
(these lists of examples are by no means exhaustive!)

Use these colloquial contractions where appropriate:

gonna (going to)**wanna** (want to)**gotta** (got to)
oughtta (ought to)**summat** (something) **gal** (girl)
fella (fellow)**'cos** (because)**'em** (them)
dunno (don't know)**penn'orth** (penny's worth)

Don't use d'you or an'/'n' - type **do you**, **and**.

Fillers and affirmatives
Rationalize:
all "yes"-like words (yup, yiss, yeh, yus etc) into:
yes yeah yep
all "er"-like hesitation/filler words (uh, eh, um etc)
into:
er erm
all "oh"-like words and sounds into:
ah (rhyme - car) **oh** (rhyme - no) **ooh** (rhyme - do)
all "grunty" sounds into
mm (one syllable, lips closed)
hm (one syllable, lips closed, starting puff of air)
mhm (two syllables, lips closed)
uh huh (two syllables, lips open)
ugh (noise of disgust, often just a grunt)
Also note:
hey (attracting attention or expressing surprise)
eh (expressing puzzlement or seeking agreement)
oi (as in oi you over there)
ah hah (expressing surprise, as when finding something you've been looking for)

Non-standard grammar
As far as possible, faithfully reproduce examples of non-standard grammar (i.e. don't standardize):
we was I be theirselves I were
But always use **my**, even when it sounds like "me", and always use **isn't** even when it sounds like "aint", "in't" etc.

Capitalization

Generally, in cases where you're not sure whether or not to capitalize a word, leave it in lower case. None of these words need upper case:

doctor professor social worker summer winter north southerly eastern

except in cases where they're part of a title:

Doctor Jones North Ridge Summer Camp

Proper nouns of all kinds do need to be capitalized: names of roads, languages, planets, specialized drugs, song titles, radio and TV stations, wars, political parties, big political or administrative bodies, marketed games:

M Forty, Spaghetti Junction, A Six One Four, Jupiter, Swahili, French, R U Four Eight Six, E Twelve, Any Dream Will Do, Channel Four, Radio W M, B B C Two, Gulf War, World War One, Conservatives, Lib-Dem, Common Market, European Community, Soviet Bloc, Trivial Pursuit, Monopoly,

Spellings

Wherever there's a choice between -ise and -ize spellings, use **-ize.**

realize intellectualize organize rationalize criticize etc.

In general, when you're not sure of a spelling, either use a guess code (if the word just occurs the once), or make a transcriber's note if the word occurs frequently.

One/two words

These lists aren't complete, but so far we've come across these words that need to be clarified:

One word:

altogetheralways almost
unemployedextraordinarystraightforward
unnexessaryroadwordstailbacks
breakdownbusinesslikeradioactive
fallout overturnelectromagnet
rearrangegrandchild

Two words:

good night all rightthank you
straight awayany moregreat-
every daygrandfather

Speaker Anonymity

To preserve the anonymity of members of the public, use **MX FX** and **XX** to replace names (including nicknames and surnames), and remove identifying telephone numbers, addresses and so on. Use transcriber comment codes to note what has been removed:

<F01>Can I give my number out?
<M01>Yeah course you can.
<F01>Okay it's Nottingham
<M01>That's 0 six 0 two isn't it?
<F01>0 six 0 two yeah. <ZZ1> gives telephone number

<ZZ0>
<M01>0 six 0 two <ZZ1> repeats telephone number <ZZ0>

People whose names are in the public domain (politicians, actors, pop stars, radio DJs, sportspeople etc) don't need to be protected in this way.

References

Blanche-Benvenise, Claire & Colette Jean Jean (1987): *Le Français Parlé: Transcription & Edition*. Paris: Didier Erudition.

Burnard L. (1992a): "TGC W30 Corpus Document Interchange Format v.1.0", Oxford University Computing Service.

Burnard L. (1992b): "The Text Encoding Initiative: A progress Report", Oxford University Computing Service.

ESPRIT Project 2589 (SAM). User Guide to ETR Tools. Multilingual Speech Input/Output: Assessment, Methodology and Standardisation. Ref. SAM-UCL-G007.

Johansson S., Burnard L., Edwards J., Rosta A. (1991): "TEI - Working Paper on Spoken Texts".

Jones S., Sinclair J.M. (1972): "English Lexical Collocations", in *Cahiers de Lexicologie*, 24/15-61.

Sinclair J. M., Jones S., Daley R. (1970): *English Lexical Studies.* University of Birmingham, for Office for Scientific & Technical Information.

Sperberg C.M., Burnard L. (1991): *Guidelines for the Encoding of Machine Readable Texts*, draft version 1.1, 2nd printing, Text Encoding Initiative, Chicago, Oxford.

Warren M. (1993): *Naturalness in Discourse*. Ph.D. Thesis, University of Birmingham.

Relevant NERC Papers

Belica C. (1992): "The TEI proposal and its feasibility. An evaluation in respect to the IDS Corpus philosophy", IDS Mannheim, Working Paper, NERC-36.

Candel D.V. (1992): "Traitment de corpus oraux - Groupe Aixois de recherche en syntaxe (Notes on Transciption)" - WOrking Paper, INALF Paris, NERC-157.

Cauldwell R. (1992): "Accessing Spoken data using compact disc and hypertext", University of Birmingham, Working Paper, NERC-48.

Cauldwell R. (1992): "Johansson et al's Proposals", University of Birmingham, Working Paper, NERC-106.

de Jong E.D. (1992): "Transcription and normalization method. Dutch spoken language", Utrecht, Working Paper, NERC-159.

Edwards J.A. (1992): "Design Principles in the Transcription of Spoken Discourse", University of California, Working Paper, NERC-164.

French J.P. (1991): "Updated Notes for soundprint transcribers", University of Birmingham, Working Paper, NERC-47.

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

Kirk J.M. (1993): "The Northern Ireland transcribed corpus of speech", Queen's University of Belfast, Working Paper, NERC-158.

Payne J. (1992): "Speaking the same language? - Listening to the speech community", COBUILD Birmingham, Working Paper, NERC-132.

Payne J. (1992): "Report on the compatibility of JP French's Spoken corpus transcription conventions with the TEI Guidelines for transcription of spoken texts", COBUILD Birmingham and IDS Mannheim, Working Paper, NERC-122.

Psathas G., Anderson T. (1992): "The 'practice' of transcription in conversation analysis", INL Leiden, Working Paper, NERC-163.

Scheiter S. (1992): "German spoken language corpora and their text representation schemes - an overview", IDS Mannheim, Working Paper, NERC-43.

Villena-Ponsoda J.A. (1992): "Representational Procedures and Schemes for Spanish oral Corpus of University of Malaga", University of Malaga, Working Paper, NERC-141.

Chapter 4

Text Acquisition and Reusability/ Access and Management Software Tools

A. Text Acquisition and Reusability

1 Introduction

This part of the chapter gives an overview of the main methods for electronic text acquisition and, for each method, presents the most appropriate techniques and tools for conversion into SGML-coded text, conforming to the minimum representation level defined by Work Package 3 (Lafon and Vignaud, 1992, NERC-150) and described in the Chapter 3 above. In the rest of this document, this minimum level is referred to as the «target level». We have also tried to evaluate the corresponding costs.

We have considered four main acquisition methods: OCR (Optical Character Recognition), photocomposition tape analysis, direct data input and text retrieval over computer networks and from databanks.

To convert the text from these various sources into SGML-coded text complying with the required model, we must first be able to read the contents of a document and retrieve all the information available without resorting to interpretation. We must then be able to retrieve all information on text structure, which, depending on each case, concerns typographical aspects or a more or less abstract level. We must also achieve an understanding of how the text is marked up, and then use all this information to generate the target level. We apply the term «retroconversion» to all techniques enabling a document in a given source format to be converted into an SGML document in compliance with a target DTD.

Depending on the source of the text, the level and type of difficulty may vary. For example, a photocomposition tape may be very difficult to read, but once it has been decoded, its typographical structure is such that, depending on the target DTD, the document can be converted to SGML. On the other hand, text transferred over a network may be very easy to read but it does not have a structure mark-up, thereby making it difficult to convert. The retroconversion process, therefore, has several facets depending on the source of the texts. We will begin by defining the general retroconversion methods and then go on to study how these methods can be applied to each type of acquisition and what specific problems may arise in each case. The final report of Work Package 7 contains an overview of the various tools available on the market (Lafon and Vignaud, 1992, NERC-151).

2 SGML Retroconversion Techniques

2.1 *Two Phases*

The term «retroconversion» covers all techniques used to convert a document in a given input format to an SGML document complying with a target DTD. In practice, the input document is not converted directly into the target SGML format. A transit format, equivalent to the initial format expressed in SGML, is defined.

Two separate operations can thus be distinguished:

-Decoding, that is source document code syntax analysis

-Interpretation, that is extrapolation of a document complying with the target model from the organization of the decoded source text

A transit DTD is used to decode the source text and convert it into SGML, thereby generating an exact reproduction of the initial structure. This transit DTD is independent of the target DTD, the purpose of using it being to create an intermediary document, with a valid SGML syntax, which is significantly easier to read and handle than the source document. Powerful tools, such as SGML handling languages, can then be used for the second phase.

During the second phase, the transit document is converted into an SGML document in compliance with the target model.

In favourable conditions, the decoding phase may become purely mechanical, in which case it can be fully automated. On the other hand, the interpretation phase is an attempt to project an abstract model on to a document, which we assume can correspond to this model. It may, therefore, imply more or less arbitrary human decisions, and may even require that the source document be reorganized for the model to be applied.

Transit DTD definition is not an exact science. The structural characteristics observed repeatedly in the source document or detected automatically are described. For a printed document to be processed by OCR and visual structure recognition, this exercise is relatively intuitive. For typeset or word processor files, the codes are analyzed: as it is difficult to be exhaustive in this analysis (particularly when faced with the lack of comprehensive documentation), the work is often

conducted iteratively, the transit DTD being gradually enhanced as the analysis progresses.

In projects like this one, where the target DTD envisaged is only a minor abstraction compared with the analyzed texts, we can question the need to divide the process into two phases. We believe, however, that it is preferable to apply this method systematically, the advantage being that the target DTD element recognition procedures can be programmed generically, independently of the origin, media and format of the source data. For example, proper nouns can be recognized and marked up by the same program on two transit files, in which the basic structures (paragraphs, titles, etc.) have been marked up by two totally different methods: visual recognition for the OCR software output in the first case and by word processing format analysis in the second case.

2.2 *Decoding Phase*

The aim of the decoding phase is to solve two problems: content acquisition (text characters) and implicit structure level recognition, as highlighted by the transit DTD.

Content acquisition methods depend entirely on the source media: data input or OCR for paper documents, computer-based reading for mechanographical documents (punched tape) or magnetic media. When describing the various acquisition methods, we will see what specific problems arise in each case.

Two approaches are generally used for the structure recognition methods:

-Lexico-syntactic formatting code analysis

- Visual page structure recognition

2.2.1 *Lexico-syntactic Analysis*

Lexico-syntactic formatting code analysis is the method which naturally springs to mind when texts are available on computer media: photocomposition tapes, word processor files, etc. For it to be applied, we must have a full set of clear documentation describing the input codes and this is not always the case. The tools used can be divided into two categories:

-Tools applying single or multiple transcoding tables activated according to the context

-Lexical and syntactic analyzer generators, such as the standard Unix utilities, lex and yacc and their variations or the INR program developed by the University of Waterloo. Before using such tools, a formal grammar must be defined to describe the input flow.

In practice, the main problem which arises with all tools is the variability of the input codes used to

express the same formatting attribute. This is the case, for instance, when analyzing typeset files, in which the code sequences vary according to the personal style and tricks of the trade applied by each operator; it is also the case when analyzing word processor files (like Word files), in which different codes can be inserted to obtain the same page presentation, depending on whether a style sheet is used or not. In more general terms, we are likely to meet this problem whenever the only way to validate a document is to examine its typographical appearance on a sheet of paper.

2.2.2 *Visual Page Structure Recognition*

Visual recognition software products set up a computerized representation of the geometric structure of the page (blocks, indentation, etc.) and the typographical variations in this structure (font, size, weight). The recognition algorithm draws on this information and some programmer-stipulated rules. This approach, therefore, only concerns documents with a regular structure.

The advantage of this approach is that it gets round the code variability problem described above. The software applying this method to coded input files limits its attempts at «understanding» the codes if it can rebuild the typographical structure of the page in memory; code variability is, therefore, «smoothed over» in the resulting file.

The FastTag (Avalanche Corp.) software and its derivation TextTagger (IBM) apply this approach. Several university research projects have also produced prototypes in this field (Coray, etc.).

2.3 ***Interpretation Phase***

The interpretation phase uses SGML tools on the transit DTD. All structure change (reorganization, data delocalization) and semantic interpretation operations are held back until this phase. When programming this phase, human intervention is necessarily required for the analysis.

The greater the differences between the source text structure and the target structure, the harder this phase becomes. The types of tools used depend basically on the degree of data delocalization between the source and target structures.

3 Main Acquisition Methods

Here we cover only the main problems met by each acquisition method; for more detailed information see (Lafon and Vignaud, 1992, NERC-151).

3.1 *Optical Character Recognition (OCR)*

3.1.1 *Character Recognition*

The main problem faced by the OCR method is the error rate in character recognition. AIS Berger-Levrault has conducted tests on original pages from the Trésor de langue française (TLF) using various products available on the market. The resulting error rates in character recognition are in the region of 2 to 3%. Faced with such error rates, the OCR pass must be followed by in-depth re-reading, which takes longer than it does after manual input. In such cases, the OCR method ceases to be economically viable.

Caution must be applied when considering the veroptimistic rates supplied by the OCR software itself. The error rate indicated only takes into account the characters that the software knows that it has not identified and not those that it mistakenly thinks it has identified correctly.

Moreover, the physical qualities of the paper (ink, state of the paper, transparency) have a significant impact on the system's performance. The OCR method should not be used unless a printed copy of excellent quality is available (this is a frequent problem as far as analyzing newspaper text is concerned).

Techniques based on artificial intelligence are used more and more frequently to remove ambiguities when identifying characters: context analysis, use of lexicons, application of grammar rules. It should, however, be noted that the rules applied have been defined for everyday texts and may lead to misinterpretations and loss of information for specific literary texts (particularly old ones). The Image-In software, which uses neural networks, is a good example. It claims a character recognition rate of approximately 99.99%, but this rate corresponds to asymptotic performance in optimal conditions.

3.1.2 *Typographic Variation Recognition*

Recent OCR software products are increasingly capable of:

-Segmenting pages, by identifying columns, blocks, illustrations, headers and footers. The techniques used are those initially developed for automatic form reading software; they have been improved and made more flexible so that they can be applied to general-purpose documents.

-Recognizing font, weight and size changes.

To enable this information to be transmitted to the rest of the processing system, a «rich text» format must be used. Some systems provide a specific format (the Kurtzweil or Calera systems, for instance), but most use market standard word processor rich text formats, such as RTF (Rich Text

Format) defined by Microsoft for MS-Word.

3.1.3 *Decoding Phase*

It seems much more logical to use visual recognition methods after the OCR systems. When the rich text format supplied by the system includes information on page structure and typography, a maximum amount of information is available for the recognition heuristics.

Methods of lexico-syntactic analysis can, nevertheless, be used on the output files. They at least have the advantage of using well-documented or known coding systems.

3.1.4 *Costs*

Cost estimates can be based on operator times, using processing speeds of around 20 000 to 30 000 characters per hour. These figures assume that the decoding phase is fully automated.

3.2 *Text Acquisition Based on Photocomposition Tape Analysis*

3.2.1*Content Reading*

Physical access to data stored on <<photocomposition tapes>> is in itself likely to be a problem. Only recently designed photocomposition tapes can reasonably be used. For text edited more than 10 years ago, we cannot ignore the problem of ageing of magnetic media; phototypesetting firms are no doubt storing hundreds of worthless tapes in their storerooms.

Once we have solved the problem of how to read the media, we are faced with another one, resulting from the great diversity of character codes and typesetting instructions from one system to another. In this area, the quality and completeness of the documentation provided is a key factor to be taken into account during the feasibility and cost evaluation study for retroconversion.

3.2.2 *Decoding Phase*

Methods of lexico-syntactic analysis have a predilection for photocomposition file analysis. We are, however, faced with the following problem. It is often possible to obtain the same typographical result using very different code sequences; file coding varies with the personal style of the operator and the tricks of the trade used. As a result, file coding is often very heterogeneous.

The use of visual recognition methods assumes that we are able to simulate the graphic effect of all typesetting codes detected, which to a certain extent comes down to redeveloping the typesetting program itself; this is likely to be a very complex task.

3.2.3 *Costs*

Photocomposition tape analysis eliminates costs associated with re-entering and correcting content. We can therefore say that retroconversion costs are limited to those required to develop the decoding

program (if we do not take into account minor costs to do with data media handling). The savings generated by this acquisition method are thus closely related to the feasibility of automating the decoding phase. If the tapes can be easily read and the input coding systems are correctly documented, it is by far the most economical solution for large volumes of text.

3.3 *Text Acquisition By Direct Data Input*

Several input techniques can be used today:
-Non-formatted input without any typographical enriching, followed by a typographical enriching mark-up phase
-Non-formatted input, with direct typographical enriching
-Direct SGML input based on the DTD, using an SGML editor

3.3.1 *Non-formatted Input*

Generally speaking, data input by a person is characterized by a good character recognition rate; normal pre-rereading error rates, without dual input, do not exceed 1 to 2 per 1 000 and fall below 1 per 10 000 with dual input (generally accepted solution today). Moreover, the tolerated error threshold is often stipulated in the contract between a client and supplier.

Non-formatted input without any typographical enriching is a standard service provided by data input agencies. A speed of 12 000 characters per hour can be achieved by an average operator when the work is easy (input of a novel from a perfectly typed copy, for example).

Non-formatted input with typographical enriching does not usually exceed 10 000 to 12 000 characters per hour for averagely enriched texts (legal text, for example). For highly enriched texts, such as a dictionary, in which there may be up to three or four font changes on one line, the average input rate may drop to 6 000 characters per hour, including re-reading.

In this area, the only way to obtain a clear idea of expected average performance is to measure input speeds over short samples.

3.3.2 *Direct Input using an SGML Editor*

Direct input via an SGML editor is possible if the DTD has been formalized and tested. Several tools aimed at a production environment are now available, such as WriterStation (Datalogics) or Textwrite (IBM). Specialized data input teams are required to use these tools. Other tools on microcomputers, such as Author/Editor (SoftQuad), seem to provide better ergonomics for data input by the author of the text. Experience gained over the last four years by Berger-Levrault/GTI has shown that input speed with this type of tool varies between 4 000 and 6 000 characters per hour, depending on DTD complexity.

It should be noted than when entering data via an SGML editor, the operator performs the content acquisition, text decoding and interpretation operations at the same time, in line with the target DCD.

It is often thought that once the text entered using an SGML editor has been parsed without generating any error messages, it is valid. In fact, errors in the way the content is allocated to SGML elements may remain as they are often only detected by the re-reading of hardcopy. As a result, major projects concerning input via an SGML structure editor often include a «semantic correction» phase conducted on text samples. The system checks that the marking up is correct for all elements to which a marker must be applied (the mandatory nature not being expressed in the DTD); proper nouns fall into this category. It also checks that the content of an element corresponds to its associated definition.

3.3.3 *Costs*

Non-formatted data input costs are now very low, in the region of 7 FF per thousand characters, thanks to the setting up of specialized workshops in countries like Taiwan, the Philippines and Madagascar. (In Europe, prices vary but average about 15 FF per thousand characters.) The cost of input using an SGML editor is around 30 FF per thousand characters for a reasonably complex DTD.

3.4 *Text Retrieval Across Networks and From Databanks*

3.4.1 *Technology, Information Sources*

Over the past twenty years, a great many research institutes and universities, especially in the United States, have entered large volumes of text in order to conduct computer-based studies. With the development of computer networks, particularly Internet, these texts, often stored on-line, can now be easily accessed by direct transfer (FTP protocol) or by electronic mail.

The Georgetown Center for Text and Technology (Georgetown University, Washington D.C.) recently set up a catalogue of projects and agencies handling machine-based textual resources. Most can be accessed across a network. This catalogue is included in (Lafon and Vignaud, 1992, NERC-151).

Amongst this list, the Oxford Text Archive collection is without doubt one of the richest and most diversified. An extract from its catalogue, corresponding to works in the French language, is also provided in (Lafon and Vignaud, 1992, NERC-151).

Finally on this same subject, we would like to point out that the Le Monde daily newspaper has a documentary database containing all articles written for the paper since the beginning of January 1987. The articles are added to the database after they have been typeset but the typesetting codes are removed. Some marking up remains, in order to enable the main logical items of an article to be

found (date, headline, author, etc.). This marking up is not checked and many variations therefore exist. Le Monde has informed us that it has already provided some research institutes with restricted database extracts, corresponding to approximately one month of publications. For more substantial extracts, commercial-type agreements are envisaged.

3.4.2 *Specific Problems*

The first problem met when using textual resources available across a network is the variety of codes used; it is precisely to resist this unfortunate proliferation of incompatible coding systems that the TEI was set up. The Oxford Text Archive lists no less than forty different coding systems.

Another problem is the possible absence of all structure and typographical markers, except for line breaks. All codes are often intentionally removed from the texts so that they can be stored as ASCII files. When markers are present, they have more often than not been entered in order to meet a very specific and specialized research objective, and the underlying abstract model is generally not explained.

On the other hand, the fact that these texts can be very easily accessed and reliably transferred means that networks are now a prime source of textual data.

3.4.3 *Costs*

The cost of accessing textual resources over the Internet network is divided up as follows:

-Cost of hooking up and transferring to the closest node on the Internet network (TRANSPAC, telephone line, ISDN)

-Internet service subscription and usage costs. Prices vary greatly according to the type of client body (university, research laboratory, private company), the type of subscription taken out (electronic mail, FTP or not) and the volume processed (commitment to an annual consumption).

4 Conclusion and Recommendations

The study that we have conducted covers the main retroconversion methods and associated tools. Appendices 1 and 3 in (Lafon and Vignaud, 1992, NERC-151) contain the respective contributions of the Birmingham and Malaga teams.

The NERC teams can refer to this document when drawing up homogeneous text databases in compliance with a target DTD, such as the one recommended in (Lafon and Vignaud, 1992, NERC-

150).

Given its nature, this technical study cannot really lead to recommendations. Its scope was limited to considering advantages and disadvantages, occasionally expressing caution with respect to certain aspects of each acquisition method, whilst attempting to evaluate costs. This study, therefore, contains the main elements used to draw up a comparative cost evaluation of the various acquisition methods, provided that the concrete conditions required to build a database are determined.

Two points do, however, require emphasis.

For the retroconversion of existing records, we recommend that the process be systematically divided into two phases: decoding and interpretation (see I. SGML Retroconversion Techniques above). This point is based on actual retroconversion experiments performed by Berger-Levrault AIS.

The implementation of a retroconversion method is still a complex and difficult operation, requiring that programs be written or general tools adapted and many checks performed. Return on initial investment only becomes possible if the texts to be processed:

-Are themselves reliable; this point must be systematically checked

-Are sufficiently well-documented to avoid expensive trial and error when implementing the process

-Are of a sufficient volume to justify the initial investment

It would not, therefore, be reasonable to envisage generalized retroconversion of small databases containing readily available text taken from a great many sources and using different mark-up systems. It is preferable to work on a small number of carefully selected textual sources.

Relevant NERC Papers

Lafon P., Chahuneau F. (1992): "Acquisition de textes et réutilization", INALF Paris and AIS Berger-Levrault, Technical Report, NERC-151.

Lafon P., Vignaud D. (1992): "Répresentation des textes écrites", INALF Paris and AIS Berger-Levrault, Technical Report, NERC-150.

B. Access and Management Software Tools

1 Introduction

The context of this part of the chapter must be understood first of all. It is proposed that realistic targets for corpus access and maintenance are identified for a network of reference corpora throughout Europe, and that these are harmonised in such a way that the user community is able to retrieve data from the corpora with identical tools and routines, regardless of the location of user or corpus, or the hardware at any particular site.

In a stable environment, and with a lot of effort and patience, this could be achieved; once achieved it could be maintained indefinitely. Difficulties could be overcome once for all, and new nodes in the network could be added with relative ease.

However, in the present environment of the NERC study, the position is fluid, dynamic and turbulent. Corpora and associated tools have been worked up for different reasons, to different specifications, and with different rationales, all over the EC. Research teams are at different stages of awareness and have many diverse relations with sponsors and funding sources.

Further, the speed of development of computer hardware and software is bewildering, and plans are routinely revised twice a year. Opportunities to co-ordinate and harmonise are on the increase, but powerful new tools are always pulling developers towards unique facilities.

It is particularly important in such an environment that a report of this nature is not used to restrict or hamper progress. The spirit of these observations and recommendations is to promote the use of corpus evidence wherever languages are studied or used in applications. It is anticipated that corpora will increase in size, and in speed of growing, until the monitor corpus (Clear, 1988) stage is reached. It is anticipated that the needs and demands of users will become greater and greater, and the sophistication of users will increase dramatically. The sheer number of users will rise exponentially until almost everyone will become, in one way or another, a direct user of a corpus (through the kind of telephone service available in some countries, or the access to FRANTEXT (see Paris Group, 1992, NERC-178) through a large number of terminals). The design problem in this area is thus similar to cutting a flight of steps in a steep rock face. The developer sees the whole flight of steps and the slope on either side; the user sees the particular step from which a particular vista can be enjoyed. The pace of development is very fast, the directions are various but virtually all are expansionist. The user, particularly the professional user, needs accurate information without the detail of the constantly changing environment. The user needs to perceive a stable platform from which he or she can observe the passing language; the access and retrieval procedures must be steady and reliable over a considerable period of time. Otherwise there is no point in investing in the time and expense of gaining access.

For example, the size and constitution of a corpus for users must remain stable over time. A user could not tolerate a corpus which was larger every time it was consulted, no matter how much it was improved on each occasion. Equally, if texts were removed, even for good reasons, the user would find the corpus unreliable. Enhancements to software might be of considerable benefit in the

medium and long term, but infuriating to the user familiar with an earlier version; what is an improvement to many users may be a distinctly retrograde step to some.

Underlying the above remarks is a long history of experience of corpus work. Every problem referred to has arisen and has caused delay and inefficiency. A small research team working closely together, reorganising their procedures almost daily, in constant consultation and with considerable personal discretion may just be able to work in a constantly changing environment. But the mass of users must have stability for substantial periods.

We know, of course, that the real situation in the corpus linguistics laboratories is of a great speed of change. The e-mail list "corpora" (organised by Knut Hofland in Bergen) produces information which is already almost beyond being organised at each site. The development of cheap storage media and electronic publishing is transforming the opportunities for corpus design, and the whole spectrum of research into knowledge bases, expert systems, NLP and content analysis is focusing gradually on natural language understanding. During the existence of NERC, corpora have moved from the periphery to the dead centre of language study.

It is most important that the needs of users do not divert or encumber this very promising area of research and development. The design of access and management systems must be such that users can operate within a stable environment, while developers can alter their environment at will.

One significant process that the NERC initiative seeks to promote is the easy upgrading of corpus dimensions and access specifications. In the proposals that follow, an assumption of "state of the art" is made. That is to say, the participating institutions of NERC are reflecting what they consider to be good practice at the present time - not too ambitious, but capable of being replicated by any serious corpus provider in a community with advanced technology. However, it is recognised that many corpus providers may not be able to meet these specifications at the present time. They may be promoting a minority language which attracts few resources; they may have little experience behind them; they may have specialised purposes which demand other specifications.

In any such circumstances, we suggest that the supporting NERC Papers listed at the end of this part are consulted, because they offer a varied record of experience. The servicing needs of a million-word corpus are quite different from those of a ten m.w. corpus, which are different again from those of a fifty-plus m.w. corpus. Simpler strategies can suit simple needs, such as the orientation of students or the basic training of lexicographers. The specifications proposed here are not intended to be forbidding, but to be an attainable target for those with plenty of experience and adequate resources.

(Lane, 1992, NERC-56, Castillo Cabezas, 1992, NERC-55) deal mainly with data input conventions, linking with Chapter 3 (Written Text Representation) and part A. of this Chapter (Text Acquisition and Reusability). (Picchi, 1991, NERC-11, Picchi, 1991, NERC-9, Van der Kamp et al., 1992, NERC-51, James, 1992, NERC-89, Johns, 1992, NERC-90) give a perspective on the operating systems in use among NERC partners, including MS-DOS, UNIX and VMS. There is a wealth of experience reported here, on which the recommendations in this chapter are based. (Picchi, 1991, NERC-11, Picchi, 1991, NERC-9, Van der Kamp et al., 1992, NERC-51, Dendien, 1991, NERC-113) go on to illustrate database structures suitable for corpus applications.

(Lane, 1992, NERC-114, Lane, 1992, NERC-101, Krishnamurthy, 1992, NERC-87,

Krishnamurthy, 1992, NERC-88, Van der Kamp, 1992, NERC-85) concentrate on the central issue of software tools for access to corpus information. Users need to be able to examine the language of a corpus, and comparison of practice in Europe suggests that a great deal of the access and retrieval required is common to most users. Hence a distillation of what is likely to be both desirable and attainable is offered in section 2 below.

The later parts of the chapter become sometimes counsels of perfection. To our knowledge, no-one is as yet achieving the standards that we propose. This situation arises for two reasons

(a)the establishment of a cycle of upgrading is still unusual among corpus providers, most of whom have only a few years' experience.

(b)the pressures of work in a corpus linguistics laboratory makes it very difficult to establish reliable working practices, because their establishment uses expert resources and needs a continuous programme of monitoring.

We make it clear when we are recommending procedures that do not presently exist, and we commit ourselves to the attainment of the standards we propose, as quickly as possible. There is no practicable alternative to this position, which is unlike virtually all the other recommendations of the report. Corpus linguistics is only just emerging from a long gestation period, when hardly anyone took it seriously. Resources were minimal, and pockets of activity were isolated. Few potential users made contact with the corpus provider. With the sudden upsurge of activity, most laboratories are severely stretched to keep up with current activities, far less to plan efficient access in the future, and far far less to update the results of older practices. Hence a general caution concerning what can be achieved in the move towards standardisation, and slow progress towards fully exchangeable corpus data.

2 Basic Access Software

This section must be read alongside the recommendations of Work Package 9 (see Chapter 6), which lists principles for the design of software for today's corpus needs. The "Desiderata" of (Lane, 1992, NERC-101, § 5) are also recommended on the technical side. For ease of access, both lists are printed here:

(a)Guidelines for software design

1.Analysis should be restricted to what the machine can do without human checking, or intervention.

2.Analysis should be done in real time

3.Operations should be designed to cope with unlimited quantities of text material

4.Software will be designed to operate at more than one level of discrimination, so as to bypass doubtful decisions

5.Speed of processing should take precedence over ultimate precision of analysis

6.Software should be robust

(b)Desiderata

1.The utility of itself should impose no limit on the size of a corpus (limits dictated by machine considerations such as the amount of hard disk available are unavoidable).
2.The relationship between corpus size and indexing time should approximate linear proportionality.
3.The nature of the reindexing task should not discourage corpus alteration.
4.If the utility expects or allows the existence of multiple corpora, simultaneous interrogation of them should be possible.
5.Software limits on, say, the maximum length of a "word" or on the number of lines that can be written to an output file should be set at generous levels.
6.The utility should permit the simultaneous retrieval of more than one word. The words to be retrieved should be selectable, by at least a subset of the Unix regular expressions, from a wordlist of the corpus. More generally, one might expect a reasonably sophisticated query language to be evolved, allowing combinations of some of the query-types mentioned in this document.
7.On-screen pruning of a set of citations should be straightforward, prior to writing to a file. It is further assumed in this chapter that corpus providers are aiming to establish monitor corpora as soon as possible (see Chapter 0 - Implementation Plan). It is further assumed that the concept of networking access to corpora is established as a medium-term working target.

2.1 *Common Functionality*

At this point in the development of access software, it is felt to be valuable to state the specifications of a suitable functionality. Individual institutions may well provide more than what is set out here; our priority is to establish that they wil not provide less. For the functions listed below are those which a generation of corpus users either has wanted badly or rejoiced to have.

2.2 *Operating Systems*

There is no prospect at present of recommending a single operating system for corpus management and processing. Indeed there is a strong body of opinion that feels that any movement at this point towards a standard would be counter-productive. Such is the pace of development, we can look forward to an ever-improving range of alternatives: if one system was selected, it might preclude future avenues of improvement.

Corpus computing has originated from three different hardware systems - mainframes, minis (now workstations) and microprocessors (now PCs). Each has developed operating systems, which

are suitable for corpus processing, but which reflect the size, scope and architecture of the machine. MS-DOS is the principal operating system of the PCs, and UNIX is the widely used workstation system. Mainframes are now less commonly used for corpora, in these days of distributed computing, but some mainframe operating systems are still in use.

Appreciations of MS-DOS and UNIX are to be found in (Picchi, 1991, NERC-11, Castillo Cabesas, 1992, NERC-55) respectively. As the demands grow, these operating systems may grow to accommodate them; if not then corpus managers will switch to something more suitable.

2.3 *A Standard Query Language*

Given that operating systems remain an important variable in corpus linguistics, it is felt necessary to propose the establishment of a standard query language for corpus access and retrieval. This language should be developed in a simple formalism which will keep it free from any one operating system. It can thus be devised, discussed, enhanced and varied by the community of users from time to time, without the restrictions of an operating system. It is recommended that the EC funds the modest cost of the initial development of such a formalism. Eventually, each corpus provider would be responsible for implementing the standard query language on his or her local operating system. It should not be EC policy to finance the implementation onto unusual or old-fashioned local operating systems. In a short time there will be sufficient shared software in one or two widely used operating systems to promote a policy of replacing inappropriate installations with a more suitable operating system, rather than the EC accepting the expense of implementing the query language on an inappropriate operating system. In this way, support will be given indirectly to those operating systems that are favoured by leading users.

A corpus provider who uses an unusual operating system may have good reason, in which case funds may be found to implement the query language. Otherwise, a case should be made on a comprehensive argument for a change in operating system to one on which the query language is already implemented. The initial content of the standard query language is set out in the section on functions below. It is written informally to encourage understanding and discussion.

The importance of full and accurate documentation is stressed. Users should be aware at all times that they are using software tools, which work usually in very simple ways, providing simple results. Results may be wrongly interpreted if the technical details are not readily available.

It is easy nowadays for machines to report in a misleading fashion, and perfectly possible for them to report information which is quite wrong. "User-friendliness" can lead to simplifications which can give false impressions of the real nature of the data. Full documentation is often difficult to access. As a result, users may be seriously misled. The design and composition of suitable documentation is a major matter of follow-up from this NERC feasibility study.

2.4 *Functions*

The statement of function is organised in seven sections, reflecting the seven basic stages of a normal retrieval request.

(a)the user specifies the item to be searched for.
(b)the computer reports on corpus holdings with respect to the item.
(c)the user specifies his/her selection of the corpus holdings, using a set of parameters of classification to apply to the internal structure of the specified item.
(d)the user optionally specifies features of the environment on either side of the search item.
(e)the computer reports on the completion of steps `c´ and `d´.
(f)the user optionally sets up a cyclical refinement by returning to step `c´
(g)the user gives instructions for the disposal of the retrieved material (simple disposal instructions can be given in advance, at the end of step `d´, in which case steps `f´ and `g´ do not take place).

2.4.1 *Item specification*

1.Any regular expression
2.(Including any character string)
3.Of any length
4.With or without gaps of variable length
5.With or without wild cards of various functionality
6.Any annotation (e.g. a word class tag)
7.The annotation at any place in the string
8.Any mixture of annotations and character strings.

2.4.2 *Report of corpus holdings*

1.The precise number of instances
2.No maxima or minima restrictions on any item or combination.
3.The full enumeration of options (where the specification contains options)
4.The ability to report on the text location of any or all instances

2.4.3 *Selection parameters - internal*

1.Definition of a sub-corpus within the corpus*
2.Editing of locations (i.e. removal of unwanted locations)

3.Selection of the number of instances required.
4.Editing of the specifications by modifying the settings at (2.4.1) above (in each case of modifying a selection parameter, the other parameters will automatically change concomitantly).

* This can be set in advance of a retrieval session, as well as at this point during a session.

2.4.4 *Selection parameters - environment*

1.Extent of environment, specifiable in:
a)characters
b)words
c)analytic units (e.g. tags)
d)structural boundaries, especially sentence
e)(default setting KWIC format, screen width)

2.Identification of text location of each instance (optional) (simple disposal instructions can be given at this point)

2.4.5 *Report in the formats specified*

2.4.6 *Refinement (optional)*

return to step (2.4.3) for further editing of parameters (steps (2.4.4) and (2.4.5) follow).

2.4.7 *Disposal*

1.Default is screen presentation as in (2.4.4.1.) - KWIC format with screen enhancements such as:
-optional wider context-scrolling up and down -text location
2. To file (with editing options)
3.To editor
4.Piped to other processes

3 Corpus Maintenance, Development and Availability

3.1 *Issues*

The issues discussed in this section are those of any substantial institution that handles electronic data. They reflect the fact that most corpus work is likely for some years yet to be stretching computing resources, both hardware and software, and the ingenuity of the expert staff involved.

The section is organised as follows:

3.1Issues
3.2Corpus maintenance
3.3Enhancements to software
3.4Development routines
3.5Corpus availability to the user community

3.2 *Corpus Maintenance*

Corpora are growing in size, viz, approximately

1960s 150,000 words
1970s 1 million words
1980s 20 million words
1990s 200 million words

These figures are in any case overtaken by the availability of text in electronic form (subject to the costs of reusability; see this chapter, part A.), which is now overwhelming in the major languages of Europe. Much of the future work of corpus providers will be the control of very large amounts of material, the correct identification and classification of data, and the encouragement of the supply of material that is not produced in electronic form.

At any time a corpus manager must know what is in the corpus, where it is in the system, what form or forms it is in, and what processes are available to apply to it. Rigorous back-up routines must be carried out, and regular checks must be made on the integrity of the corpus.

Electronic data is prone to disturbance of many kinds, and modern corpora are of a size beyond the possibility of human checking. Once established, a corpus or sub-corpus is likely to be in use for many years, and users will assume that it is unchanged in size, consistency and format.

It is a major responsibility of corpus providers to ensure that the specification of a corpus remains accurate, and to protect the integrity of the corpus. Since a corpus is characteristically open to inspection by large numbers of users, and very bulky in a computer system, its security is a non-trivial issue. Users should expect to know what measures are taken to protect it, and standard conventions should be established to guarantee the long-term investment of users.

The issues of maintenance of a large modern corpus are discussed further from the point of view of a corpus manager in (Clear, 1993, NERC-153).

3.2.1 *Protection of Rights holders*

The question of copyright and other rights in corpora is addressed for NERC by the Institut de Recherches Comparatives sur les Institutions et le Droit, CNRS, Paris. It is anticipated that for some

time to come the responses of rights holders will be highly variable. All sorts of restrictions will be imposed on corpus providers, including the securing of extensions of permissions at varying intervals, small payments in some cases, and different requirements for acknowledgement. In a multi-million word corpus this is not a trivial matter, and though burdensome is a necessary response to what have been, in the main, generous permissions.

Most of the handling of rights and permissions can be automated if it is anticipated from an early stage in a project design. Correspondence can be generated automatically, and date prompts. The maintenance problem can be kept to a minimum of human intervention.

3.3 *Enhancements to access software*

In the early days of corpus linguistics, it was sufficient for the software to retrieve only instances of surface features e.g. character strings in contexts, making concordances. More sophisticated requirements extended the concept of basic requirements to the functionality set out in section 2 of this part of the chapter. It must be expected that this process of enhancement will continue indefinitely, and that corpus providers will have to adopt a policy of steady upgrading. (James, 1992, NERC-130) introduces this topic from a practical point of view, and (Monachini and Picchi, 1992, NERC-105) illustrates a further step. This paper is also a bridge to Chapter 6 (Annotation Tools), where forward planning opens up new horizons for the user, which then generate pressure to upgrade standard facilities.

From a European perspective it is sensible to plan and implement the upgrades in a co-ordinated fashion. Even if the development of annotations is not even throughout the community, the software can be kept compatible. For example, it is likely that in the next few years a new interest in research into statistical analysis will lead to proposals for enhancements to the basic software. Such software will be adaptable in any language.

Many of the enhancements that can be expected will arise from individual research projects, and may be difficult to implement, idiosyncratic and sometimes commercially sensitive. It is to be hoped that a consortium of corpus providers will have the resources and initiative to maintain contact with the evolving corpus activity, and identify valuable, generally applicable enhancements to the basic software.

3.4 *Development Routines*

Each new development should be carefully planned and should not be introduced until
1.the software is robust and user-friendly
2.it is compatible with existing established practices
3.it is very fully documented

4.it has been substantially piloted and is popular with selected users
5.provision is made for publicity, information and orientation of the user community
6.it is available simultaneously on all sites
7.users can for a period ignore the update and continue to use previous routines.
8.the new facilities are evaluated by the providers and the users from the start, and there is regular feedback and report.

3.5 *Availability*

The primary assumption of corpus availability in this report is that it will be arranged through communications networks rather than by physical copying of corpora (e.g. on CD-ROM). The latter process is seen as a valuable aid to the study of long texts in a language, and NERC wishes to promote and support the preparation and distribution of CD-ROM collections for a variety of purposes, given below. However, given the need for stability, balance, steady growth, clear information on corpus contents etc., the distribution of CD-ROMs does not compare with the network of clearing houses, which is proposed by NERC.

All the technology is in place for a user anywhere in Europe (or the rest of the world) to log in through a network to a corpus located in a clearing house; to carry out investigations and to bring over the results. Some considerations of cost and time argue that very large files should be sent physically and not electronically, and the experience so far of interchanging data and software suggests that advances should be made in small steps to avoid expensive failures.

Here is seen the need for as much common convention as possible, so that users can consult any language using essentially the same conventions. The same user identifications and systems of mailing and charging are further simplifications. Up till now corpus students and networkers have been mainly expert computer operators, patient and interested in the technical problems, like vintage car enthusiasts. It is the aim of NERC to make corpora available to all who wish to find out about language, and so the user interface must be simple, clear and cheap to run.

The current issues of availability are further discussed in (Clear, 1993, NERC-154) from a practical point of view.

References

Clear J. (1988): "Trawling the Language: Monitor Corpora", *ZüriLEX '86 Proceedings*, M. Snell-Hornby (ed.), 383-389, Tübingen, Francke.

Relevant NERC Papers

Castillo-Cabezas M. (1992): "Tool-set for the corpus processing in PC", University of Malaga, Working Paper, NERC-55.

Clear J. (1993): "Corpus Availability", COBUILD Birmingham, Working Paper, NERC-154.

Clear J. (1993): "Corpus Maintenance", COBUILD Birmingham, Working Paper, NERC-153.

Dendien J. (1991): "Acces a l'information dans une base textuelle Functions d'acces et index optimaux", INALF Paris, Working Paper, NERC-113.

James Z. (1992): "Cobuild use of the UNIX Operating system", COBUILD Birmingham, Working Paper, NERC-89.

James Z. (1992): "Enhancement of standard facilities (eg annotation tools)", COBUILD Birmingham, Working Paper, NERC-130.

Johns T. (1992): "Corpus Linguistics in a PC-compatible environment", COBUILD Birmingham, Working Paper, NERC-90.

Krishnamurthy R. (1992): "Basic access software, section A: word lists", COBUILD Birmingham, Working Paper, NERC-87.

Krishnamurthy R. (1992): "Basic access software, section B: basic concordancing 2: the user's perspective", COBUILD Birmingham, Working Paper, NERC-88.

Lane T. (1992): "Data input", COBUILD Birmingham, Working Paper, NERC-56.

Lane T. (1992): "Basic access software, section B: basic concordancing 1", COBUILD Birmingham, Working Paper, NERC-101.

Lane T. (1992): "Access and Management Software Tools. Task: Development Routines", COBUILD Birmingham, Working Paper, NERC-114.

Monachini M., Picchi E. (1992): "Tagged corpora: A query system", F. Kiefer, G. Kiss and J. Pajzs (eds.), in *Proceedings of the 2nd International Conference on Computational Lexicography, COMPLEX 92*, Budapest, Hungarian Academy of Sciences, NERC-105.

Paris Group (1992): "FRANTEXT", Working Paper, CNRS Paris, NERC-178.

Picchi E. (1991): "DBT: Data Base Testuale", ILC Pisa, Working Paper, NERC-9.

Picchi E. (1991): "DBT: a textual data base system", *Linguistica Computazionale*, 7: 2, pp. 177-205, NERC-11.

van der Kamp P. (1992): "Proposals for standard to access corpora", INL Leiden, Working Paper,

NERC-85.

van der Kamp P., Boom A., van der Voort van der Kleij J. (1992): "Overview of and experiences with hard- and software used at the Institute for Dutch Lexicology", INL Leiden, Working Paper, NERC-51.

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[10].

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.

---

10 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).

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 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[11]. 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:

-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[12] resources, standardization

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

12 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).

in this field has become an issue of vital importance[13].

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.

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

13 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).

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[14]:

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

[14] 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).

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[15] (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
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

[15] 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.

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, pronc
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

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.

	Pc	BNC	Go	Br	LOB	La	SUC	It	PrS	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

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).

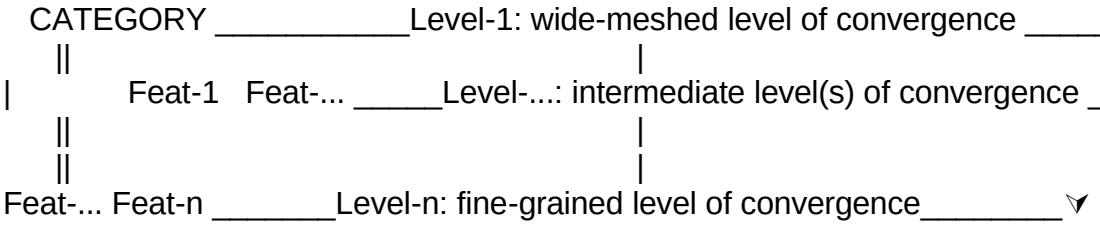
3.4.2 *Consensual Categories*

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)[16]



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.

3.4.4 *Transduction between Existing Tagsets and the Proposed Scheme*

16 It is worth reminding that the features of a lower level add new information.

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 for

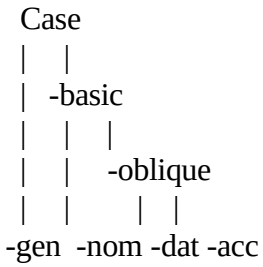
Nouns; in the Romance languages there is the feminine, the masculine and often the common gender, while the Scandinavian languages have the genders neutrum and utrum for Nouns. It was decided to leave out the values `common' and 'unmarked' from the proposed set of shareable values, since it can be seen 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:



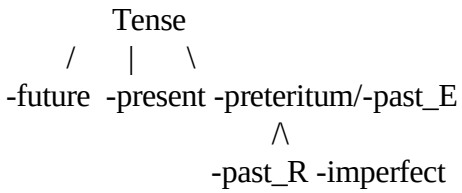
It must be stressed that each language system will use its own appropriate set of values. For the Noun category:
-German: `nominative', `genitive', `accusative' and `dative'
-Dutch and Scandinavian languages: `basic' and `genitive'
-English: `basic', `genitive'. The value `genitive' refers to the Saxon 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[17] 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*. Stockholm.

[17] 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.

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.

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.

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.

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).

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.

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.

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.

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 recognized by different theories and for which there is substantial agreement about the boundaries. The disagreement is related instead to the internal structure of the noun phrase. It is therefore reasonable, as Leech affirms, "for a syntactic annotation scheme to distinguish the boundaries of the noun phrase without being too much concerned about its constituency". Skeletal parsing goes in this direction, and therefore can be seen as a possible answer to the theory neutrality requirement.

Skeletal parsing is also connected with the training process of stochastic grammars. As can be noticed by examining our sample of syntactic annotation schemes, variable degrees of granularity of linguistic information can be added to a raw corpus. The delicacy of the analysis should not be seen as a value in itself; instead, the more granulated the analysis the scheme offers, the larger the corpora that are required in training stochastic grammars. Therefore, the tendency to adopt more granulated analysis schemes is now being reversed at all linguistic levels of description (i.e. there is a move from more detailed annotation schemes to more simplified 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 eliminate sources of error, inconsistency, and uncertainty in annotating, and increases the speed of both 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).

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.

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".

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.

5 Annotation beyond the Syntactic

5.1 *General introduction*

It is clear from the earlier sections dealing with annotation that the job of assigning meaningful linguistic category labels to portions of texts is not easy. Even the simplest disambiguation routines give odd results in some cases, and the experience of annotating a text tends to turn the spotlight on the analytical framework, and call it into question. Few if any annotation schemes have ended as they began.

This is healthy in a young subject. The chapter on annotation tools argues that we must cultivate a flexible approach to the whole business of annotation, as corpus evidence is gradually built in to our work. Starting out with conventional word classes and syntactic categories, the computers are trained in how to recognise them: in this process new distinctions and groupings suggest themselves; if they are confirmed then the nature of the annotation system has changed, and the whole system has to be updated.

Structural annotation up to the level of sentence is largely the automation of a well understood process. The picture changes when we move beyond the well-known fields of syntax, morphology and phonetics; in areas like semantics, discourse and pragmatics there is no long text-based tradition of analysis. The categories and the data are much further apart, and there is even less consensus about what constitutes an adequate method, or acceptable results.

The categories of semantics are abstract, and are not readily related to any physical entities. Conventionally the word is the unit most central to semantics, but the importance of phrases in making meaning is being seen as a quickly growing area, and one that will be fuelled by corpus study. The environment of a word helps in disambiguation, which is one of the principal aims of semantic annotation. A disambiguated text provides powerful information as input to other applications of language technology. Hence it is worthwhile to try to automate semantic annotation.

In the case of discourse annotation, there is another dimension of difficulty in the alignment of category and text - that of the size of the unit of analysis. The physical events of language are small in scale - short disturbances of the air or tiny marks on a page: but these are combined and recombined to form the structural units of syntax. Each step up in size adds a lot of complexity. Discourse deals in large units, like sentences. The number of different sentences is unlimited, and they range in size from one or two words to hundreds. This makes it exceptionally difficult to relate the analytical categories to stretches of the text.

With pragmatics there is a further problem, because pragmatic meaning seems to depend on the precise transaction that is taking place, and the interpretation of the people involved. It is only in part associated with words and phrases, and seems to relate data and meaning in an extremely complex and obscure way.

Despite the inherent difficulties in these areas, there is growing interest in solving the linguistic and computational problems involved in adding them to the annotation systems already available. Since NERC is a planning and feasibility study, it is important that, however nebulous the state of research may be, provision is made for developments of this kind.

5.2 *Semantic annotation in text corpora*

Semantic annotation developed later than other kinds of annotation (morphosyntactic, syntactic etc.). As a result, there are not yet many publications in the field. Also, it is not yet very clear what the semantic annotation operation should consist of. Leech presents it as the "obvious next step" from grammatical and syntactic tagging (Leech, 1992a) and, as an example of its uses, says: "Semantic word tagging can be designed with the limited (...) goal of distinguishing the lexicographic senses of the same word". Martin also attempts to formulate a definition (Bon and Martin, 1992, NERC-70): "Within the process of semantic tagging context information leads to knowledge about semantic values of objects in the text". In practice, in the projects analyzed within the NERC study (see Spanu, 1992, NERC-66), semantic annotation deals with the attribution to a word of a label which refers to a semantic feature (also called content or conceptual category, or semantic class, or other names) which belongs to a more or less rich and structured set of semantic features.

We examined the ways in which semantic annotation is used in different projects, analyzing, where possible, the theoretical motivations underlying the practical choices, and trying to identify the problems posed by the semantic annotation process. This was done as a preliminary effort, pending further study aimed at examining the feasibility of proposing any minimal standard at the level of semantic description.

The survey pointed out clearly that there has been a surge of activity in the last year, with announcements of several new sponsored projects in this area, in Europe, in US, and in Japan. Among the others we can mention ET-10/51, on extracting semantic information from a corpus based dictionary; LRE Delis, on developing descriptive lexical specifications and tools for corpus-based lexicon-building; Acquilex-II, on the construction of a substantial multilingual lexical knowledge base starting from lexical and textual sources; and Hector on tools for corpus semantic annotations and meaning analysis in a Fillmore approach (Fillmore, 1992).

In (Spanu, 1992, NERC-66) a survey of commonly used annotation tagsets was performed; this brief review permits a first sketch of the situation of semantic annotation, at least in the projects examined.

Semantic annotation has been employed in the framework of different projects with, among others, the following goals:

-automatic content analysis of spoken discourse;
-analysis of word associations;
-disambiguation of word senses;
-feasibility study of a rule-based semantic tagger.

Furthermore, a semantic tagging procedure for trees is employed in the framework of a Generalized Probabilistic Semantic Model for semantic preference assignment.

We must point out that semantic annotation, while playing a very central role in all the projects considered, is always associated with a previously accomplished syntactic tagging process. Even though the semantic annotation procedure is not used independently from other annotation procedures, its centrality - and the growing interest in it - demonstrates the importance of defining more precisely the semantic annotation procedure of a text, keeping in mind the projected area of application.

There is one point which clearly emerges from this brief survey: the choice of semantic features is

seldom supported by a theoretical background, unlike in a number of lexical projects reviewed in a MULTILEX paper (see Spanu, 1992) which was produced in order to make a comparison of the semantic features used with a view to evaluating the possibility of defining a minimal set of shared semantic tags. Rather, it is determined by practical needs and by the problems encountered during the working phase. This is true above all for those research projects where a set of domain dependent features is used.

Another criterion for feature selection is linked to the necessity of having a set of features answering to requirements which can vary from exhaustiveness and completeness to generality. In each of these cases, a rather general level of subcategorization is given, but it can be rendered more fine-grained if necessary.

5.2.1 *Standardization of annotation tagsets*

It is premature to think about the possibility of proposing standardization in semantic annotation. The studies on it have not yet been deeply explored, and above all there is a lack of theoretical underpinning. Nevertheless, a first attempt towards a minimal standardisation might be desirable in the near future, and should be feasible at least as far as the more general semantic concepts are concerned (see also Spanu, 1992, for a discussion of semantic tags in lexical projects).

5.2.2 *Future directions for research*

The late development of semantic annotation as compared with other kinds of annotation, and the very few examples of its use, are an obstacle to an exact evaluation of the state-of-the-art. But precisely because of the work in this field is so scarce, it might be possible to lay down minimal guidelines that could inspire future research. It is therefore very important to look carefully at the short-term results of recently started projects which make use of a semantic annotation procedure.

In this way a high level of homogeneity in the production of semantically annotated corpora could be obtained at a lower cost than is necessary in order to obtain the same results in other levels of linguistic description of text corpora. For this purpose we can take into consideration the proposals already made in a general framework as well as in a more specific semantic framework, and look at them from the following angles:

-revision - in view of a semantic annotation - of the desiderata proposed by Leech (Leech, 1992) (speed, consistency, accuracy etc.) in order to evaluate them in the framework of semantic annotation;

-an attempt to avoid subjectivity in the choice and assignment of semantic categories, for example by employing a reliable categorisation; this is particularly desirable in the framework of lexicographic research (Sinclair, 1991, NERC-19) and also in the field of Natural Language Analysis (Sampson, 1989). To move in this direction, more work has to be done in comparing, on the one hand, existing semantic tagsets, and, on the other hand, the few minimal standards deriving from lexical projects.

5.3 ***Discourse and pragmatics***

In these areas we can only report on a few recent studies which show some potential for automatic or semi-automatic annotation. The vexed question of topic analysis is raised by (Sinclair, 1992b, NERC-134) and some pointers are given in two contrastive studies. The fundamental organisation of text, often called coherence, is also approached by (Sinclair, 1992a, NERC-133), and there is some prospect of at least partial automation when the underlying linguistic structure is better understood.

Channell (Channel, 1992, NERC-149) reports on an annotated database of pragmatic observations about English, which came about as a by-product of the Cobuild lexicography initiative. No place could be found in the published Dictionary for these thousands of observations, a fact which indicates how difficult it is to present material of this kind in a conventional manner. Equally it shows the beginnings of an annotation scheme which could aid the understanding of the interpretation of texts.

A major obstacle to progress in devising annotation schemes in this area is the incalculable effect of inference. Unless the mechanism of inference is accurately understood it will remain unpredictable and unlimited in its effects. Using arguments and examples from corpus study Tognini-Bonelli (Tognini-Bonelli, 1992, NERC-148) is devising ways of associating apparently inferential variables with structural choices in texts.

Although language is an exceedingly complex phenomenon, ways are emerging for it to be studied in new and penetrating ways, all involving the computer annotation of texts. The interaction between analysis and text will improve understanding, and that in turn will improve the next application of analysis. For the wider community of users, it is the analysis of meaning that is the main concern; syntax and morphology are seen simply as steps towards a generalised account of what a text means, and not as ends in themselves. The pursuit of research such as that reported here may have far reaching consequences.

References

Basili R., Pazienza M.T., Velardi P. (1992): "Computational Lexicons: the Neat Examples and the Odd Exemplars", in *Proceedings of ACL '92*, Trento, pp.96-103.

Chang J., Luo Y., Su K. (1992): "GPSM: A Generalized Probabilistic Semantic Model for Ambiguity Resolution", in *Proceedings of the 30th Annual Meeting of ACL*, University of Delaware, Delaware, pp.177-184.

Fillmore C.J., Atkins B.T. (1992): "Toward a Frame-Based Lexicon: The Semantics of RISK and its Neighbors", in Lehrer A. and Kittay E. (eds.) *Frames, Fields, and Contrasts*, Lawrence Erlbaum, Hillsdale, NJ.

Gale W., Church K.W., Yarowsky D. (1992): "Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs", in *Proceedings of the 30th Annual Meeting of ACL*, University of Delaware, Delaware, pp.249-256.

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

Leech G. (1992a): "Corpus Annotation Schemes", Paper presented at Corpus Workshop, Pisa 1992.

Roget's Thesaurus, (1975) Longman.

Sampson G. (1989): "SABRE - A Benchmark System for (English) Natural Language Analysis", Extracts from a proposal made to the Information Engineering Directorate.

Spanu A. (1992): "Proposal for a core set of basic semantic features for syntactic disambiguation", MULTILEX Report, Pisa.

Velardi P., Pazienza M.T., Fasolo M. (1991): "How to encode semantic knowledge: A method for meaning representation and computer-aided acquisition", in *Computational Linguistics*, 17, 2, pp. 153-170.

Wilson A. (1992): "Lancaster Semantic Tag Set (SEMTAGS)".

Wilson A., Rayson P. (1992): "The Automatic Content Analysis of Spoken Discourse". A Report on Work in Progress, to appear in Souter C. and Atwell E. (eds.) *Corpus Based Computational Linguistics* (working title), Amsterdam: Rodopi.

Yarowsky D. (1992): "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", in *Proceedings of COLING-92*, Nantes.

Yokota E. (1990): "How to Organize a Concept Dictionary", International Workshop on Electronic Dictionaries.

Relevant NERC Papers

Bon A., Martin W. (1992): "On Semantic Tagging. A feasibility study of the design and function of semantic taggers", INL Working Paper, NERC-70.

Channel J. (1992): "The coding and extraction of pragmatic information in a dictionary database", Working Paper, University of Birmingham, NERC-149.

Sinclair J.M. (1991): "The Automatic Analysis of Corpora, for the Nobel Symposium on Corpus Linguistics", Working Paper, University of Birmingham, NERC-19.

Sinclair J.M. (1992a): "Coherence in Text", Working Paper, University of Birmingham, NERC-133.

Sinclair J.M. (1992b): "The Analysis of Topic", Working Paper, University of Birmingham, NERC-134.

Spanu A. (1992): "Semantic Annotation in Text Corpora: short overview", Working paper, ILC Pisa, NERC-66.

Tognini-Bonelli E. (1992): "Use and functions of the adjective 'real'", Working Paper, University of Birmingham, NERC-148.

Chapter 6

Corpus Annotation Tools

1 Introduction

This is a forward-facing chapter. In the Report as a whole there is a clear emphasis placed on the creation and management of corpora which are much larger than those in general use today. At present only a few languages are served by corpora running to tens of millions of words (French, German, English, Italian, Dutch, for example), and only English for hundreds of millions (Bank of English). The vast majority of languages can only be studied using small corpora or even no corpus at all. However, we believe that these restrictions will fall away fairly rapidly as the value of large corpora is recognised, and that the corpus providers and users should plan to cope with corpora whose size will double every few years.

Further, it is envisaged that the acquisition of text evidence has no natural end point, and that text collections will enlarge until the state of a monitor corpus is reached - that is, when there is so much incoming data that there is little point in storing previous data in a raw form, but instead a priority is placed on the sophisticated processing of incoming text, and previous states of the corpus are stored in processed form.

The importance of size in a corpus must be outlined at this point. The accumulation of data which is on the surface highly repetitive must be supported by good reason, if possible by demonstration that there is information to be gained which is not obtainable elsewhere yet and which is essential to the understanding of language and the success of applications of that understanding to practical problems.

1.The distribution of word forms in a corpus is exceptionally uneven. Since (Zipf, 1935) it has been clear that the vast bulk of the vocabulary of a language has a low frequency of occurrence, and therefore requires an extensive amount of text material to capture an adequate number of occurrences. The so-called "grammatical words", on the other hand, tend to occur with great frequency, and although their relationship with other elements can be complex, there is soon far too much evidence for them - or at least that is the state of current perceptions.

2.Different occurrences of the same word show an exceptional variety of environmental patterning. The repetitions do not by any means confirm each other with any regularity. This arises for several reasons.

(a)Many words have several different meanings. Since the meaning of a word has a profound effect on its environment (Sinclair, 1991), it is not reasonable to expect the successive environments of a word to show any connection at all with each other.

(b)The word is not always the unit of meaning; a proportion of the patterning of a word may be its

contribution to idioms and compounds. This will have a double effect on the environment, by adding with some consistency other components of the idiom, and by setting an environment appropriate to the meaning of the idiom rather than any of its components.

(c)Texts and classes of texts have distributional patterns that affect the environment of individual words, even though there is not necessarily a perception of a radical shift of meaning.

(d)Each occurrence of each word has a particularity of usage that is appropriate to the precise conditions of the text, the occasion, the purposes of the speaker and writer and the meaning intended. It is thus distinguished in principle from any other occurrence of that word, and consequently the environment is likely to reflect this uniqueness in being unlikely to recur exactly ever again. Obviously the uniqueness is in turn dependent on a number of factors such as the extent of the environment, the type of occasion, etc. Nevertheless, one of the first lessons of corpus linguistics is the appreciation of the amazing variety of usage within broadly repetitive frameworks.

These reasons, particularly the last, provide a powerful argument for large corpora. Between the peculiarity of a single instance and the broad sweep of useful linguistic generalisations lies a network of relationships which are largely undescribed, because they are neither centrally grammatical nor centrally semantic.

Point (c) above hints at an extra dimension of organisation which is required in the processing of large corpora. Layers of classification of texts will be necessary to make the best use of the corpora. In the first instance these will no doubt follow the lines of the design parameters (see Chapter 2), enhanced by internal typology studies such as those of (Biber, 1988 and Nakamura, 1992, NERC-137). Users will often require specific sub-corpora to be constructed for a job, and some classes of texts may be deemed unsuitable for some operations. Further, it is reasonable to foresee the convergence of researchers who concentrate on severely restricted sub-corpora of specialised varieties of a language, with those who begin with a comprehensive review of the characteristic patterns of the language.

It is a formidable job to design software tools to exploit the new corpora. The size of the data store is so great that only fully automatic tools can be contemplated, and the patterns to be retrieved are not clearly delineated in research so far. It follows that the tools proposed here are exploratory, and the list is provisional. As the early ones are developed and applied, the first results will help future targeting. The background paper to this chapter is (Sinclair, 1992, NERC-19).

2 Current software tools: Parsers etc.

One of the most valuable prospects for the use of corpora is in the evaluation of analytic tools. Parsers can in principle be compared by running them over a corpus in competition with each other. Software under development can be improved by testing against different corpora.

There is a clear move in computational linguistics to establish standards, targets and other criteria in the corpus area (e.g. the EAGLES project). The use of corpora in evaluation will increase our understanding of the characteristics of corpora, while at the same time acting as increasingly sensitive test-beds for the software of the future.

At present the differences in performance and goals is so great that realistic competition is unlikely to be achievable. Chapter 5 offers a thorough description of the categories and results of the leading taggers for most of the European languages and of parsers for English, from which it can be seen that some degree of comparison can in any case be achieved. At the level of tagging, the comparison, in spite of having to take many variables into account, is quite feasible, at least for a core set of features. At the level of syntactic analysis, the comparison - because of the number and nature of the variables and of their interrelationships - is restricted to the definition of a range of parameters with respect to which a classification of various parsing schemes can be attempted. From the information presented in Chapter 5, it is possible to match a parser with a particular job, and to see the characteristics of the main types available. Each has some important features, whether fine discrimination of categories or the proportion of automatic routines to manual, or the complexity of processing.

It is very likely that the performance of the existing parsers will improve with the corpus experience that they are currently getting. The Helsinki parser (Karlson, 1990) has begun the mammoth task of parsing the Bank of English; the Oxford-Longman-Lancaster corpus is being manually tagged in Lancaster, with machine participation; the TOSCA (Oostdijk, 1988a, 1988b) is parsing the ICE corpus (Greenbaum, 1990). There are substantial issues of pre-editing, handling of spoken transcriptions, dealing with output etc., with the likelihood of useful feedback to the Text Encoding Initiative.

Because of the models of language that lie behind these systems, success can only be self-determined. An error of automatic analysis arises when the machine makes a decision that the human researcher disagrees with, usually when the analysis does not correspond with the dictates of a pre-existing grammar. Rather than the corpus being used to test the grammar, it is as if the grammar (which was written before any corpus experience was obtained) is being used to test the corpus.

There are signs that some revisions are being made with corpus experience; decisions made in advance are found to be less effective than patterns suggested by the close study of the data. For example, (Marcus and Santorini, 1992) report that the usual division of the word *to* into preposition and infinitive marker is not supported by the experience of the Penn Tree Bank Project, which is a heavily automated and hand finished parser. This confirms the direction of thinking in corpus grammar, that the most frequent words in a language may best be treated as unique word classes and not squashed into word classes which are set up for the mass of regular vocabulary (Sinclair, 1991).

The grammars, then, will gradually adapt to the corpora. But what about the structural patterns that are noted in corpus studies but that do not fit the usual kind of grammar? What if there are form classes that conventional linguistics has just not noticed? What is ultimately the best way of handling the very frequent words automatically? (They form the majority of running words in most texts). It is clear that in addition to the job of tuning and automating existing parsers etc. there must be new initiatives.

3 Current software tools: Applications

In addition to the taggers and parsers, there are a number of software routines which are designed to operate on a large corpus and provide output that is part of an actual or potential application. The power of the statistical analysis of large corpora is impressive. Pioneering work in speech recognition, devising self-organising models for predicting the language that is likely to follow a certain state-of-text (Jelinek, 1985a, 1985b) has been followed by a growing range of imaginative applications.

There are simple tools for lexical analysis, such as those set out in (Church and Hanks, 1989), and for semantic discrimination (Yarowsky, 1992). There are statistical tools for aligning parallel texts in different languages (Church and Gale, 1991) and for automatic translation (Della Pietra, 1992). These have some of the essential characteristics of the new generation software. They are indifferent to the length of the texts on which they are operating; the design of the software is of a throughout process that continues as long as there is text left to process. Some are organised to accept corpus input in big batches - of a few million words at a time - so that baseline statistics can be established.

They operate on "raw" text - usually text that is being reused. Because they need a great deal of text, and sometimes a considerable variety of it, they cannot make heavy demands on input configurations, or expect pre-editing processes to take place.

They can handle unusual or corrupt text with reasonable efficiency; that is to say, they are robust. If they encounter something unexpected in the text they do not break down or suspend operations; they recover a working equilibrium quickly. This means that they can be put to work on large quantities of text of unknown origin with reasonable confidence that they will perform their functions with an acceptable accuracy, at very high speed and without the need to monitor.

It follows that they consist mainly of essentially simple and statistical routines at the present time. Unlike the internal complexity of a parser, where interlocking levels depend on each other, these tools are fairly straightforward. No-one disputes the need for a lot of statistical analysis with long texts, either as an end in itself or as input to other processes.

The motto of leading researchers in this area is "There's no data like more data!" (Liberman, oral contribution, Pisa Workshop on Textual Corpora, January 1992). The reliability of statistical work is improved by increasing the total amount of data processed. For many fundamental patterns of distribution, the size of the text collection is more important than where it all comes from.

This work is performed on whatever text is available, and is thus dependent on social and institutional factors that determine what texts are put in electronic form, and can then be made public, and even put into the public domain. Up till recently there has been a preponderance of legislative and legalistic prose, and texts of unusual structure like dictionaries and catalogues. 1992 saw the publication of some newspapers in electronic form, enabling a better balance to be achieved, and as it

becomes more and more standard practice to make and keep an electronic version of all text, and to transcribe speech at least semi-automatically, the representativeness of a comprehensive collection of everything available will improve.

The interests of researchers in the "more data" school and the providers of major corpora are on a course of convergence. For the former, as the tools for statistical analysis become more sophisticated and the research tasks more discriminating, there will be pressure for text material to be of known provenance, and for the quantities of it to have some proportional justification.

This is already an important area in collocational studies. Collocations, in their simplest state, are statistically significant associations between pairs of words. Because of the way words are distributed, any one collocation will tend to cluster in a particular set of texts, probably in one or more particular genres. Hence the constitution of a corpus has a direct effect on the pairs of words which will be identified as collocations. If one text is replaced by another, the idiolectal collocations will presumably also be replaced by the individual collocations of the new author. Assuming the two texts are from the same genre, there should be only a slight effect if any on the broader collocational picture. A mass of texts from the same genre will prioritise collocations from that genre and obscure collocations from other genres.

The growing diversity of reusable data will enable such problems to be overcome. Meanwhile the movement, already noted several times in this Report, towards the creation of monitor corpora, brings the corpus providers into line with the most voracious of the "more data" school. For the task of the management of a monitor corpus is to keep relevant data flowing in (and through) the machines in proportions determined by the design, making as much use as possible of what is available.

The distinction between parsers and other analysis tools is also ripe for scrutiny. In contrast to the new software advocated here, this distinction presupposes that it is possible to separate the major lines of syntactic structural classification from all the other patterns that are capable of extraction from the text. Whereas "applications" such as machine translation are likely to remain considered as applications for a very long time to come, the claims of collocational studies to be received as linguistic-structural, ultimately on a par with syntax, are becoming stronger as the evidence grows.

4 New corpus software tools: grammatical analysis

The remarkable growth of interest in studying corpora is a clear indication that they are providing new kinds of information to a lot of researchers and applied language scientists. Corpora do not just offer a string of sample sentences on which existing hypotheses and coding systems can be tested, and with which they can be refined. That they do, but the trade would not need anything like the amount of text material proposed for such contained goals. They do not just offer language fodder of an indeterminate kind for gross statistical analysis. For that they would not need the care taken in their design, the efforts made to gather quantities of informal conversation and ephemera, their classification internally and the time-consuming business of mark-up to facilitate retrieval.

The new corpora suggest a new kind of enquiry into the nature and structure of language -not one

where the main aim is confirmation of what is already fairly well agreed, but one where the exploration is likely to uncover facts about language and languages that have not been available before. What kinds of structures emerge from the data with the minimum of preconceptions?

There must be, of course, preconceptions in any scientific study, and linguistics is no exception. Indeed it is of all sciences probably the one most determined by preconceptions, in the way it is most characteristically practised. For many years now the evidence for linguistics has mainly come from the introspection of the native speakers of a language, or of competent non-native speakers. Grammars built from these insights are projected onto corpora, and the discrepancies between prediction and actuality are dealt with by amending the grammar or correcting the analysis as a mopping-up operation.

"Minimal Assumptions" is a phrase which is attracting some attention in corpus computing. The successes and weaknesses of "maximal assumption" analysis of language are fairly well known; what about returning to a stage of conceptual innnocence and approaching the description of a corpus by tracking down what appear to be repeated patterns of any kind?

4.1 *Lemmatisation*

At the earliest relevant stage, the computer regards language text as a linear string of characters, chosen from an approved set. One character has special significance attached to it - the space character is given as a word boundary. This is one of the very basic assumptions. Then punctuation characters are identified, and numerals and any layout codes. Then the computer is informed that certain groups of words are actually variant forms of the same WORD. In English most of the forms of a WORD have a lot in common with each other - all but one or two letters at the end, usually. A WORD is called a lemma. The assumption is that the forms of a lemma differ only because of their gross syntactic environment, and that the same meaning persists in the morpheme despite the superficial change in form. In other words, a fairly large assumption is made regarding the persistence of meaning through changes in form.

Further, while the regularities of lemmatisation can be expressed for most languages by a small set of rules, there are many exceptions, and of these many are among the commonest words. No computer would decide that English *good*, *better* and *best* are instances of the same event unless it was arbitrarily instructed to do so, nor Italian *andare* and *vado*. But more important for the long term, early studies of word forms in a corpus suggest that the relationship between meaning and word form is not at all straightforward (Council of Europe, 1991).

The whole of grammar, semantics and beyond is built up on the notion of a stable lemma. Corpus work is signalling that the notion of a lemma is problematic. Experience of automatic lemmatisation, at least for English, shows the need for a degree of arbitrariness which is barely acceptable. There is clearly a need to develop automatic routines in this area to underpin future work.

4.2 *Tagging*

It is assumed throughout this section, the previous one and those to follow that the software routines for corpora in the many millions will operate entirely automatically. The cost in expert time and money of preparing by hand, monitoring the processes or checking and correcting at the end will be prohibitive. Furthermore, such emendations to what the computer reports introduce an uncontrolled factor into the process, thus reducing the scientific value of the work (though it may be acceptable with small corpora for some applications).

The inevitable restriction to what is retrievable by automatic means will initially lead to an apparent dip in the quality of analysis provided. There are well known areas of ambiguity in which computers have great difficulty in matching human judgments, and statistical methods are frequently used in the resolution of doubtful circumstances. Large strides will need to be made in software before the new tools will achieve the utility value of present tools, or even basic credibility.

Since there is ultimately no other way, it is possible to see such a restriction as a step forward. The reason that the performance of the computer is currently disappointing is because it is being compared directly with a human being, who has access to an indefinitely large store of relevant information which is denied to the computer. The computer can only report on the state of the text. Once ingenuity has been exhausted, we have to assume that the human is able to make distinctions and classifications which are actually impossible to do on the basis of text evidence alone. There is thus an important boundary drawn, between what information is in principle deducible from the text, and what is not.

In addition to familiar analyses redone with this restriction, it is more than likely that new types of analysis will suggest themselves. These will align closely with the text evidence and will be simple to identify and to process further. New insights into linguistics may well follow.

In the area of word class tagging there are grave problems in replicating anything like the received classifications of English. The old idea of a small number of major word classes with a few secondary ones has few adherents in the field of computerised tagging; most of the well known systems require a large number of basic tags. Progress is made by interaction between the abilities of the machine, the results of passes through text samples, and the specifications of the tagging exercise (Garside et al., 1987).

The general kind of specification is that each occurrence of each word in a text should be supplied with an intuitively satisfying word class tag, unambiguously. The state of the art is reported in Chapter 5 of this Report (see also Federici and Pirrelli, 1991, NERC-20 and Saba et al., 1991, NERC-22).

It should be noted that tagging is a kind of partial parsing, in that it is sensitive to the precise textual position of an occurrence. Because many words in English may realise more than one word class, tagging is not usually seen as the simple matching of a set of words with a set of tags. Rather, it is a process which evaluates a textual position and on that basis selects from the available alternatives the most likely or appropriate tag. Probabilities are often invoked, because the textual position does not give sufficient discrimination for a clear choice to be made. (Bindi et al., 1991, NERC-103) open up a statistical approach, starting from a tagger.

Tagging has value both as an end in itself and as a stage in more complex analyses. It also has potential value in applications. As an end in itself, it powers a lot of studies of the relative frequency of word classes, and of words in the classes; and it opens up the study of phraseology, including idioms and other fixed or fairly-fixed phrases. A system able to disambiguate occurrences of *that* and *to*, two of the very commonest English words, would speed up the identification of idioms. *Mean* adjective, verb and noun are three different lemmas, and *means* is a member of the verb and the noun lemmas, but is also a separate noun lemma on its own, capable of differentiation by a tagger because of the different number choices of the two nouns. As a stage in a complex analysis, tagging is normally seen as the first part of a parsing strategy, a platform from which the higher structures are shaped. The word classes are the building blocks of the syntactic categories. While this is not the only possible way of constructing a parser, it is the time-honoured one. Its lineage goes back to traditional grammar, pivoting around the word. The concept of word which is used goes back to the grammars of the classical languages, where it is simultaneously the lemma, preserving meaning through changes of declension and conjugation, and the unit of phrase, clause and sentence syntax.

The output of a tagger is felt to be of value in practical, including commercial, applications of language analysis. The ability of a computer system to move from dealing with unclassified words to nouns, verbs etc. should improve the efficiency of spelling checkers, speech recognisers, translation aids, thesauri and style guides, to name but a few. Tagged text will help in the identification of technical terms, particularly multi-word terms (Yang, 1986).

As with lemmatisation, there is a "minimal assumption" approach to tagging, as yet largely unexploited. In such an approach, the same tag would apply to each occurrence of a word, so that tagging was strictly separated from parsing. A word such as *dwell* in English is always a verb, and *duvet* is always a noun. Alphabetically in between them is *dwarf*, which can be used as verb or noun. That third type is a common feature of English, often remarked on in books about the structure and history of the language. Hundreds of words in English occur in verbal positions in syntax, and with verb inflections, and also in noun positions, with inflections where they are count nouns. Only tradition prevents us recognising them as a word class distinct from verb and noun, for which we do not have a technical term such as "norb". It is possible to build up a different set of tags using arguments like this (see Francis, 1993, NERC-176). The tag set for English is not likely to be very large, and there will only be a few oddities such as *like* or *round*, which seem to fit into most of the traditional word classes.

A text tagged with such a tagger will not be as informative as one that evaluates each textual position, assigning each "norb" to noun or verb. It will not be as satisfying to those who wish to replicate their intuitive judgments, whatever the difficulty from a machine perspective. It will not act efficiently as a front end to a conventional parser. But it will yield statistical results about the language which are every bit as interesting as the conventional ones; it may well inspire new approaches to parsing, keeping closer to the textual patterns; it will be simpler to use and more easy to compare languages with than a tagger dedicated to one language only. The relation between such a tagger and a conventional one will be worth studying, to see how often the composite nature of a tag like "norb" will cause real problems.

Most of this line of argument will be familiar to students in this field, since typically a tagger starts with a dictionary that lists the (conventional) word classes available to a particular word, before going on to set out the conditions under which one will be preferred to another. The only deviation

proposed here is that a composite tag is assigned before the environmental conditions are investigated.

This example of a minimal assumption tagger brings into focus one characteristic of the software tools of the future. By following the literal textual discrimination of English, the tagger requires a tag which ranges across both noun and verb. In an application of the tagger as an analytic device, there might be occasions where the analysis asks for greater discrimination. In this instance, is the word acting as a noun or a verb? Sometimes the state of the text will provide a clear answer, and at other times it will not. The hierarchical relation of "norb" to noun and verb is then a source of strength and flexibility; the system is not forced to make a decision for which there may not exist enough evidence in the text; but the tag "norb" can be recorded and the analysis can proceed. A tagger that insisted on discrimination in such circumstances would be stuck, or would have to use a fairly blunt statistical instrument to impose a fairly arbitrary decision, running the risk of introducing a mistake at an early stage of analysis.

For many years to come it is likely that the state of a text will not always be able to provide the evidence that an analyst needs for maximum discrimination. Perhaps it is a permanent condition. It seems then good policy to design analytical tools so that there are several layers of fall-back in case it is not possible to discriminate.

4.3 *Parsing*

The current state of corpus parsing schemes, both automatic and assisted, is set out in Chapter 5. The movement of corpus studies is in two directions from the current state.

(a)As mentioned above, the efficiency of existing parsers will be improved by experience of corpus analysis. However it is unlikely that the output of different parsers, constructed on different principles and serving different goals, will converge, so that they will be comparable in any strict sense. As with tagging output, there will be a variety of applications for well parsed text.

(b)New kinds of parsers will be developed for the particular circumstances of large corpora. Parsers are notoriously fussy about the text they require as input; a corpus parser has to cope with the inherent unreliability of text in the real world. Parsers with different specifications derived from work on lemmatisation and tagging (see above) will no doubt adopt different strategies and come up with new results.

Because of the lack of uniformity about the precise aims of parsing, the whole concept of parsing will be reinterpreted for future corpus work. The various stages will be isolated and arranged to operate independently of each other, linked by clear input and output specifications. A large number of existing and anticipated tools will be grouped together as "partial parsers", and integrated into the toolkit. This atomisation of parsing is necessary partly because of the complexity and inherent difficulty of the job;

partly because of the extra information available in a corpus, and partly because of the variety of user needs that can be anticipated. High on the list of user needs is adaptability to the parsing of sub-corpora, for which considerable flexibility is required.

The algorithm that (Yang, 1986) proposes is a partial parser of one kind; given any text (but preferably a scientific one) the device will search for structures which match a mini-grammar. Those that do are putative multi-word technical terms.

The work of Coniam (see Coniam, 1991, NERC-21) offers another model of partial parsing. Coniam's "Boundary Marker" examines all the word spaces in a text and evaluates them with respect to the two words on either side, to see which indicate boundaries of higher units than the word. The study covers three lines of enquiry:

(i)How robust can a parser be? The boundary marker consults only a three-word window; hence if it fails or makes a mistake it will pick up quickly because of forgetting previous states of text almost immediately.

(ii)What contribution does a sound boundary marker make to a full parse? In many cases, an indication of word class through a tagger and an indication of boundary may be enough to assign clause structure.

(iii)Are the conventional units above the word confirmed by the text evidence? Are there any other putative units suggested? Are the beginings and ends of units clearly marked?

This work is all part of an efficiency drive, to see how quickly parsing can be done. It has been notoriously slow in operation, even without counting human intervention. The large corpora have estimates of about 0.5m words per working day, which at today's speed of acquisition does not keep up with input.

Special parsers for sublanguages are popular at present, because sublanguages hold out hope of good results in the shorter term. Given the great complexity of human language, and the open-endedness of texts, it is necessary to plan over a long period to achieve improvements. But many applications require mastery over only small corpora of severely restricted language. For those, it is sensible to develop partial parsers in yet another sense - parsers which are comprehensive for the texts they parse, but good for only a strictly defined subset of the language. The reduced number of options, it is held, makes the grammar essentially simpler, easier to write, cheaper to develop and neater to incorporate into some system.

An instance of this is the parser being written for the ET10 project 51 on semantic analysis (First and Second Reports, submitted to the Commission of the European Communities, 1992). The project begins with an analysis of the defining style of the *Collins Cobuild Student's Dictionary* (Sinclair, 1990b), which uses ordinary English sentences but a highly restricted syntax, and a vocabulary which is also fairly specific to the task of defining. Instances of a defining word having more than one meaning are unusual, and there are a number of words which have specialised uses in definition.

The job of explaining meaning is very particular, and the grammar of the definitions can be written specially for greater efficiency. Nouns and verbs, subjects and objects are not needed in the first

instance, because there are more immediate structural categories - the definiendum, or headword; its co-text, sometimes on either side of it; the superordinate and discriminator of the definiens. Within the last of these we encounter "normal" grammar.

The simplicity of the parser is remarkable, and its speed of operation is impressive. This is because it is designed for one kind of language only and it only uses those scraps of conventional grammar that it needs.

It is clear that the techniques of parsing are becoming diversified as they shape up to the challenge of corpora. They are opening up to the information contained in corpora and they are continuing with the eclecticism that has characterised them so far.

5 Lexical tools

The chief driving force towards very large corpora was the argument of lexical analysis, summarised at the beginning of this chapter. With suitable tools, the lexical structure of a language could be worked out in parallel with the grammatical structure. Obvious lexical structures are collocations, compounds and idioms; all of them problematic to identify, to demarcate, to relate to other structures and meaning.

Although basic collocational software has been in use for many years (e.g. Reed, 1977), it has remained fairly basic, probably because of the small corpora that were available until recently. The opportunity now arises for major developments in this field of study.

A compound is a restricted type of collocation, worthy of special treatment because the two elements together are held to constitute a single lexical item. The introduction of the results of a study of compounds into a text would lead to a recalculation of the other collocational patterns, and a greater accuracy of result. Work on the automatic identification of compounds is part of the AVIATOR project (Blackwell, forthcoming; Collier, forthcoming; Renouf, forthcoming), and (Yang, 1986) is also relevant here, since one major class of technical terms consists of compounds.

Idioms and like phrases are particularly difficult to study with today's corpora and computers. Very little is known about their formal characteristics, and not much about their place in linguistic structure. They have been on the periphery of linguistic vision because they do not comply with the regular rules of construction - or if they do they do not deliver the expected meaning.

The main computational problem is identifying them amid all the variation that surrounds them. Researchers disagree about their constitution, and suitable computer models of their construction have yet to be devised. Osborne (see Osborne, 1993, NERC-175) has made a start by providing a basic tool which can be used to assess the problem as a whole. The algorithm is a fuzzy matching one, allowing for a wide range of variation, but not variable sequence. Further refinements and additional tools can be expected in this area.

One potentially fruitful enquiry is the collocational behaviour of the frequent grammatical words. By the simple device of ignoring or merging the less frequent words, a notion of "frames" emerges; because of the great frequency of the grammatical words, long repeated strings can be identified

(Renouf and Sinclair, 1991).

One important characteristic of the lexical tools is that they operate directly on character strings, and are thus independent of any particular language. This increases the value of investment in such tools, as compared with the language-specific nature of most lemmatisation, tagging and parsing. They will also facilitate language comparisons.

The concept of a lexical "parse" needs to be given substance. The tools indicated above, and no doubt others, would be used on a corpus to identify the units of lexical meaning and the ways in which they fit together. Another dimension of linguistic patterning would be harnessed towards understanding the organisation of language.

Applications of the lexical tools include such tasks as selecting examples of words in context according to given criteria; disambiguating words on the basis that each distinct meaning of a word is likely to attract a distinct group of collocates; contributing to the automation of translation.

6 Lexicogrammar

Grammar is gradually becoming lexicogrammar. The separation of grammar and lexis is being seen as artificial and unnecessary. Corpus linguistics is not the only force in this movement (see Halliday, 1985), but it is certainly involved. There is a mass of lexis to get through before the outlines of grammar can be perceived, and the corpus provides a new and exciting view of grammar (Francis, forthcoming) on the way. Both in moving from the detail to the generalities, and in the reverse direction, the likelihood of strong lexicogrammatical regularities becomes strong. A somewhat neglected concept, the lexical set (Halliday, 1966) may be revived as an in-between category for groups of words and/or phrases which have a status simultaneously in the lexical and the grammatical structure of the language.

Sophisticated software will be required to track down the sets of a language, and most of the other software described here will need to be used in the hunt. It is impossible to say what proportion of the structure of a language can be accounted for in purely lexicogrammatical terms, and what has to be left for separate analysis by grammar and lexis. The *Collins Cobuild English Grammar* (Sinclair, 1990a) gives a number of examples of possible sets.

The key to lexicogrammar seems to be the ability of a corpus to push grammatical patterns down the scale of delicacy - to get beyond the massive primary distinctions of grammar like past and present, singular and plural, and into the intricate web of subcategories. There, where it is essential to have a corpus for guidance because the intuition fails, the groupings of grammar and lexis begin to overlap.

Applications of lexicogrammar include all that has been put forward for both lexis and grammar. Additionally, there will be contributions to text typology. The internal patterns of a text can be used to aid in its typological classification, and the work of (Biber, 1988 and Nakamura, 1992, NERC-137) can be cited as early evidence of the impressive results that can be obtained. A more comprehensive picture wil be obtained when lexicogrammatical criteria are available.

Lexicogrammar will also be used in fundamental research on such issues as the fundamental classification criteria that can be used in linguistics; on the value and role of statistics in building up a linguistic model of a language; on the accuracy of beliefs such as that language is well described within an item-and-arrangement framework. Each of these issues will probably involve a major effort in corpus study.

7 Multilingual software

It has been mentioned several times in this chapter that certain tools are language-independent. Also the notion of software that makes the minimal possible assumptions about the specific structure of a language has been aired on several occasions. Both of these directions are favourable to the development of mulitlingual corpus linguistics, which is one of the more exciting current trends.

The availability of parallel corpora in different languages has sparked off a number of impressive attempts to associate words and phrases in the two languages together thus laying a basis for automated translation (e.g. Church and Gale, 1991). The study of concordances of putative translation equivalents in seven languages has led to proposals for multilingual lexicography that is of the "minimal assumption" style (Council of Europe, 1991). There is every reason to foresee a surge in activity in this area as the raw materials become more readily available to researchers.

The general kind of software advocated in this chapter is or will be sophisticated, expensive and complicated. It is unlikely that all of it will be available at each research site, any more than all the corpora are likely to be gathered together. Appropriate software will have to be accessed remotely, by networking. Therefore a high standard of documentation will be required, and a considerable discipline of standardisation. In addition the software will have to meet standards of robustness, portability and flexibility that are not commonly found.

The toolkit advocated here is not intended to be more than illustrative of what will be possible with corpora, and what will be made available to a user group which will rise wiftly in numbers and diversify in needs, aims and skills. The qualities of the tools - speed, full automation, hierarchical structure, lack of size restrictions - are as important as the particular tools themselves.

References

Biber D. (1988): *Variation Across Speech and Writing.* Cambridge, Cambridge University Press.

Blackwell S. (forthcoming): "From Dirty Data to Clean Language" in *Proceedings of the 13th ICAME Conference,* P. De Haan, N. Oostdijk and J. Aarts (eds.), Amsterdam, Rodopi.

Chruch K., Hanks P. (1989): "Word Association Norms, Mutual Information and Lexicography". *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver,

British Columbia. 76-83.

Chruch K., Gale W. (1991): "Concordances for Parallel Text". *Using Corpora: Proceedings of the Seventh Annual Conference of the UW Centre for the NEW OED & Text Research*. St.Catherine's College, Oxford.

Collier A. (forthcoming): "Issues of Large-scale Collocational Analysis". *Proceedings of the 13th ICAME Conference,* P. De Haan, N. Oostdijk and J. Aarts (eds.), Amsterdam, Rodopi.

Council of Europe Multilingual Lexicography Project (1991): Final Report for the Period up to 31st December 1991, John Sinclair (Co-ordinator).

Francis G. (forthcoming): "Corpus-driven Grammar and its Relevance to the Learning of English in a Cross-cultural Situation", *English in Education: Multi-cultural Perspectives,* A. Pakir (ed.), Singapore, Uni Press.

Garside R., Leech G., Sampson G. (eds.) (1987): *The Computational Analysis of English.* London: Longman.

Greenbaum S. (1990): "The International Corpus of English", in *ICAME Journal* 14, 106-8.

Halliday M.A.K. (1966): "Lexis as a Linguistic Level", in *Memory of J.R. Firth,* C.E. Bazell, J. C. Catford, M.A.K. Halliday and R.H. Robins (eds.), 148-62. London, Longman.

Halliday M.A.K. (1985): *An Introduction to Functional Grammar.* London: Arnold.

Jelinek F. (1985a): "Self-Organized Language Modeling for Speech Recognition", *IBM Research Report*. T. J. Watson Research Center, Yorktown Heights, NY.

Jelinek F. (1985b): "The Development of an Experimental Discrete Dictation Recognizer", in *Proceedings of the IEEE* 73, 1616-24.

Karlsson F. (1990): "Constraint Grammar as a Framework for Parsing Running Text". *Papers Presented to the 13th International Conference on Computational Linguistics, Helsinki 1990.,* H. Karlgren (ed.), 168-73.

Marcus M.P., Santorini B. (1992): "Building Very Large Natural Language Corpora: the Penn Treebank." Paper Presented at the Pisa Workshop on Textual Corpora, January 1992.

Oostdijk N. (1988a): "A Corpus for Studying Linguistic Variation", in *ICAME Journal* 12, 3-14.

Oostdijk N. (1988b): "A Corpus Linguistic Approach to Linguistic Variation", *Literary and Linguistic Computing* 3, 12-25.

Reed A. (1977): "CLOC: A Collocation Package", in *ALLC Bulletin* 5.

Renouf A. (forthcoming): "A Word in Time: First Findings from the Investigation of Dynamic Text", in *Proceedings of the 13th ICAME Conference,* P. De Haan, N. Oostdijk and J. Aarts, (eds.), Amsterdam, Rodopi.

Renouf A., Sinclair J. (1991): "Collocational Frameworks in English", *English Corpus Linguistics*, K. Aijmer & B. Altenberg (eds.), 128-44. London, Longman.

Sinclair J. (ed.), (1990a): *The Collins Cobuild English Grammar*, London and Glasgow, Collins.

Sinclair J. (ed.) (1990b): *The Collins Cobuild Student's Dictionary*, London & Glasgow, Collins.

Sinclair J. (1991): *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

Sinclair J. (1992): "The Automatic Analysis of Corpora", *Directions in Corpus Linguistics*, J. Svartvik (ed.), 379-97, Berlin, Mouton de Gruyter.

Yang H. (1986): "A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts", *Literary and Linguistic Computing* 1, 93-103.

Yarowsky D. (1992): "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", in *Proceedings of COLING '92*, Nantes, France.

Zipf G.K. (1935): *The Psychobiology of Languag*, Boston, Houghton Mifflin.

Relevant NERC Papers

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", *7th Annual Conference of the UW Centre for the NEW OED and Text Research, Using Corpora*, pp. 170-196, Oxford, Oxford University Press, NERC-103.

Coniam D. (1991): "Boundary marker: system description", University of Hong Kong, Working Paper, NERC-21.

Federici S., Pirrelli V. (1991): "Tagger SECS: a neural environment for corpus-driven tagging", ILC Pisa, Working Paper, NERC-20.

Francis G. (1993): "Look-up Tagger", Working Paper, COBUILD Birmingham, NERC-176.

Nakamura J. (1992): "Hayashi's Quantification method type III. A Tool Determing text typology in large Corpora", University of Tokushima and COBUILD Birmingham, Working Paper, NERC-137.

Osborne G. (1993): "CHOC: a Text Analysis Tool", Working Paper, Apricot Ltd. Birmingham, NERC-175.

Saba A., Ratti D., Catarsi M.N., Cappelli G. (1991): "MORFSIN", ILC Pisa, Working Paper, NERC-22.

Sinclair J. (1991): "The automatic analysis of corpora", University of Birmingham, Working Paper, NERC-19.

Chapter 7

Knowledge Extraction

1 Introduction

This chapter is mainly concerned with the process of acquiring linguistic information directly from textual data, and the use of this linguistic information in a number of non-trivial NLP tasks/applications. It begins with some general remarks on the relation between rule-based approaches to NLP and linguistic knowledge acquisition from texts, stressing the importance of the diversity of

goals for NLP research, and the contributions both types of approach can make. The point is made that there is a vast and yet largely unexplored common area where a number of hardly tractable *cruces* in NLP can be promisingly tackled through a combination of the two approaches. The first half of this chapter discusses in some detail the methodological novelty of many current strategies for knowledge extraction. The second half shows how some well-known difficulties of rule-based approaches can be partly addressed by applying knowledge-extraction software to large amounts of text in Machine Readable Form.

1.1 *The relation between corpus-based and rule-based work in NLP*

A rule-based system is one which relies entirely on the formulation of explicit rules for relating (linguistic) objects. Corpus-based approaches to NLP rely on systems which either do not use explicit linguistic knowledge at all and make direct use of information automatically (or semi-automatically) extracted from large corpus resources, or rather integrate linguistic knowledge already explicitly modelled by a set of rules by running these rules on a large "training" corpus, and assessing their performances. Rules are thus checked, amended or assigned "preference scores" (see infra). In spite of some apparent differences in the way linguistic information is represented, both approaches aim at **describing** domains of linguistic knowledge. But how does one know whether most of what is relevant to a certain domain has been taken into account for description? We refer to this issue as the problem of **linguistic knowledge acquisition**.

1.2 *Linguistic knowledge acquisition: a major bottleneck in NLP*

Clearly, corpus-driven testing of grammars would in principle be dispensable if linguistic knowledge were already fully and explicitly available for writing rule-based grammars. The most serious bottleneck for rule-based systems is the problem of completeness and coherence of the linguistic knowledge to be covered in a certain domain. The concentration on problem cases and carefully chosen tricky examples which mostly characterizes rule-based approaches has been extremely valuable in highlighting limitations and inadequacies of theoretical proposals or computational architectures. However, the scaling-up problem of using a rule-based grammar on some unconstrained large-scale corpus resources for concrete applications has made more and more acute the problem of grammar coverage, and the related problem of acquiring linguistic information in a complete and coherent way. For example, for MT systems the acquisition and maintenance of bilingual lexical knowledge, and the considerable difficulty of capturing explicitly the precise conditions under which elements stand in a translation relation are key problems. In rule-based parsing systems, complete disambiguation of elements in context is seldom attempted, since it leads to a serious problem with multiple output. Certainly, the use of mainstream formalisms is important and helpful in providing elegant solutions. But this is not sufficient. Ultimately one needs a way of semi-automatically or automatically deriving linguistic information from texts, if only to determine what the language in the intended input texts is actually like.

1.3 *On the impact of corpus-based work in Linguistics and NLP applications*

Under the big umbrella of Linguistic Knowledge Extraction (or Knowledge Acquisition, henceforth LKE) lies a whole variety of different approaches (e.g., rule-based vs stochastically driven), purposes (e.g., tagging, acquisition of lexical resources, building up of grammatical tools or NLP systems), types of textual/linguistic sources being exploited (e.g., raw corpora, annotated corpora, general *vs* specific corpora, dictionaries, grammars and the like), methodologies of extraction (manual, semi-automatic, fully automatic etc.), and uses/applications (Speech Recognition, Optical Character Recognition, Machine Translation etc.). Nevertheless, all such approaches revolve around a common basic assumption: the complex task of language description draws more heavily on the process of inductive elicitation of regularities from real instantiations of linguistic usage (i.e. real texts), than on the privileged access to experts/informants' internalized linguistic competence through intuition/introspection, prior to the evidence provided by actual data. This assumption has a number of implications at the level of linguistic methodology:

i)the **content** of linguistic knowledge takes priority over the **formal representation** of items of linguistic knowledge. If the study of grammar formalisms has made a fundamental contribution to the development of a rigorous approach to language description during the eighties, a LKE oriented study of language attempts to make up for the lack of attention given to descriptions of the linguistic content of actually occurring individual words/constructions.

ii)the role and relevance of restrictions on language use due to contextual factors such as specific subject domains, communicative purposes, etc., are given more attention: a *functional* or *linguistic-external* mode of explanation complements an exclusively *linguistic-internal* approach to language description.

iii)accordingly, much more emphasis has been laid recently on the exception-ridden nature of language.

iv)the task of generalizing about linguistic data, as conceived so far in linguistic theorizing (basically as a form of descriptive parsimony) is far less emphasized.

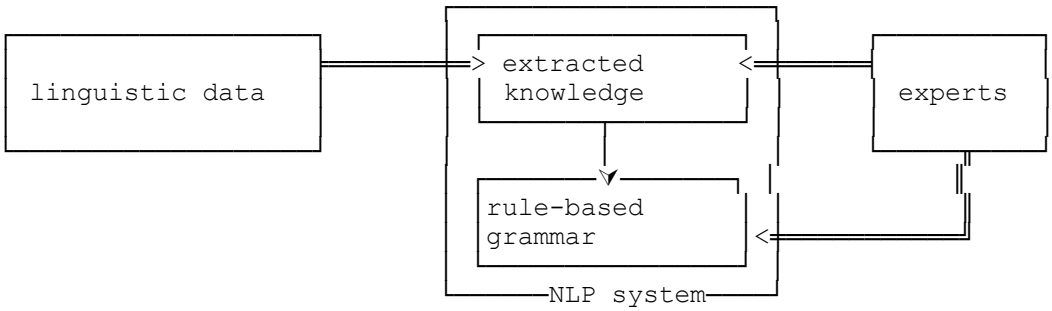
v)all linguistic phenomena, and not only those about which it is easier to make general statements, are given attention.

Generally speaking, all such concerns involve laying stress on the importance of connecting linguistic knowledge with **real texts**. This task has proven to be so challenging "... as to be beyond the range of human conscious awareness and descriptive capability" (Leech, 1992). "We have apparently reached the limit of the ability of human analysts to hand-craft such linguistic knowledge resources as grammars and dictionaries needed for particular NLP applications" (Leech, *ibidem*). In 1979, Lightfoot pointed out: "*The crucial factor in science is depth of explanation, not data-coverage*". The methodological uneasiness which has led to an increase of interest in the field of LKE sheds a different light on this claim, which can be rephrased as follows: "*In linguistics, Data Coverage poses a real problem of explanatory adequacy to linguistic theories*". This stance has made more urgent the need for tapping electronic resources on an unprecedented scale, both in terms of the amount of textual material to be collected, and of the computational resources (both hardware and software) required for the process of accessing and exploiting such linguistic material.

1.4 *Three knowledge acquisition models*

The relation between extracted knowledge and NLP systems' architecture can be mediated by a number of factors. Following (Tsujii et al., 1992), three basic models of such interaction can be seen as emerging over the last few years in the literature. They can be illustrated by the following diagrams.

a) *the add-on model*



a) represents a class of "hybrid" NLP systems, which results from the simultaneous combination of the "formalistic" approach to language parsing of the eighties, and the new data-oriented/corpus-driven perspective we want to investigate here. Within this model, the role of experts' intuition is not utterly rejected, but just curtailed. Usually, linguistic data is automatically acquired. Statistical techniques play an important role in bringing out what is looked for, or, at least, in assisting the linguist in the phase of selecting and assessing what is relevant for his/her own purposes, on a more objective basis.

b) *the parasitic model*

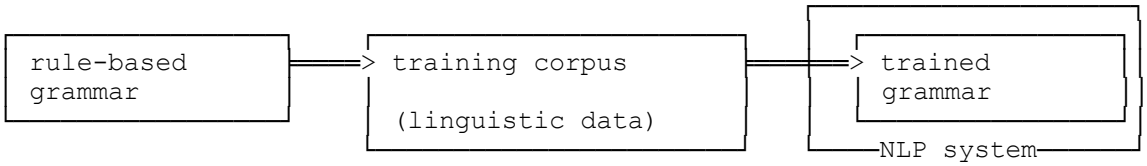
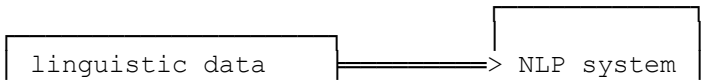
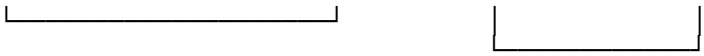


Diagram b) above models LKE as "training". A fixed set of knowledge is assumed: rules have been manually crafted. Training consists in running a parser on a sample text in order to elicit the relative frequency of application of individual rules on that corpus type. Frequency results are then used **i)** to devise preferential cues, which yield the more plausible result in case of multiple parses, **ii)** to prune out the number of branching choices of a non-deterministic grammar at processing time, by ranking them in order of probability of use (the algorithm attempts first the most high-ranked instruction). A variant of this approach is the so called "supervised" training, or "tuning". In supervised mode, training consists of assigning probabilistic scores to rules on the basis of frequency measures elicited from a "treebank", that is a substantial portion of running text, where each sentence has already received an appropriate structural representation (a phrasal tree or an equivalent one-dimension surrogate of it) either (semi-)automatically or manually.

c) *the direct model*





c) reflects the view that, whatever form of abstract linguistic knowledge is added to real texts (either directly, e.g. as a form of labelling interspersed with text, or indirectly, e.g. as a hierarchical structure superimposed on instances of text as in "thesaurus trees"), the process of combining such knowledge into a coherent NLP system is too complex to be handled by human analysts alone. The main claim is that such a process should rather rely on fully automatic, self-modelling routines of some kind: e.g., in the form of an intricate distribution of weights over nodes in a "parallel distributed network", estimated Markovian probabilities over state transitions in a transition network, etc.

1.5 *Structure of the chapter*

In the first half of this chapter, we will consider different strategies which are grouped under the headings "add-on", "parasitic" or "direct" models, as illustrated above. For each method the following points will be focused on:
INPUT DATA: the amount and nature of data required by each technique;
TECHNIQUES: specific techniques (Mutual Information, Clustering, Neural Networks etc.) and individual tools (concordance tables, inverse indexes etc.);
OUTPUT DATA: the nature, transparency and controllability of the extracted information; the structure of the results obtained from corpus study;
REUSABILITY: the integration of the acquired information into an effective NLP system as relevant to the application-oriented side of our investigation: this aspect will be further expanded in the second half of this chapter, where the effectiveness of some methodologies of knowledge extraction is tested in terms of their performances in practical applications.

It should be appreciated that no review such as this one can claim to be anywhere near exhaustive in its coverage: the intention is to give the reader a flavour of some of the methodologies/applications attempted so far, with a view to highlighting the main implications rather than offering a complete repertoire of the knowledge extraction systems currently available. For more details, the reader is referred to (Pirrelli, 1993, NERC-79).

2 Methodologies of Linguistic Knowledge Extraction

2.1 *"Direct models"*

2.1.1 *Acquisition of Morphological Information*

Most "direct" systems for extraction of morphological information work on carefully selected lists of character-based strings (usually word-forms), rather than texts. In PDP models (Rumelhart and McClelland, 1986) pairs of stem/inflected word-form are fed into an untrained pattern associator, modelled as a connectionist network. Analogy-based systems (Skousen, 1989) take as input pairs of word-form/morphological class, to provide a statistically based procedure which models the process of classifying a given input word- form on the basis of its surface analogy with already acquired (and classified) *exempla*. In Hidden Markov models (Gilloux, 1991), morphological analysis is carried out by a program designed to construct **finite state transducers** automatically, thus dispensing with the

onerous and lengthy task of hand-crafting the set of needed translation rules. Because of these input requirements, direct-models of LKE for morphology will not be considered in detail here. More on this important and fast-growing research area can be found in (Pirrelli, 1993, NERC-79), to which the reader is referred for an in-depth overview. A **corpus-driven** morphological analyser is described by (Federici and Pirrelli, 1992, 1993). This strategy is very practical, since one can use as input any ordinary, already morphologically disambiguated text, and it resembles the actual learning process of a child gradually going through a text, improving its knowledge as it goes on reading.

2.1.2 *Acquisition of Morphosyntactic Information*

(Nakamura et al., 1990) illustrate the architecture of a **neural network** for word category prediction in English texts. The system, called NETgram, contains a core network for bi-gram disambiguation. Because this network is trained to guess the next word-tag as output for a given input word-tag, hidden layers are expected to learn some "linguistic" structure from the relationship between one word-category and the next in the training text. The system can be expanded to tackle more complex input: trigrams, 4-grams etc. Performances are evaluated by comparing the NETgram and the statistical trigram model. Accuracy rates of NETgram, however difficult to evaluate, are claimed to be higher than those of stochastic trigram models[18]. Furthermore, the statistical trigram model cannot learn tag sequences which do not appear as a trigram in the training data. NETgram will always tentatively provide an output: in many cases its own interpolation fares well. This aspect connects with another one: even if training data are insufficient to estimate accurate trigram probabilities, NETgram performs effectively. NETgram interpolates sparse trigram training data using bigram training memory.

2.3 *Acquisition of Syntactic Information*

In this context, "direct systems" are capable of eliciting phrasal structures from a "raw" (i.e., non pre-parsed) text corpus.

A Mutual Information Parser

(Brill et al., 1991) demonstrate how **unlabelled** constituent boundaries (called "distituents") can be extracted from a sequence of n categories, or an n-gram, by analysing the mutual information values of the part-of-speech sequences within that n-gram. They make use of a "generalized mutual information" statistic, an extension of the bi-gram (pairwise) mutual information of two events into n-space. Tested on unconstrained text from a reserved test corpus, the parser brackets unlabelled constituents, averaging about two errors per sentence for sentences under 15 words in length. On sentences between 16 and 30 tokens in length, it averages between 5 and 6 errors per sentence.

---

18 Testing is performed indirectly: the NETgram system is "plugged into" a speech recognition system; accuracy rates thus refer to the task of successful word recognition, when tagging is used as an add-on facility:

training sentences	NET gram	Statistical model
512	86.3	85.5
1024	86.9	85.4

Unfortunately, we cannot compare such results to analogous "goodness of fit" measures provided by other authors.

2.2 *"Add-on Models"*

"Add-on models" imply human interaction. Linguistic knowledge acquisition is not carried out through unsupervised, fully automatic procedures, but relies on the interaction with an "expert" or "informant".

2.2.1 *Acquisition of Morphological Information*

Expert systems for semi-automatic extraction of morphological information take a rather prominent position in the literature (see, for example, XMAS by (Zhang and Kim, 1990)), or the system described by (Brasington et al., 1987), on which more in (Pirrelli, 1993, NERC-79). However, like most "direct models" for Morphology, they work on stem/inflected-form pairs, which are also "commented" in many cases. Hence, they are not further delved into in this chapter.

2.2.2 *Acquisition of Morphosyntactic Information*

(Marshall, 1983) describes the LOB corpus tagging algorithm developed in Lancaster (later named CLAWS) as similar to that employed in the TAGGIT (Greene and Rubin, 1971) program, where the task of contextual disambiguation of word tags assigned to word forms taken out of context is carried out on the basis of a rather substantial set of handcrafted "context frame rules" (about 3,300 of them). Each rule in TAGGIT, when its context specification is satisfied, has the effect of deleting one or more candidates from the list of possible tags for one word. However, despite some similarities (in the tag set used, in the dictionary used, in the disambiguation routine, etc.), three fundamental novelties make CLAWS more than a simple offshoot of the TAGGIT system:

1) word spans of unlimited length can be handled (subject to machine resources);
2) a precise mathematical definition is possible for the main algorithm of CLAWS;
3) the CLAWS algorithm is quantitative and analog, rather than artificially discrete.

Bi-gram relative probabilities are derived from a square matrix of **collocational probabilities**, which indicate the relative likelihood of co-occurrence of all ordered pairs of tags. This matrix is mechanically derived from a representative **pre-tagged** portion of the Brown Corpus (of about 200,000 words). CLAWS has been applied to the entire LOB corpus with an accuracy of "between 96% and 97%". Without the use of the hand-crafted idiom-list, the algorithm was 94% accurate.

(Federici and Pirrelli, 1991a, NERC-20) illustrate a hybrid analogy-based parallel system for disambiguating ambiguously tagged texts, which exploits **both** self-modelled associative links acquired through a learning phase, **and** rule-driven syntactic constraints on admissible tag sequences implemented as **pre-wired inhibitory links** among units in the network. The model is first trained on a comparatively small sample of a given pre-tagged text. Then, it attempts to disambiguate an ambiguously tagged text of the same type as the training excerpt on the basis of the acquired

knowledge. When pre-wired rule-like constraints fail to meet the input context, the inferential routine is invoked, and an analogical response is attempted by the system on the basis of a "best fit" criterion.

2.2.3 *Acquisition of Syntactic Information*

Current tagging systems do not take account of differing frequencies of occurrence of lexical entries; for example, in the LOB corpus the verb *believe* occurs with a finite sentential complement in 90% of citations (Briscoe and Carroll, 1991), although it is grammatical with at least a further five patterns of complementation. This type of lexical information, which is very likely to vary between sublanguages, should be integrated into a probabilistic model. However, the acquisition of the statistical information from which these probabilities can be derived is problematic. (M.R. Brent, 1991) suggests looking for **clear, unambiguous cases** of a certain **type** of subcategorisation pattern. For example, a pattern like "V PRONOUN to V" cannot possibly backfire in the way "V NP to V" does, since, while "I expected him to eat ice-cream" is a well-formed sentence in English, "I doubted him to eat ice-cream" is not. Testing is carried out on a 2.6 million word Wall Street Journal corpus provided by the Penn Treebank Project. Of approximately 5,000 verb tokens found, there were 28 disagreements only between the verb detecting routine and the Penn tags.

2.3 ***Parasitic models***

2.3.1 *Acquisition of Morphosyntactic Information*

VOLSUNGA (De Rose, 1988) has been developed to address some of the problems left open in CLAWS. VOLSUNGA does away with the preprocessing phase of idiom look-up. Because of this, manually constructed special case lists are not necessary. Secondly, the optimal path is the one whose component collocations multiply out to the highest probability. No other add-on probability is used. Thirdly, the VOLSUNGA bi-gram transition matrix is calibrated by reference to 100% instead of the 20% of the Brown Corpus, and has been applied to the entire Corpus for testing with remarkable accuracy rates (about 96%). This is a particularly important test because the Brown Corpus provides a long-established standard against which accuracy can be measured. Last but not least, where CLAWS scales transition probabilities by 1/2 for rare tag-word pairs, and by 1/8 for extremely rare tag-word pairs, VALSUNGA uses relative frequencies of tag-word pairs themselves (called *Relative Tag Probabilities*) as a factor in the equation which defines transition probabilities.

2.3.2 *Acquisition of Syntactic Information: Probabilistic grammars*

(Fujisaki et al., 1989) describe an experiment in corpus parsing by training a context free grammar onto a sample corpus. The training corpus consists of 4206 sentences. The grammar is initially converted into Chomsky Normal form (only binary branching rules are allowed), and then into Griebach Normal form (which requires that each production contains a terminal category as its leftmost member). The double conversion process simplifies the statement of the training and parsing algorithms. The training process involves the assignment of a probability score to each CF rule on the basis of the frequency of

its application in the training corpus. The training procedure is unsupervised, which means that data concerning frequency of use include both correct and incorrect parses of the corpus. The re-estimation process is carried out using some version of the inside-outside algorithm (Baker, 1982) which is guaranteed to converge, in the sense that the probabilities assigned to rules tend to stabilise. A success rate of 85% on a testing corpus of 84 sentences is reported. However, results are difficult to evaluate in the absence of full details on the nature of the training corpus.

An ID (Immediate Dominance) LP (Linear Precedence) Grammar factors out the two types of information encoded in CF rules - immediate dominance and linear precedence - into two rule types which together define a subset of CF languages. (Sharman et al., 1990) conducted a training experiment on a grammar in ID/LP format. A restricted set of 16 non-terminal symbols were derived from the 64 actually used in the "treebank" (a preprocessed collection of unrestricted sentences of English taken from the Associated Press newswire material, consisting of about one million words of text). A restricted set of 100 terminal symbols were derived from the 264 actually used in the treebank. The treebank was divided into two parts, one for adapting the grammar, and one for testing. The word list of the training data was extracted, with the unigram probability of each entry. This is used to generate the probability of tag assignments to words when those words are found in a sentence to be parsed. Initial estimates of the dominance relations for non-terminal symbols and of the precedence relations for non-terminal and terminal symbols were derived from the treebank. The resulting grammar was used to parse 42 sentences of 30 words or less. 18 of the parse results were identical to the original manual analysis, while a further 19 were 'similar', yielding a success rate of 88%.

An LR parser is a shift-reduce parser guided by a parse table indicating what action should be taken next after shifting a given term in the string to be parsed. The table consists of two parts: an action table and a go-to table. Both parts consist of a square matrix, having states as rows, and non-terminal symbols as columns. At each row-column cross, the parser is told what kind of action to perform, when the pre-terminal symbol in the column is shifted, and the parser is in the state represented by the row. Unlike other probabilistic LR based techniques, where frequency estimates concern CF rules in the grammar, Briscoe and Carroll extract probabilities by running the LR parse on a "raw" corpus in interactive mode. A number of optimization techniques are used to limit, as far as possible, the need for interaction between the parser and the human user, which corresponds to the number of alternative parses of the same input sentence. This way, probabilities are associated in order to parse states and actions, so that a certain amount of context-sensitivity is learnt; this totally eludes probabilistic CF grammars. The parse table is a non-deterministic finite state automaton, so it is possible to apply Markov modelling techniques to the parse table in a way analogous to their application to lexical tagging. If the results of a test carried out on a relatively small sample (63 parsed definitions) are interpreted in terms of a goodness of fit measure such as that of (Sampson et al., 1989), the measure would be better than 98%. If we take correct parse/sentence as our measure, then the result is 81%. This rate falls to a disappointing 57.4%, however, when the parser is applied to a further 54 noun definitions not drawn from the training corpus.

One way of extending the constituent likelihood approach from word-tagging to parsing (Marshall, 1985) involves assuming that the parser is trying to find the likeliest assignment of labelled brackets for an input sentence, with the opening bracket symbol and closing bracket symbol for each "hypertag" treated as two separate symbols, called "hyperbrackets". The parser is run over two samples of the tagged LOB Corpus (totalling over 250 sentences), achieving a success rate of approximately

50% .

2.4 *Summary and conclusions*

INPUT DATA: very few self-modelling systems can work directly on "raw" data; the majority of them use **pre-processed data** only, which have already been enriched with some form of linguistic information, either through tags, or tree-structures, or feature structures etc. It is generally acknowledged that pre-processed textual material enhances the systems' performance. This is also true of those systems which can work on raw data. As to the **amount of data** required for training, the range of variation is wide, depending on a) the amount of linguistic knowledge already embedded in the system, b) the type of linguistic information to be extracted, c) the speed of the self-modelling routine, and d) the "rawness" of the training text. Statistical training is the most data consuming (over 1 Mb of forms, there being virtually no upper limit to the optimal size of the training corpus). Interactive add-on models are obviously less costly. PDP approaches take a position half-way between the two. Fast-learning software has been designed which requires tens of thousands of training forms only, but still needs some thorough testing. It is clear that all techniques will require either growing quantities of running texts as training material, or less large but fairly **domain-specific** sample corpora, which will hopefully guarantee quicker convergence of the learning routine.

OUTPUT DATA: the nature, transparency and controllability of the information extracted depends on the technique adopted and the input data. In PDP techniques, only the input and output layers are available to inspection: the hidden layer, which arguably models the acquired linguistic knowledge, is totally transparent to the user. At the opposite end, rule-extraction and grammar-tuning software provides a full set of explicitly encoded rules and a formal characterization of the language represented in the training corpus. Both pieces of linguistic information are of great value for enhancing the performance of NLP systems.

IMPROVEMENTS ON CURRENT TECHNIQUES: For most techniques, training requires manually pre-tagged input. This can be a very laborious and error-prone task, especially with training corpora of considerable size (over one million word-forms). One way to get around massive manual pre-tagging is to rely on **bootstrapping routines**: at first a small amount of text is manually tagged and used to tag more text; then the tags are manually corrected and used to retrain the model (Derouault and Merialdo, 1986). This procedure raises the level of consistency and saves time. However, more should be done to speed up bootstrapping routines of this sort. In statistical methods, dealing with n-gram models of an order higher than two means having to face two big stumbling blocks: **parameter estimation** and **sparse data**. The two issues are obviously correlated: the more parameters to be counted in, the more likely it is that some of them will not be reliably represented in our training corpus. But while the problem of sparse data can be, in theory, avoided by using bigger and bigger corpora, the computational complexity of the estimation process for parameters coming from even a small set, soon becomes intractable for $n > 3$. Interpolated estimation seems to be a workable solution: a mixture of 1-gram and bi-gram models has proved to fare reasonably well in CLAWS and VOLSUNGA. More accurate estimates of less reliable n-gram models turn out to give better results than less accurate

estimates of more reliable n-gram models. Bigger corpora of specific sublanguages are likely to further impove estimation accuracy. As a result, it turns out that, in practice, a response to the problem of computational inefficiency of parameter estimation has made more acute the problem of availability of larger corpora. The importance of dealing with texts from **restricted subject domains** is also stressed by connectionistic approaches working on (simulated) neural networks. These models are germane for two main reasons: because of their parallel architectures, they are based on algorithms whose order of complexity is generally lower than exponential; secondly, they appear to be able to interpolate unknown data better than statistical approaches do. "Calibrated" training, that is training carried out on language specific corpora, provides for quicker learning performances. The crucial requirement here seems to be a fairly small corpus which best fits the target population.

EXPORTABILITY: using the acquired knowledge to parse data outside the scope of the training material (the "exportability" of acquired knowledge) is still rather problematic, at least with respect to lexical entries. The problem of "interpolating" new evidence by means of "old" knowledge either through "simulated annealing" (Atwell, 1985, Sampson et al., 1989) or through other statistically defined measures of "closest fit" has not been properly and successfully addressed to date. In order to deal with "grammar interpolation", pieces of software which are cleverer than those reviewed so far appear to be needed, such as "inductive grammars" (Berwick, 1985) or analogy-based grammars. Promising insights on this front come from parallel processing techniques (Nakamura et al., 1990, Federici and Pirrelli, 1991b).

SCALING-UP: in this context "scaling-up" means applying the extracted grammar to different domains of linguistic analysis (e.g. from a syntactic parse to a semantic interpretation). Attempts have been made in this direction. (Grishman, Hirshman and Nhan, 1986), and (Grishman and Sterling, 1992), illustrate an interesting set of discovery procedures for the acquisition of "selectional patterns" automatically from a domain specific, syntactically pre-parsed sample corpus. These experiments show that scaling-up from syntactic parses to "proto-semantic" disambiguation in **specific domains** is possible. Moreover, it is a promising solution to the problem of selecting the correct analysis from the set of multiple parses licensed by a grammar. Clearly, the problem of "sparse triples" can again be addressed by using **larger training data**, and cleverer ways of combining triples to give better estimates.

3 Applications

3.1 *Introduction*

In this section the following task domains will be covered:

i)Dictionary Construction: techniques for developing large on-line dictionaries, either by exploiting already existing ordinary dictionaries in Machine Readable Form, or by scouring large textual databases (either by statistical means, or syntactic parsing, or a combination of the two), or

"parallelling" two comparable sources of the same linguistic information (i.e., two "comparable" dictionaries, two translationally equivalent texts etc.).

ii)Speech Applications: including both Text-to-Speech conversion and Speech recognition; such applications do not cover the use of phonological models in a strict sense (usually hidden-Markovian models); these were omitted from the foregoing review of the state-of-the-art.

iii)Word Processing: including spelling-checkers and text-inputting.

iv)Document Retrieval: including text-indexing, text-retrieval and database query.

v)Machine Translation: including various sorts of so-called "example-based", "case-based" or "similarity-based" approaches.

3.2 *On-line dictionary construction*

The production of a large on-line dictionary, often for some particular topic-domain, can benefit greatly both from pre-existing dictionaries and from suitable textual corpora.

3.2.1 *Using morphology*

An example of the former kind is (Wolff, 1984), where a broad-coverage on-line medical dictionary is produced as a result of expanding a large, but far from complete, pre-existing dictionary, through morphological decomposition of the orginal set of lexical entries. The morphemes thus obtained, together with their semantic properties and restrictions on their concatenation, are then recomposed to construct possible entries for an expanded version of the same dictionary. This example pictures a fairly straightforward scenario: information about morphological analysis is extracted from manually tagged texts (or lists of words) according to the techniques illustrated in the first part of this study; this knowledge is more or less directly exploited to devise word-structure rules (or regularities); at a second stage, the morphological engine thus arrived at through training is applied to texts of a rather specific type (e.g. dictionaries) in order to produce enhanced dictionaries.

However, as we shall see in a moment, there are a number of research domains in which a similar scenario is a long-term goal rather than a reality.

3.2.2 *Using syntax and semantics: the egg-and-chicken bottleneck*

For practical research in NLP, it is indispensable to develop large-scale semantic dictionaries for computers. It is particularly vital to improve techniques for compiling semantic dictionaries from natural language texts. However, there are at least two difficulties in analyzing existing texts: the problem of syntactic ambiguities and that of polysemy. Hence, researchers are faced with a typical egg-and-chicken situation: they need to extract semantic and syntactic information by parsing pre-existing texts, but, in order to do that accurately, they need to be already equipped with the sort of information they are looking for.

Various ways out of this vicious circle have been put forward in the literature. All of them hinge on some sort of bootstrapping strategy, or gradual approximation, either in a supervised mode or in an unsupervised one. Basically, some more or less "incomplete" parse is carried out on a text as a first

approximation. Results can then be checked manually for revision and added to the parsing engine for improvement (supervised mode): hopefully this would produce an enhanced version of the parser, which is now able to detect subtler regularities in texts and to widen its coverage. Alternatively, results are automatically updated in the process of parsing itself, as new evidence accrues by parsing more and more text (unsupervised mode). Both supervised and unsupervised bootstrapping strategies require that considerable amounts of texts be scoured. This follows from the fact that, in general, "incomplete" parses aim at locating neat unambiguous examples, since ambiguous ones are exactly those which the parser cannot tackle yet. Thus, only large texts can guarantee that a significant number of neat examples will be stumbled upon. Clearly, such a process of gradual approximation can also benefit from restricting the type of input text to be processed. Restricted input texts are mainly drawn either from fairly specific text types (e.g. dictionaries) which encode textual information in a hopefully more systematic way, or from specific knowledge domains (so-called "sublanguages"), whose lexicon and syntactic constructions are allegedly less unpredictable than in unrestricted domains. In the next subsection we will consider only some of these applications. For simplicity, the overview is split into two parts: extraction from on-line dictionaries, and extraction from other text types.

3.2.3 *Using a dictionary as input-source*

The extraction of lexical semantic information from MR dictionaries has been attempted in several ways, using different strategies. Most of them converge onto the objective of locating the genus terms in (mainly) noun and verb definitions, the underlying assumption being that the genus term represents inherent features of the word it defines. A great deal of effort has been put into the individuation of recurrent patterns emerging from the way lexical defintions are structured (Calzolari, 1984), and into heuristics for their automatic elicitation (Chodorow and Byrd, 1985). Lexical semantic relations are lexically expressed in dictionary definitions by means of so-called *defining formulae*, that is, wordings or phrases such as "characterized by", or "a state of", or "a group of", etc. (Evens, 1988). (Alshawi, 1987) uses a hierarchy of patterns which consist mainly of part-of-speech indicators and wildcard characters; (Markovitz et al., 1986), (Jensen and Binot, 1987) and (Nakamura and Nagao, 1988) also use pattern recognition to extract semantic relations such as taxonomy from various dictionaries. On the other hand, (Montemagni and Vanderwende, 1992) suggest that a minimum of "loose" syntactic parsing (as opposed to pattern matching) is an indispensable prerequisite of the process of extracting reliable semantic information from dictionaries. For specific subject domains, it has been suggested that conceptually (as opposed to syntactically) driven parsing strategies are likely to achieve more accurate results (Reedijk, 1991, Martin, 1992). Conceptually driven parsing techniques rely on the existence of a pre-defined "conceptual system", meaning a set of unspecified relational predicates (e.g., *caused_by (disease, etiology)*) which are intended to exhaustively cover the conceptual space in a specialized dictionary. Such relational predicates get specified by the particular instantiations provided by dictionary definitions, by mapping "deep" conceptual structures onto their surface realizations within lexical definitions. In fact, it remains to be seen how mappings of this sort can be elicited. It has been claimed that the special word-classes and relations of a particular sublanguage provide the basis for a variety of natural language processing applications that would not be practicable in the language as a whole. In (Guthrie et al., 1991) sets of consistently contiguous words are extracted from machine readable dictionaries to help semantic disambiguation in information retrieval (see also below). In

(Farwell et al., 1992), linguistic information contained in on-line lexical entries from the Longman Dictionary of Contemporary English (LDOCE) is extracted and formatted in the form of a standardized LISP structure.

3.2.4 *Using more than one dictionary as input-source*

The attempt to integrate information coming from different dictionary sources through word-sense matches is likely to fail in a significant number of instances (Calzolari and Picchi, 1986, Atkins 1987, Boguraev and Pustejovsky, 1990). This is simply because dictionaries seldom describe the same word using the same sense distinctions. However, it has been claimed (Sanfilippo and Poznanski, 1992) that, when dealing with sources which use entry definitions which are not too dissimilar, a correlation technique based on word sense merging can be made to yield useful results, given a set of appropriate computational tools, whose intended purpose is to supervise a process of gradual integration. Some tools of this sort have been developed, such as a linguistically motivated database aimed to facilitate the interrogation of dictionary entries, and "comparators", that is, pattern associators which take as arguments pairs of "normalized" field values relative to the senses of the two dictionaries under comparison, and return a "correlation score" plus an advisory interface about the next steps to be taken.

3.2.5 *Using ordinary texts as input-source*

Ideally, knowledge about word relations should be acquired directly from considerable amounts of text with a minimum of manual pre-processing. This idea underlies many recent studies of word association. Results of these studies have important applications in lexicography, in the detection of lexico-syntactic regularities, such as, for example, support verb constructions (e.g. "draw conclusion"), phrasal verbs ("come up with"), subcategorization frames ("rely upon"), semantic relations ("part of"), different degrees of idiomaticity of expression ("to a *modifier* extent", "kick the bucket" etc.), and the identification of the general meaning of unfamiliar noun-noun compounds (Church and Hanks, 1990, Calzolari and Bindi, 1990, Hearst, 1992). Co-occurrence analysis augmented by syntactic parsing has also proved to be a useful tool for the purpose of word classification (Zernik, 1989, Hindle, 1990). All these studies are based on the assumption that syntactic similarity in word patterns implies semantic similarity.

Statistically collected associations provide pragmatic clues for lexical choice in sentence generation. For example, (Smadja and McKeown, 1990) can predict that "make a decision" is a better choice than, say, "have a decision" on the basis of purely statistical evidence elicited from large corpora. Basically, syntactic augmentation of statistical counts allows for lexically-based association techniques to be more carefully guided by syntactic clues. For example, (Hindle and Rooth, 1991) propose that a syntactic disambiguation criterion for PP attachment problems can be elicited by comparing the probability of occurrence of noun-preposition and verb-preposition pairs in V NP PP structures. Thus, for example, the preposition *to* occurs frequently in the context *send NP* __, i.e., after the object of the verb *send*, and this is evidence of a lexical association of the verb *send* with *to*. Similarly, *from* occurs more frequently in the context *withdrawal* __, and this is evidence of a lexical association of the noun *withdrawal* with the preposition *from*. Incidentally, for occurrences of V NP PP structures to be counted, no sentence-level parsing is needed, but only fairly "local" phrasal parsing.

However, there are reasons for doubting that surface distributional analyses are robust enough for disambiguation purposes. A preposition may or may not be attached to a verb, even if it frequently occurs with it, depending on the underlying semantic relations which are expressed in context. Apart from cases of strongly subcategorized complements such as those just mentioned, it is often the semantic category of the noun following a preposition that ultimately influences the choice of the proper attachment (cf. "vendo scarpe per uomini" English 'I sell man shoes', and "vendo scarpe per beneficenza" English 'I sell shoes for charity'). (Basili et al., 1992) make frequency counts of so-called "clustered associations", i.e. triples of the form C_1 *synt rel* C_2, where *synt rel* is a prepostion, and C_1 and C_2 are fairly gross semantic classes (e.g., PHYSICAL_ACT, HUMAN_ENTITY, PLACE, MACHINE etc.) which are extracted from a manually tagged (as to semantic word classes), and automatically pre-parsed (as to syntactic patterns) corpus. With fairly generic prepositions such as *of* and *with*, however, even clustered associations give no strong indications. Thornier cases of PP attachment such as "I saw the man with a telescope" as opposed to "I saw the man with a scarf" can be only tackled by relying on real world-knowledge clues. Unfortunately it is not clear where the necessary information about "men with scarves" is to be found. In order to tackle PP-attachment ambiguity (but the strategy is claimed to be usefully extendable to any other kind of structural ambiguity) (Sekine et al., 1992a) count "meaningful" co-occurrences between verbs and nouns, that is, co-occurrences where the nouns actually appear in the position of the head-noun of PP's which can be attached either to verbs or to other nouns. This poses the familiar "bootstrapping" paradox: in order to obtain frequencies of "meaningful" co-occurrences in sample texts, we have to know the correct attachment positions of PP's, and determining the correct attachment of PP's in a sample text requires knowledge of frequencies of "meaningful" occurrences. The following algorithm is set up to get around this chicken-and-egg situation: credits are assigned to candidate "instance-tuples" in such a way that the sum of the credits assigned to competing instance-tuples (those which show different attachment positions of the same PP) is equal to 1. In an ambiguous sentence such as "I saw a girl with a scarf", for example, the instance-tuples [girl, WITH, scarf] and [saw, WITH, scarf] are assigned 0.5 credit each. It is assumed that tuples corresponding to "intrinsic" ontological relations occur more often in texts than "accidental" ones. By iteration of the above mentioned algorithm it is then expected that there should be an increase in the credits assigned to instance-tuples which correspond to correct attachment positions, accompanied by a decrease in the credits assigned to competing ones. However we usually cannot expect to have a corpus of sample sentences which is large enough for "intrinsic" relations to appear significantly more often than "accidental" ones. One possible way to increase the number of co-occurrences is to introduce semantic similarity measures between words. Semantic (dis)similarities are counted on the basis of the (dis)similarities of their patterns of co-occurrence with other words (in short, two nouns are judged to be close to each other, if they often co-occur with the same words). Again we face another chicken-and-egg situation here. (Sekine et al., 1992b) illustrate an algorithm which uses as input the output of a Japanese tagging program which finds word boundaries and puts all possible parts-of-speech for each word under adjacency constraints. For the purposes of the illustrated application the system treats only noun sequences (generally referred to as "compounds", although they are not always compounds in a strict sense). A parser produces all possible syntactic descriptions among words in the form of syntactic dependency structures. The description is represented by a set of tuples, of the type [HEAD WORD, syntactic relation, ARGUMENT]. The only syntactic relation which is assumed to hold in this scenario is MOD, short for "modified". For example, the description

of a compound such as "file transfer operation" contains three triples:

a [*transfer*, MOD, *file*]
b [*operation*, MOD, *file*]
c [*operation*, MOD, *transfer*]

In order to resolve this sort of ambiguity a system may have to be able to infer extralinguistic knowledge. The best that a system can do, without full understanding abilities, is to select more plausible triples and reject less plausible ones. The self-learning algorithm computes the plausibility values of hypothesis-tuples like (*operation*, MOD, *transfer*) basically by counting frequencies of the instance-tuples [*operation*, MOD, *transfer*] as generated from the input data. At the first cycle credit scores are assigned to competing triples (b and c above) on the assumption that they are all equally probable: 1/2 each. Now, if an instance-tuple occurs frequently in the corpus or if it occurs where there are no alternative tuples, the plausibility value for the corresponding hypothesis must be large. At this stage "word-distances" are used to modify the plausibility values of the hypothesis-tuples. Word-distances are either defined externally using human intuition, or calculated in the previous cycle on the basis of already encountered tuples. Intuitively, the plausibility value of a given tuple (say [*operation*, MOD, *transfer*]) is increased by the plausibility value of any other tuple of the form [*operation*, MOD, x], where x is any word semantically similar to *transfer*. In fact, only the highest such value is allowed to increase the plausibility value of [*operation*, MOD, *transfer*].

It now needs to be specified how the distance between two words is automatically computed. Again, this is done by using already encountered tuples: intuitively two words are close to each other if they occur as arguments of the same tuples. Once all such distances are gauged, a clustering program will produce word clusters based on them, which will be further revised by human intervention. (Bindi et al., 1991, NERC-103) illustrate a statistical, lexically-based strategy for the individuation of clusters of semantically related words. Given a certain subset of "key-words" belonging to a certain semantic area, the algorithm applies Multidimensional Scaling techniques to vectors which represent each key-word in question in terms of the highest mutual information values with other words. The strategy exploits the intuition that similarities in word meanings can then be ascertained by the determination of coincidences in the contexts in which the words are used in different text passages. In the resulting representation, word distances are calculated on the basis of the typical contexts the key-words happen to occur in. Contexts are represented as clusters of content-words distributed around the centroid of the corresponding key-word.

(Utsuro et al., 1992) claim that difficulties arising from PP-attachment ambiguities in texts can be overcome, to a certain extent, by making use of translation examples in two distinct languages, the more distinct the better. The basic idea is that what looks like an ambiguous construct in one language might be rendered unambiguously in the other. For a pilot experiment, 50,000 translation examples have been collected from a machine readable Japanese-English dictionary and an English learner's textbook. In these bilingual corpora, more than 70 distinct Japanese verbs appear in more than 100 examples. The case slots of Japanese 'write' for example are extracted from 207 translation examples and described through features in a unification-based notation. In the process of extraction, bilingual feature label pairs are quite useful to find different case slots which are marked by the same postpositional particle in Japanese. On the assumption that 200 translation examples are required for

acquiring case frames for one verb, 100,000 translation examples are necessary for 70 verbs. If a bilingual corpus of 1,000,000 translation examples is obtained, it is possible, the authors claim, to compile a semantic dictionary for verbs with little interaction with a human instructor. It has been observed that there are a number of useful hyponym relations that are not likely to appear in a dictionary. This may be because the corresponding hyponyms are common knowledge and so do not need to be mentioned in a dictionary (like "broken bone"), or because the ISA relation underlying hyponymy is, in a sense, "non canonical" (Hearst, 1992). This applies even to extremely common locutions, although their automatic acquisition would be of great interest for a number of NLP applications (information retrieval, recognition of topic boundaries, automatically built thesauri). (Hearst, 1992) defines a strategy whereby a set of carefully identified "lexico-syntactic" patterns (e.g., *such NP as*, *NP or other NP*, *NP and other NP*, *NP including NP*, *NP especially NP* etc.) are used as indicators of a hyponym relation holding between two NP's in a naturally-occurring text. The reader should note that, unlike statistical techniques such as those mentioned above, in this approach only one sample need be found in order to determine a salient relationship. Hearst's methodology is similar to Brent's in its effort to distinguish clear pieces of evidence from ambiguous ones. The assumption is that, given a large enough corpus, the algorithm can afford to wait until it encounters clear examples. Even a more granular lexicographic analysis can profit from the use of co-occurrence analysis as carried out on large text corpora. This point is strongly made by (Sinclair, 1991). Most actual lexicographic examples, as extracted from corpora through concordance routines, are in fact unrepresentative of the pattern of the word or phrase for which they have been chosen. The vast majority of them can thus be safely discarded when their statistical contribution to the concordance as a whole has been recorded. Therefore, it is necessary to have access to a large corpus because the normal use of language is highly specific, and good representative examples are hard to find. Accordingly, a procedure is defined to locate 'citation forms' for linguistic and lexicographic purpopes, and carve them out of the body of 'non-citation forms'. The procedure begins with a machine-generated concordance of a large corpus. A list is then compiled in frequency order of all the word-forms in the concordance. These are called the collocates of the key-word. Very infrequent collocates are then cut off. Each of the remaining collocates is given a weighting by relating its frequency in the concordance to its overall frequency in the full corpus. So a common word gets a low rating, and a word which makes a distinctive collocation with the node will score high. The concordance is now re-sorted in order of typicality: the most typical instances should come at the top. From this point onwards an automatic procedure is not yet fully established, and the study continues largely on a subjective basis. Finally, a number of unexpected insights can be gained by running the "converse" of a KWIC concordance program (so-called CIWK, Arad 1991) on a comparatively small sample corpus representative of a specific sublanguage. CIWK shows groups of words which occur in a pre-defined context, as in the following:

TO MATCH THE * PORTION

where the asterisked position can be taken, in the restricted domain of UNIX manuals, by either *leftmost* or *rightmost*. CIWK is a useful trigger for the discovery of the linguistic structure of a certain Knowledge Domain for both NLP purposes and information retrieval operations (Tsujii et al., 1992).

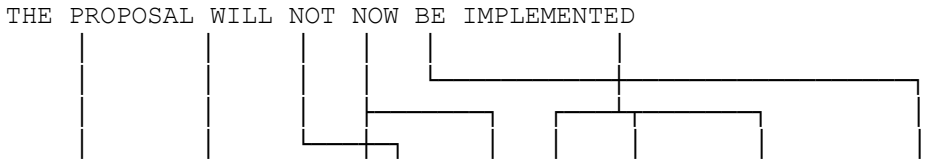
3.3 *Speech applications*

The interest in recorded spoken corpora is growing in the field of Phonology/Phonetics, as witnessed by the important work carried out in the framework of EC-funded projects such as SAM and SUNDIAL. Here, we only mention possible uses of written corpora to enhance the performance of speech recognition systems. We have already illustrated the contribution of direct models of LKE for Morphosyntax to the task of spoken word recognition (Nakamura et al., 1990). (Hosaka and Takezawa, 1992) evaluate the incidence of corpus-based syntactic rules on the performances of a Japanese speech recognition model. They describe phrase-based syntactic rules which are used as constraints in the Japanese speech recognition module which is used in an experimental speech-to-speech translation system. In devising rules, they take into account the error tendency in speech recognition. They end up treating precisely those syntactic categories which tend to be recognized erroneously. To increase the efficacy of each rule, rule construction is strongly motivated by an existing dialogue corpus. By applying corpus-driven, phrasal rules, the speech recognition rate for the top candidates in the sample dialogue corpus improves from 37.2% up to 70.1%, and for the top 5 candidates from 73.7 % up to 83.9%.

3.4 *Machine Translation*

3.4.1 *Statistically based MT*

A "source language Model" and a "translation Model" furnish a probability distribution over source-target pairs (S, T). The joint probability P(S,T) of the pair (S,T) is the product of the probability P(S) computed by the language model and the conditional probability P(T|S) computed by the translation model. The parameters of these models can be estimated automatically from a large database of source-target sentence pairs (bilingual corpus) using a statistical algorithm which optimizes the fit between the models and the data, that is, which maximizes the probability of a sentence S and its translation T co-occurring in the same corpus. In (Brown et al., 1991), a trigram model is used as "source language model". As to the translation model, we can construct it statistically by starting from a bilingual corpus of a suitable size. (Brown et al., 1991) perform their experiments on the bilingual proceedings of the Canadian Parliament (called Hansards), which are kept in both English and French. In Hansards, however, sentences are not conveniently paired; it is possible, however, to produce a parallel text where each sentence in the English proceedings is identified, and paired with exactly one sentence in the French proceedings (see Church, 1991, and, for a different approach, Picchi et al., 1991). However, what is needed for parameter estimation is a set of "aligned" sentences, one example of which is illustrated below:



In order to express alignment relations between two sentences, we need to know:

1)a set of so-called *fertility* probabilities P(n|e) for each English word *e* and for each *fertility n* from 0 to some moderate limit: fertility indicates the number of French words that an English word produces in a given alignment;
2)a set of translation probabilities P(f|e), one for each member *f* of the French dictionary and each member *e* of the English one;
3)a set of distortion probabilities P(i|j,l) for each target position i, source position j, and target length l.

These parameters are generally unknown. The method used for their estimation is called the *forward-backward* algorithm.

3.4.2 *Example-based MT*

Statistical approaches to MT are not the only ones using large bilingual databases as a spring-board. So-called example-based Machine Translation has gathered considerable momentum over the last ten years, and has so far shown promising results, at least as some sort of fall-back strategy that can be resorted to when classical rule-based approaches fail. In (Sumita et al., 1990) the translation of Japanese noun phrases of the form *N1 no N2* into English noun phrases is demonstrated by the use of examples. Sumita suggests using examples of the configuration *N1 no N2* taken from a bilingual database of parallel texts and measuring the distance between a given input and matching examples. Such a distance is expressed as a linear sum of the distances between the subparts of the corresponding expressions multiplied by their respective weights. In case of lack of full matching (when the input noun phrase is not already present in the example database), the distance between two lexical items is calculated by using a thesaurus and a set of weights derived from the database of example parallel texts. (Sato and Nagao, 1990) illustrate a more general strategy which revolves around the same basic idea: translate a source sentence by imitating the translation example of a similar sentence in a bilingual database. In many cases, they observe, it is necessary to imitate more than one translation example and combine some fragments of them. Take the translation of the sentence "He buys a book on international politics". If we happen to have examples i) and ii) below in our bilingual database, then we can translate *He buys a book on international politics* by imitating the translation of i) and ii) and combining the resulting fragments:

i)**He buys** a notebook
Kare ha nouto **wo kau**

ii)I read **a book on international politics**
Watashi ha **kokusaiseiji nitsuite kakareta hon** wo yomu

The process would give the following translation:

Kare ha kokusaiseiji nitsuite kakareta hon wo kau.

The algorithm can be factored out into three steps:

1)parallel translation examples are converted into dependency trees, and links associating each node in one example with the corresponding node in the other example are set up.
2)the input sentence is also turned into a dependency tree; in case of failure of full overlapping between the input sentence and translation examples in the bilingual database, parts of the dependency tree of the input sentence are matched against parts of the dependency trees of the translation examples; a "clone" dependency tree of the input sentence is thus generated, as the result of the process of "cutting and gluing" already existing translation-example dependency trees.
3)nodes in the input clone are replaced by their translational equivalents as defined by translation links in the example pairs, and a thorough checking of the well-formedness of the resulting translation is performed by comparison with already existing structures in the example database of the target language.

A similar approach, based on a labelled directed graph architecture (which allows the expression of both syntactic similarities between input graphs and examples in terms of dependency relations, and semantic similarities in terms of type hierarchies) is illustrated in (Watanabe, 1992). At this juncture, however, the problem arises as to how example pairs can be constructed automatically, in such a way that a full dependency analysis of them is provided, and, even more crucially, an accurate set of links between Japanese and English nodes is defined. (Kaji et al., 1992) illustrate an algorithm whereby the process of machine learning of so-called translation templates (that is, translation pairs where some content words are replaced by variables) from bilingual texts is modelled. The entire process thus consists of two phases: the creation of translation templates and their use in the translation phase. The algorithm is based on the assumption that syntactic ambiguities cannot be resolved completely in the syntactic parsing. Syntactic ambiguities are resolved during the mapping process, together with other types of ambiguities such as ambiguities in the correspondence between words. The algorithm however is based on surface constituent analyses, by which phrases are formed through concatenation of adjacent words. This is inconvenient for languages like Japanese in which word order is flexible, and makes translation templates far less suitable than the matching dependency sub-trees developed by Sato and Nagao. Extensive use of multilingual databases for MT purposes is documented by the work carried out in the framework of a feasibility study on multilingual lexicography, funded by the Council of Europe, and co-ordinated by Sinclair (1990-91), on which more in (Baker, 1993, NERC-143). A number of very common words of mixed word classes were selected for each language involved, on the basis of their prima facie translation equivalence: e.g. *say* (English), *sagen* (German), *dire* (Italian), *saga* (Swedish). The idea was to explore regularities in the textual environment of each member of an equivalence pair by studying concordances in the various languages, in order to detect a method for the automation of translation correspondences based on: valency, word order, use of anaphors, quantifiers and other determiners, prepositions, negation etc.

3.5 *Information retrieval*

In conventional information retrieval, the stored records are normally identified by a set of keywords or phrases known as index terms. Requests for information are typically expressed by Boolean combinations of index terms. However, conventional Boolean logic is rigid in a retrieval setting mainly because it treats all terms as equally important and all retrieved documents as equally useful. In a vector processing system, Boolean queries are replaced by multidimensional queries. If *t* distinct terms are available for identification, a document D_i is representable as a *t*-dimensional vector of pairs, D_i (d_{i1}, w_{i1}; d_{i2}, w_{i2}; ... ; d_{it}, w_{it}) where d_{ij} represents the jth terms assigned to documents D_i and w_{ij} is the corresponding term weight. When queries are expressed accordingly as *t*-dimensional vectors, then a global composite vector comparison can measure the degree of similarity between a query-document pair on the basis of the weights of the corresponding matching terms.

All such approaches, however, start from a pre-defined, closed list of terms, and completely ignore the problem of assessing their relevance with respect to the specific query at hand, or, analogously, the relevance of the document being retrieved by term-matching to the expectations of the query user. The needed term probabilities can be estimated by accumulating a number of user queries containing term T_k and determining the porportion of times document D_i is found relevant to the respective queries. The expected usefulness of a retrieval system is optimized when the item with the highest probability of relevance is extracted from the file at each point. In order to achieve this optimization, large quantities of documents representative of different classes are needed, for the process of extracting the set of more relevant terms to converge onto an optimal solution. An optimal query vector can usefully be seen as multidimensional, special class representatives, or *centroids*, around which all relevant document items which belong to a certain class revolve. Centroids are also exploited in retrieval strategies, whereby groups or clusters of documents can be built. File searches are then confined to those document clusters whose centroids exhibit large "query-centroid" similarities.

Another possibility for the formulation of viable text analysis systems that are valid for unrestricted text environments is to perform detailed analyses of the available texts and to incorporate in the analysis process the multiple contexts in which the words and expressions are used in the available texts. Similarities in word meanings can then be ascertained by the determination of coincidences in the contexts in which the words are used in different text passages. When sufficiently large contextual similarities are detected, the conclusion follows that the word meanings in the corresponding texts are homogeneous (Salton and Buckley, 1991). There are indications that such methods can operate with a certain degree of accuracy. It is clear that extensive applications would require the availability of representative text samples, and of fairly sophisticated memory-based strategies. (Morris and Hirst, 1991) deal with the notion of lexical cohesion that arises from the semantic relationships between words in text, and compute it by finding so-called lexical chains, that is chains of related words that contribute to the continuity of lexical meaning. These lexical chains are the result of units of text being "about the same thing", and consist of sequences of related words which co-occur within a given span. Lexical chains do not stop at sentence boundaries. They can connect a pair of adjacent words or range over an entire text. Lexical chains are computed by using an abridged version of *Roget's Thesaurus* (1977) and a set of candidate words which excludes: a) repetitive occurrences of closed-class words such as pronouns, prepositions, and verbal auxiliaries; b) high-frequency words like *good*, *do*, and *taking*. The two most frequently used thesaural relations are the

following:

1)two words have a category in common in their index entries. For example, *residentialness* and *apartment* both have category 189 in their index entries;
2)one word has a category in its index entry that contains a pointer to a category of the other word. For example *car* has category 273 in its index entry, and that contains a pointer to category 276, which is a category of the word *driving*.

There are two main reasons why lexical cohesion is important for computational text understanding systems:

a)lexical chains provide an easy-to-determine context to aid the resolution of ambiguity and the narrowing of a word to a specific meaning.
b)lexical chains provide a clue to the determination of coherence and discourse structure, and hence the larger meaning of the text.

(Hoey, 1991) illustrates the use of "lexical cohesion" in text for the automatic abridgement of non-narrative texts. The underlying principle is that the most important sentences of a text are those having the maximum number of "bonds" with other parts of the text. Bonds are created when two sentences share a threshold level of lexis, the threshold being variable according to the type of text.

Another promising application to specific knowledge domains is illustrated in (Coultard, 1993, NERC-142), where police reports, police interviews and the like are utilized for the attribution of authorship to disputed texts, through an assessment of the linguistic similarities over a significant amount of forensic texts.

3.6 *Summary and Conclusion*

Amount of textual data required

In order to analyse, describe, produce or understand language for NLP applications, one major hurdle has to be overcome: language ambiguity. Ambiguity (both lexical and structural ambiguity) is pervasive in language, but not ubiquitous. If enough text is considered, the chances are that an unambiguous instance of a certain construction will be encountered. Such an instance can be capitalized on, and used to tackle ambiguous cases. For example, knowledge extraction techniques by pattern matching are successful in recovering a fair amount of non-trivial semantic information with a minimum of string processing. Pattern matching is a fairly primitive parsing resource: but if enough textual material is provided, the number of successful matches is likely to be significantly high, and results can be generalized. By the same token, subcategorization frames can, in many cases, be neatly isolated if they are tracked down in those contexts which allow for one (unambiguous) interpretation only (for example, in English, if and when pronouns are present, which are uniquely marked for case). Real world knowledge associations necessary for interpreting a sentence like "I saw the man with a scarf" can be retrieved in texts, as long as different enough contexts of the use of analogous

expressions are provided, and either ambiguous or unrepresentative contexts are discarded.

In corpus-oriented work, stress is currently being laid on **tools for corpus exploitation**. All of them require the availability of massive quantities of textual data for processing. Quantity is not only a pre-requisite of statistically-based approaches, but of virtually any corpus-based process of linguistic knowledge elicitation. In a sense, it is literally true that the more data the better. However, quantity is not the only dimension that matters in corpus-oriented work and related applications.

Work on sublanguages: the need for domain-specific texts in NLP

Sublanguage-oriented research can be looked at from an analogous perspective: in sublanguage domains, linguistic ambiguity is drastically reduced due to the context provided by the particular topic dealt with in text. If specific enough contexts are considered, only certain readings of a word/construction are likely to be found. It has been argued that, for some hardly tractable NLP problems like compound intepretations for MT, results can be obtained only for fairly constrained subdomains. Self-modelling systems attain reasonable performances when trained on texts containing a limited amount of lexical ambiguity, and presupposing a specific world-knowledge domain. The availability of texts coming from well-selected domain-specific areas in a variegated corpus is indispensable for this type of demand to be met. Sublanguage-oriented investigations are too important for practical applications to be neglected.

A further dimension to corpus composition: text genres

For some specific applications, like speech recognition, it makes sense to concentrate on those linguistic phenomena which are likely to take place in dialogue exchanges, thus narrowing down the area of investigation in a non-arbitrary, effective way. It is recommended, therefore, that different genres be suitably represented in a corpus, to be recovered on demand according to the specific research needs in question.

Multilingual parallel corpora

Recent interesting applications in so-called example-based, or case-based Machine Translation have shed light on the importance of paralleled multilingual corpora for linguistic knowledge elicitation. Their use has not been limited to MT-oriented applications, so that multilingual corpora can now be regarded as free-standing sources of linguistic knowledge in their own right for a number of useful applications (word-sense disambiguation, lexicon construction, etc.).

Corpus-based work: a fast-growing area

Corpus-based linguistics has gathered considerable momentum over the last few years, as shown by the sheer number of papers which have been devoted to work on real texts. To give but one example, the Proceedings of COLING '92 contain over twenty papers which document work on corpora. This figure would go up considerably if papers whose content is related, albeit indirectly, to corpus-work were counted in. A growing number of private companies have shown interest in corpus-related

applications. Again, if we consider only those companies whose research labs have been concretely involved in corpus-work documented by contributions/submissions to the last COLING, the list would include AT&T Bell Labs., NEC, XEROX (Palo Alto Research Center), ATR, HITACHI, IBM and others.

Rule-based approaches and corpus-work

Stress should be laid on an often neglected point in the on-going debate between rule-based NLP systems and corpus-work: the two strategies, far from being impervious to mutual integration, have recently discovered a vast and yet mostly unexplored common ground where a number of unsolved cruces in NLP can be promisingly tackled through a complementary approach. This is particularly evident in those research areas in which concrete applications have pride of place over more theoretical interests. For example, there are relatively few either purely "empirical" approaches to MT, or purely "rule-based" ones. A term has been coined (*hybrid approaches to MT*) to refer to those systems which involve a traditional rule-based core and add-on modules based on more empirical techniques. Examples of such extensions would be statistically based preference mechanisms, and large scale lexical and corpus-based resources deployed essentially as bilingual lexical disambiguation components. Generally speaking it has been acknowledged that two major bottlenecks for rule-based NLP systems are the laborious process of "acquisition of linguistic information", and the elusive nature of language ambiguity. Both aspects have been thrown into sharp relief throughout this overview. We have reasons to believe that problems of lexical selection in generation, or lexical disambiguation in analysis can be more appropriately and efficiently solved through corpus-based techniques like those reviewed in the second section of this study, than by setting up a framework of conceptual primitives, which is still a huge undertaking for any non-trivial domain, at least in the relatively short-term.

Tools, uses and annotated corpora

A shift of perspective has recently emerged in NLP from "exhaustive parses" (according to a certain existing grammatical standard) to "partial analyses", seen as intermediate steps in the process of gradual approximation to wide-coverage NLP systems. It has been argued that there are simply too many things which we still ignore about language to enable us to embark on the daunting task of carrying out a full (mostly manual) either syntactic or semantic parse of a whole corpus. Two trends seem to provide interesting indications about promising developments in the field: a) the design of so-called "bootstrapping" systems, which start from fairly crude processing stages and gradually reach considerable complexity and wealth of linguistic information; b) the development of case-based strategies, which stress the importance of storing chunks of text as such, rather than chopping them into more elementary but less significant linguistic units, in order to preserve to the largest possible extent the context-specificity of language. Both a) and b) picture a scenario where such tools will be used on domain-specific corpora for particular applications, in order to arrive at linguistic representations defined at the level of granularity required by the target application. It remains to be seen whether the extracted knowledge should be annotated in the text and made available with the text through circulation, or just elicited on demand. Be that as it may, it will be vital to provide researchers with a battery of tools for corpus interrogation.

References

Alshawi H. (1989): "Analysing dictionary definitions", in: B. Boguraev, T. Briscoe (eds.) *Computational Lexicography for NLP*, Longman, London.

Alshwede T., Evens M. (1988): "Generating a Relational Lexicon from a Machine Readable Dictionary", in *IJL*, vol.1, 214-37.

Atwell E. (1985): "Constituent-likelihood grammar", in R. Garside, G. Leech and G. Sampson (eds.) *The Computational Analysis of English*, Longman.

Bahl L., Jelinek F., Mercer R. (1983): "A Maximum Likelihood approach to continuous speech Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol PAMI-5.

Baker J.K. (1979): "Trainable Grammars for speech recognition", in *Speech Communication Papers for the 97th Meeting of the Acoustic Society of America*.

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 *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, Oxford.

Black A.W., van de Plassche J., Williams B. (1992): "Analysis of Unknown Words through morphological Decomposition", in *COLING 92 Proceedings*, Nantes.

Brasington R., Biggs C., Jones S. (1987): "Automated knowledge Acquisition: a linguistic paradigm". Paper for the Symposium on Knowledge Acquisition and Knowledge Abstraction Communication and Cognition 20, Ghent, Belgium.

Brent M.R. (1991): "Automatic Acquisition of Subcategorization Frames from Untagged, Free-Text Corpora", in *Proceedings of the 29th meeting of the ACL*.

Brill E. (1992): "A simple Rule-based Part of Speech Tagger", in Proceedings of the *Third Conference on Applied Natural Language Processing*, Trento.

Brill E., Magerman D., Marcus M., Santorini B. (1991): "Deducing Linguistic Structure from the Statistics of Large Corpora".

Briscoe T., Carroll J. (1991): "Generalised Probabilistic LR Parsing of Natural Language (Corpora) with Unification-based Grammars", University of Cambridge, Technical Report No. 224.

Brown P.F., Cocke J., Della Pietra S.A., Della Pietra V.J., Jelinek F., Lafferty J.D., Mercer R.L.,

Roossin P.S. (1990): "A Statistical Approach to Machine Translation", *Computational Linguistics* 16:2.

Brown P.F., Della Pietra V.J., DeSouza P.V., Lai J.C. (1990): "Class-based n-gram models of natural language".

Calzolari N. (1984): "Detecting Patterns in a Lexical Data Base", in *Proceedings of COLING-84*, Stanford University, California.

Calzolari N., Bindi R. (1990): "Acquisition of Lexical Information from a large textual Italian corpus", in H.Kalgren (ed.), *Proceedings of COLING 90*, Helsinki.

Chodorow M.S., Byrd R.J., Heidorn G.E. (1985): "Extracting Semantic Hierarchies from a Large on-line Dictionary", *ACL Proceedings*.

Church K., Gale W. (1991): "Concordances for Parallel Texts", in *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, Oxford.

De Kock J., Bossaert W. (1978): "The Morpheme", Van Gorcum, Amsterdam.

DeRose S. (1988): "Grammatical Category Disambiguation by Statistical Optimization", *Computational Linguistics*, 14:1.

Evens M.W. (editor) (1988): "Relational Models of the Lexicon: Representing Knowledge in semantic networks", CUP, Cambridge.

Federici S., Pirrelli V. (1991a): "Tagger SECS: a Neural Environment for Corpus-driven Tagging", unpublished.

Federici S., Pirrelli V. (1991b): "Doing Morphology without Rules: an Approach to Linguistic Knowledge Acquisition Through Examples", Internal Report, ILC-CNR NLP-1991-2.

Federici S., Pirrelli V. (1992): "A Bootstrapping Strategy for Lemmatisation: Learning through Examples", in *Proceedings of COMPLEX-92*, Budapest.

Garside R., Leech G., Sampson G. (eds.) (1985): *The Computational Analysis of English*, Longman.

Garside R., Leech G. (1985): "The UCREL probabilistic parsing system", in R. Garside, G. Leech, G. Sampson (eds.), *The Computational Analysis of English*, Longman.

Gilloux M. (1991): "Automatic Learning of Word Transducers from Examples".

Grishman R., Sterling J. (1992): "Acquisition of Selectional Patterns", in *Proceedings of COLING 92*,

Nantes.

Guthrie J., Guthrie L., Wilks Y., Aidinejad H. (1991): " Subject-dependent Co-occurrence and Word sense disambiguation", in *Proceedings of ACL*.

Hearst M.A. (1992): "Automatic Acquisition of Hyponyms from Large Text Corpora", in *COLING 92*, Nantes.

Hoey M. (1991): "Patterns of Lexis in Text", Oxford University Press.

Hosaka J., Takezawa T. (1992): "Construction of corpus-based syntactic rules for accurate speech recognition", in *COLING 92*, Nantes.

Jelinek F. (1990): "Self-organized Language Modeling for Speech Recognition", in Weibel and Lee (eds.), *Readings in Speech Recognition*, Morgan Kaufmann Publishers, California.

Jelinek F., Mercer R. (1980): "Interpolated estimation of Markov source parameters from sparse data", in *Proceeding of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North Holland.

Jensen K., Binot J.L. (1987): "Disambiguating Prepositional Phrase Attachments by using on-line Dictionary Definitions", *Computational Linguistics*, 13.

Kaji H., Kida Y., Morimoto Y. (1992): "Learning Translation Templates from Bilingual Text".

Leech G. (1992): "Manual, Automatic and Machine-Assisted Corpus Annotation: the Lancaster Experience", in *Proceedings of the International Workshop on Fundamental Research for FGNLP*, Manchester.

Marinai E., Peters C., Picchi E. (1991): "Bilingual Reference Corpora: a System for Parallel Text Retrieval", in *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, Oxford.

Marshall I. (1985): "Tag selection using probabilistic methods", in R. Garside, G. Leech, G. Sampson (eds.), *The Computational Analysis of English*, Longman.

Martin P. (1990): "Automatic Assignment of Lexica Stress in Italian", in *Proceedings of the ESCA Workshop on Speech Synthesis*.

Martin W. (1992): "Concept-oriented parsing definitions", in *COLING 92*, Nantes.

Markovitz J., Ahlswede T., Evens M. (1986): "Semantically significant patterns in dictionary definitions", in *Proceedings of the 24th Annual Meeting of the ACL*, pp.112-119.

Morris J., Hirst G. (1991): "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", in *Computational Linguistics*, vol.17 n.1.

Nakamura M., Maruyama K., Kawabata T., Shikano K. (1990): "Neural Network Approach to Word Category Prediction for English Texts", in *Proceedings of COLING 90*, Helsinki.

Pirrelli V., Federici S. (forthcoming): "An analogical way to Language Modelling: MORPHEUS", in *ACTA HUNGARICA*".

Redijk M. (1991): "A conceptual parser for definitions of medical terms", Free University of Amsterdam.

Roe D., Pereira F., Sproat R., Riley M., Moreno P., Macarron A. (1991): "Toward a spoken language translator for restricted-domain context-free languages", in *EUROSPEECH 91*.

Salton G. (1992): "Developments in Automatic Text Retrieval", in *Science*, vol.253, pp.974-80.

Salton G., Buckley C. (1991): "Global Text Matching in Information Retrieval", in *Science*, vol. 252, pp. 1012-15.

Sanfilippo A., Poznanski V. (1992): "The Acquisition of Lexical Knowledge from Combined Machine-Readable Dictionary Sources", in Proceedings of the *Third Conference on Applied Natural Language Processing*, Trento.

Sato S., Nagao M. (1990): "Toward Memory Based Translation", in *COLING 90*, Helsinki.

Sekine S., Carroll J.J., Ananiadou S., Tsujii J. (1992a): "Automatic Learning for Semantic Collocation", in Proceedings of the *Third Conference on Applied Natural Language Processing*, Trento.

Sekine S., Carroll J.J., Ananiadou S., Tsujii J. (1992b): "Linguistic Knowledge Generator", in *Proceedings of COLING 92*, Nantes.

Sinclair J. (1991): *Corpus Concordances and Collocations*, Oxford University Press.

Tsujii J., Ananiadou S., Arad I., Sekine S. (1992): "Linguistic Knowledge Acquisition from Corpora", in Proceedings of the *International Workshop on Fundamental Research for FGNLP*, Manchester.

Utsuro T., Matsumoto Y., Nagao M. (1992): "Lexical Knowledge Acquisition from Bilingual Corpora", in *COLING 92*, Nantes.

Watanabe H. (1992): "A similarity-driven transfer", in *COLING 92*, Nantes.

Wolff (1984): "The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding", *Methods of Information in Medicine* 23: 195-203.

Wothke K. (1986): "Machine Learning of Morphological Rules by Generalization and Analogy".

Zhang B.T., Kim Y-T. (1990): "Morphological analysis and synthesis by automated discovery and acquisition of linguistic rules", in *COLING 90*, Helsinki.

Relevant NERC Papers

Baker M. (1993): "Multilingual Databases", Working Paper, COBUILD, NERC-143.

Coultard M. (1993): "On the Need for a Corpus of Forensic Texts", Working Paper, COBUILD, NERC-142.

Hoey M. (1993): "Report on a research funded by British Telecom into automatic abridgement of non-narrative text", Working Paper, COBUILD, NERC-144.

Pirrelli V. (1993): "Report on Knowledge Extraction", Technical Report, ILC Pisa, NERC-79.

General NERC Bibliography

[1] M. Alvar. "Recommendations for a corpus of Spanish". Strategic paper, University of Malaga, December 1992. NERC-STRATPAP-139.

[2] M. Alvar-Ezquerra and G. Corpas-Pastor. "Design criteria". Working paper, University of Malaga, October 1992. NERC-WP6-84.

[3] M. Antona. "A comparison of EUROTRA ECS grammars". Working paper, ILC Pisa, December 1992. NERC-WP8-64.

[4] M. Antona. "The treatment of subordinate clauses in EUROTRA: an overview". Working paper, ILC Pisa, December 1992. NERC-WP8-63.

[5] M. Baker. "Multilingual Databases". Working paper, COBUILD Birmingham, February 1993. NERC-WP10-143.

[6] C. Belica. "The TEI proposal and its feasibility. An evaluation in respect to the IDS corpus philosophy". Working paper, IDS Mannheim, September 1992. NERC-WP3/WP4-36.

[7] R. Bindi, N. Calzolari, M. Monachini, and V. Pirrelli. "Lexical knowledge acquisition from textual corpora: a multivariate statistic approach as an integration to traditional methodologies". In *7th Annual Conference of the UW Centre for the NEW OED and Text Research, Using Corpora*, pages 170–196, Oxford, September 1991. Oxford University Press. NERC-WP9/WP10-103.

[8] R. Bindi, N. Calzolari, M. Monachini, V. Pirrelli, and A. ampolli. "Corpora and Computational Lexica: Integration of Different Methodologies of Lexical Knowledge Acquisition". Paper presented at the pisa workshop, ILC Pisa, January 1992. NERC-WP9/WP10-177.

[9] R. Bindi and M. Monachini. "Building corpora: text acquisition". Working paper, ILC Pisa, May 1992. NERC-WP7-58.

[10] R. Bindi, M. Monachini, and P. Orsolini. "Italian reference corpus: key for consultation". Working paper, ILC Pisa, 1991. NERC-WP7-13.

[11] Birmingham. "Access and Management Software Tools". Technical report, University of Birmingham, December 1992. NERC-WP5-155.

[12] Birmingham. "Annotation Tools". Technical report, University of Birmingham, December 1992. NERC-WP9-138.

[13] Birmingham. "Byte and Type: Cobuild corpus coding convention". Working paper, University of Birmingham, July 1992. NERC-WP3-40.

[14] Birmingham. "Text Representation: Spoken Language". Technical report, University of Birmingham, December 1992. NERC-WP4-156.

[15] Birmingham. "WP5 access and management software tools". Interim technical report, University of Birmingham, October 1992. NERC-WP5-98.

[16] M.J. Blanco-Rodriguez. "Criteria for Morphosyntactic Labelling of Spanish". Working paper, University of Malaga, July 1992. NERC-WP8-112.

[17] A. Bon and W. Martin. "On semantic tagging". Working paper, INL Leiden, June 1992. NERC-WP8-70.

[18] L. Burnard. "TGC W30 Corpus Document Interchange Format v.1.0". Working paper, Oxford University Computing Service, March 1992. NERC-WP4-166.

[19] L. Burnard. "The Text Encoding Initiative: A progress Report". Working paper, Oxford University Computing Service, 1992. NERC-WP4-167.

[20] L. Burnard and C.M. Sperberg-McQueen. "Guidelines for Text Encoding and Interchange". Working paper, ACH, ACL, ALLC, April 1992. NERC-WP4-162.

[21] N. Calzolari. "Issues for lexicon building". Strategic briefing paper, NERC Consortium, June 1992. NERC-STRATLEX-100.

[22] D. Candel. "Rapport sur les besoins exprimés par les utilisateurs virtuels de corpus linguistiques français". Working paper, INALF Paris, September 1992. NERC-WP2-31.

[23] D. Candel. "Traitement de corpus oraux - Groupe Aixois de recherche en syntaxe (Notes on Transcription)". Working paper, INALF Paris, December 1992. NERC-WP4-157.

[24] D. Candel. "Classification des documentes textuels par domaines". Working paper, INALF Paris, February 1993. NERC-WP6-127.

[25] M. Castillo-Cabezas. "Tool-set for the corpus processing in PC". Working paper, University of Malaga, October 1992. NERC-WP5-55.

[26] R. Cauldwell. "Accessing spoken data using compact disc and hypertext". Working paper, University of Birmingham, January 1992. NERC-WP4-48.

[27] R. Cauldwell. "Johansson et al's Proposals". Working paper, University of Birmingham, March 1992. NERC-WP4-106.

[28] R. Cauldwell. "Start the Week - using TEI P2 34 codes". Working paper, University of Birmingham, May 1992. NERC-WP4-107.

[29] J. Channel. "The coding and extraction of pragmatic information in a dictionary databse". Working paper, University of Birmingham, December 1992. NERC-WP8-149.

[30] J. Clear. "Corpus Availability". Working paper, COBUILD Birmingham, February 1993. NERC-WP5-154.

[31] J. Clear. "Corpus Maintenance". Working paper, COBUILD Birmingham, February 1993. NERC-WP5-153.

[32] D. Coniam. "Boundary marker: system description". Working paper, University of Hong Kong, September 1991. NERC-WP9-21.

[33] D. Coniam. "Different parsing strategies and parsing models". Working paper, University of Hong Kong and University of Birmingham, October 1992. NERC-WP9-77.

[34] O. Corazzari. "Phraseological units". Working paper, ILC Pisa, December 1992. NERC-WP8-68.

[35] G. Corpas-Pastor. "Semantic annotation". Working paper, University of Malaga, October 1992. NERC-WP8-62.

[36] G. Corpas-Pastor. "Semantic knowledge extraction". Working paper, University of Malaga, December 1992. NERC-WP10-120.

[37] P. Cotoneschi and M. Monachini. "An empirical experience on the utilization of the Italian reference corpus in meaning analysis". In D. Ross and D. Brink, editors, *ACH-ALLC '91, Making Connections*, pages 77–80, Tempe, March 1991. Arizona State University. NERC-WP10-102.

[38] M. Coultard. "On the Need for a Corpus of Forensic Texts". Working paper, University of Birmingham, February 1993. NERC-WP10-142.

[39] H.M. Dahan. "Beyond the exchange: a look at communicative competence". Working paper, ITM Malaysia and University of Birmingham, 1991. NERC-WP8-17.

[40] E.D. de Jong. "Transcription and normalization method. Dutch spoken language". Working paper, Utrecht, December 1992. NERC-WP4-159.

[41] J. Dendien. "Acces a l'information dans une base textuelle Functions d'acces et index optimaux". Working paper, INALF Paris, 1991. NERC-WP5-113.

[42] M.W.F. Dutilh and J.G. Kruyt. "Feasibility experiment design criteria: Investigation into text typological classification tools". Working paper, INL Leiden, November 1992. NERC-WP6-94.

[43] M. Dutilh-Ruitenberg. "Corpus annotation schemes in the Netherlands". Working paper, INL Leiden, June 1992. NERC-WP8-69.

[44] J.A. Edwards. "Design Principles in the Transcription of Spoken Discourse". Working paper, University of California, December 1992. NERC-WP4-164.

[45] B. Endres and F. Wagner. "Synoptic Report on the Needs of Corpus Users". Interim technical report, IDS Mannheim, December 1992. NERC-WP2-119.

[46] S. Federici and V. Pirrelli. "Tagger SECS: a neural environment for corpus-driven tagging". Working paper, ILC Pisa, November 1991. NERC-WP9-20.

[47] G. Francis. "The Helsinki tagger and parser". Working paper, COBUILD Birmingham, September 1992. NERC-WP9-78.

[48] G. Francis. "A Look-up Tagger". Working paper, COBUILD Birmingham, February 1993. NERC-WP9-176.

[49] J.P. French. "Updated notes for soundprint transcribers". Working paper, University of Birmingham, October 1991. NERC-WP4-47.

[50] J.P. French. "Transcription proposals: multi-level system". Working paper, University of Birmingham, October 1992. NERC-WP4-50.

[51] J.M. Garcia-Platero. "Answers to the questionnaire about textual corpora needs in Spain". Working paper, University of Malaga, October 1992. NERC-WP2-28.

[52] M. Hoey. "Report on a research project funded by British Telecom into the automatic abridgement of non-narrative text". Working paper, University of Birmingham, February 1993. NERC-WP10-144.

[53] Y. Huizhong. "A new technique for identifying scientific/technical terms and describing science texts". *Literary and Linguistic Computing*, 1(2):93–103, 1986. NERC-WP6-95.

[54] . James. "Cobuild use of the UNIX Operating system". Working paper, COBUILD Birmingham, October 1992. NERC-WP5-89.

[55] . James. "Enhancement of standard facilities (eg annotation tools)". Working paper, COBUILD Birmingham, December 1992. NERC-WP5-130.

[56] S. Johansson, L. Burnard, J. Edwards, and A. Rosta. "TEI - working paper on Spoken Texts". Working paper, October 1991. NERC-WP4-165.

[57] T. Johns. "Corpus linguistics in a pc-compatible environment". Working paper, COBUILD Birmingham, October 1992. NERC-WP5-90.

[58] F. Karlsson. "Frequency considerations in morphology". *PSK*, 39(1):19–28, 1986. NERC-WP8-73.

[59] F. Karlsson. "Lexicography and corpus linguistics". Opening address at 5th congress of EURALEX 92, August 1992. NERC-WP8-74.

[60] F. Karlsson. "SWETWOL: a comprehensive morphological analyser for Swedish". *Nordic Journal of Linguistics*, 15:1–45, 1992. NERC-WP8-75.

[61] F. Karlsson, A. Voutilainen, A. Anttila, and J. Heikkila. "Constraint grammar: a language-independent system for parsing unrestricted text, with an application to English". Helsinky, 1992. NERC-WP8-92.

[62] J.M. Kirk. "The Northern Ireland transcribed corpus of speech". Working paper, Queen's University of Belfast, December 1993. NERC-WP4-158.

[63] R. Krishnamurthy. "Basic access software, section A: word lists". Working paper, COBUILD Birmingham, October 1992. NERC-WP5-87.

[64] R. Krishnamurthy. "Basic access software, section B: basic concordancing 2: the user's perspective". Working paper, COBUILD Birmingham, October 1992. NERC-WP5-88.

[65] R. Krishnamurthy. "Data collection". Working paper, COBUILD Birmingham, September 1992. NERC-WP6/WP7-57.

[66] R. Krishnamurthy. "Pisa TEI workshop: recoding a Cobuild spoken text". Working paper, COBUILD Birmingham, March 1992. NERC-WP4-49.

[67] R. Krishnamurthy. "Recoding a TEI text in Cobuild codes". Working paper, COBUILD Birmingham, March 1992. NERC-WP3-41.

[68] R. Krishnamurthy and S. Smith. "Text Encoding at Cobuild". Working paper, COBUILD Birmingham, January 1992. NERC-WP3-38.

[69] J.G. Kruyt. "Evaluative report on design criteria for corpora construction I: selection principles". Interim technical report, INL Leiden, October 1992. NERC-WP6-93.

[70] J.G. Kruyt. "Evaluative report on design criteria for corpora construction II: availability of text for corpus building". Interim technical report, INL Leiden, December 1992. NERC-WP6-115.

[71] J.G. Kruyt. "Design Criteria for Corpora construction in the Framework of a European Corpora Network". Technical report, INL Leiden, February 1993. NERC-WP6-168.

[72] J.G. Kruyt and E. Putter. "Corpus Design Criteria". Working paper, INL Leiden, December 1992. NERC-WP6-129.

[73] J.G. Kruyt and J.J. van der Voort-van der Kleij. "Papers in Computational Lexicography - Towards a Computerized Historical Dictionary of Dutch: from Printed Dictionary to Correct Text File". Technical report, Budapest, 1992. NERC-WP7-152.

[74] J.G. Kruyt and W. Vercouteren. "Corpora: user needs in the Netherlands". Working paper, INL Leiden, April 1991. NERC-WP2-5.

[75] P. Lafon. "Dictionaires machine et lexicométrie". *Études de Linguistique Appliquée*, pages 85–86, 1992. NERC-WP8-72.

[76] P. Lafon. "Encoding and acquisition techniques: cost evaluations". Interim technical report, INALF Paris, September 1992. NERC-WP7-59.

[77] P. Lafon. "Text representation". Interim technical report, INALF Paris, September 1992. NERC-WP3-37.

[78] P. Lafon and D. Candel. "Entretien avec Simon Sabbagh et ses collaborateurs responsable Eurolang societe site". Working paper, INALF Paris, February 1993. NERC-WP2-145.

[79] P. Lafon and F. Chahuneau. "Acquisition de textes et réutilization". Technical report, INALF Paris and AIS Berger-Levrault, December 1992. NERC-WP7-151.

80] P. Lafon, J. Lefevre, A. Salem, and M. Tournier. *Le Machinal Principes d'enregistrement informatique des textes.* INALF, Paris, 1985. NERC-WP3-8.

[81] P. Lafon and D. Vignaud. "Répresentation des textes écrites". Technical report, INALF Paris and AIS Berger-Levrault, December 1992. NERC-WP3-150.

[82] T. Lane. "Access and Management Software Tools. Task: Development Routines". Working paper, COBUILD Birmingham, November 1992. NERC-WP5-114.

[83] T. Lane. "Basic access software, section B: basic concordancing 1". Working paper, COBUILD Birmingham, October 1992. NERC-WP5-101.

[84] T. Lane. "Data input". Working paper, COBUILD Birmingham, October 1992. NERC-WP7-56.

[85] T. Lane. "TEI headers". Working paper, COBUILD Birmingham, May 1992. NERC-WP3-42.

[86] W.A. Liebert. "Textual Reference Corpora: User needs. A report on the relevant literature in the years (1985-1992)". Working paper, IDS Mannheim, December 1992. NERC-WP2-125.

[87] W.A. Liebert. "Textual Reference Corpora: User needs. Abstracts of main articles from the bibliography Textual Reference Corpora". Working paper, IDS Mannheim, December 1992. NERC-WP2-128.

[88] W.A. Liebert. "Textual Reference Corpora: User needs. Indexed Bibliography of the years 1985 to 1992". Working paper, IDS Mannheim, December 1992. NERC-WP2-126.

[89] W.A. Liebert. "User Needs: survey on IDS corpus users". Technical report, IDS Mannheim, December 1992. NERC-WP2-121.

[90] J.M. Lopez-Guzman. "Acquisition and reusability of material for corpus generation". Working paper, University of Malaga, December 1992. NERC-WP7-83.

[91] J.M. Lopez-Guzman. "Representation levels of written texts". Working paper, University of Malaga, October 1992. NERC-WP3-35.

[92] B. MacWhinney. "CHAT. The CHILDES Project: Tools for analyzing talk". Technical manual, Carnegie Mellon University, 1991. NERC-WP4-44.

[93] B. MacWhinney. "CLAN. The CHILDES Project: Tools for analyzing talk". Technical manual, Carnegie Mellon University, 1991. NERC-WP4-45.

[94] Malaga. "Design of a Spanish corpus within the framework of a European corpus". Working paper, University of Malaga, 1991. NERC-WP6-12.

[95] Malaga. "Questionnaire Juridique (with Spanish Lawyer's answers)". Working paper, University of Malaga, November 1992. NERC-WP11-118.

[96] Mannheim. "Analysis on IDS service activities on corpora (1986-1991)". Working paper, IDS Mannheim, 1991. NERC-WP2-4.

[97] E. Marinai, C. Peters, and E. Picchi. "Bilingual reference corpora: a system for parallel text retrieval". In *7th Annual Conference of the UW Centre for the NEW OED and Text Research, Using Corpora*, pages 63–70, Oxford, September 1991. Oxford University Press. NERC-WP9-104.

[98] M. Monachini. "Italian corpus - Needs of actual and potential users". Working paper, ILC Pisa, June 1992. NERC-WP2-32.

[99] M. Monachini and A. Oestling. "Morphosyntactic corpus annotation - A comparison of different schemes". Technical report, ILC Pisa, September 1992. NERC-WP8-60.

[100] M. Monachini and A. Oestling. "Towards a minimal standard for morphosyntactic corpus annotation". Technical report, ILC Pisa, September 1992. NERC-WP8-61.

[101] M. Monachini and E. Picchi. "A query system for tagged reference corpora". Working paper, ILC Pisa, November 1991. NERC-WP5-10.

[102] M. Monachini and E. Picchi. "Tagged corpora: A query system. In G. Kiss F. Kiefer and J. Pajzs, editors, *Proceedings of the 2nd International Conference on Computational Lexicography, COMPLEX 92*, Budapest, 1992. NERC-WP5-105.

[103] S. Montemagni. "Syntactically annotated corpora: comparing the underlying annotation schemes". Technical report, ILC Pisa, December 1992. NERC-WP8-67.

[104] I. Moreno-Torres. "Tagging textual corpora at the pragmatic level". Working paper, University of Malaga, 1991. NERC-WP8-15.

[105] J. Nakamura. "Determing text typology by means of Hayashi's quantification method type III". Working paper, University of Tokushima and COBUILD Birmingham, July 1992. NERC-WP6-53.

[106] J. Nakamura. "Hayashi's Quantification method type III. A Tool Determing text typology in large Corpora". Working paper, University of Tokushima and COBUILD Birmingham, December 1992. NERC-WP9-137.

[107] J. Nakamura. "On the structure of the Bank of English based on the distribution of pronominal forms". Working paper, University of Tokushima and COBUILD Birmingham, May 1992. NERC-WP6-97.

[108] J. Nakamura. "Quantitative comparison of Modals in the Brown and LOB Corpora". Working paper, University of Tokushima and COBUILD Birmingham, December 1992. NERC-WP9-131.

[109] NERC. "Activity summary report". Report to EC, NERC Consortium, Pisa, December 1991. NERC-MANAGINTR1-80.

[110] NERC. "Activity report". Report to EC, NERC Consortium, Pisa, October 1992. NERC-SUMMREP2-117.

[111] NERC. "Corpus Questionnaire". Working paper, CETH, PISA and MANNHEIM, August 1992. NERC-WP1-111.

[112] NERC. "Introduction (to the 2nd Report)". Report to EC, NERC Consortium, Pisa, October 1992. NERC-MANAGINTR2-116.

[113] NERC. "Policy for corpus provision for Europe". Strategic briefing paper, NERC Consortium, June 1992. NERC-STRATCOR-99.

[114] NERC. "Preliminary draft report". Report to EC, NERC Consortium, Pisa, July 1992. NERC-DRAFTFINREP-81.

[115] NERC. "The Third NERC Interim Report". Report to EC, NERC Consortium, Pisa, December 1992. NERC-MANAGINTR3-160.

[116] NERC. "Workshop on textual corpora". Report to EC, NERC Consortium, Pisa, January 1992. NERC-WORKREP-82.

[117] NERC. "Implementation Plan". Strategic paper, NERC Consortium, February 1993. NERC-STRATPAP-147.

[118] G. Osborne. "CHOC: A Text Analysis Tool". Working paper, Apricot Computers Ltd, Birmingham, February 1993. NERC-WP9-175.

[119] N. Ostler. "Survey of user needs". Working paper, DTI, May 1992. NERC-WP2-34.

[120] M. Panhuijsen, J. van der Voort-van der Kleij, and P. Wagenaar. "Automatic lemmatization experiment". Working paper, INL Leiden, June 1992. NERC-WP9-76.

[121] Paris. "FRANTEXT". Working paper, CNRS Paris, February 1993. NERC-WP5-178.

[122] J. Payne. "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, December 1992. NERC-WP8/WP4-122.

[123] J. Payne. "Speaking the same language? - Listening to the speech community". Working paper, COBUILD Birmingham, December 1992. NERC-WP4-132.

[124] E. Picchi. "DBT: a textual data base system". *Linguistica Computazionale*, 7(2):177–205, 1991. NERC-WP5-11.

[125] E. Picchi. "DBT: Data Base Testuale". Working paper, ILC Pisa, 1991. NERC-WP5-9.

[126] V. Pirrelli. "Report on knowledge extraction". Technical report, ILC Pisa, February 1993. NERC-WP10-79.

[127] Pisa. "Transcription of 3 minutes of spoken Italian (according to NERC Recommendation)". Working paper, ILC Pisa, December 1992. NERC-WP4-161.

[128] N. Poujol. "Questionnaire juridique". Working paper, CNRS Paris, December 1992. NERC-WP11-140.

[129] G. Psathas and T. Anderson. "The 'practice' of transcription in conversation analysis". Working paper, INL Leiden, June 1992. NERC-WP4-163.

[130] E. Putter and J.G. Kruyt. "Evaluation TEI-Guidelines draft version 1.1". Working paper, INL Leiden, June 1992. NERC-WP3-91.

[131] A. Renouf. "Collaboration between CELEX and Birmingham University". Working paper, University of Birmingham, August 1992. NERC-WP2-33.

[132] A. Renouf. "Profile of users and uses of the Birmingham corpus". Working paper, University of Birmingham, September 1992. NERC-WP2-30.

[133] A. Renouf. "The SHAPE corpus". Working paper, University of Birmingham, August 1992. NERC-WP2-29.

[134] H. Rettig. "Comments on Corpus Questionnaire". Working paper, IDS Mannheim, February 1992. NERC-WP1-110.

[135] H. Rettig. "Evaluative report on the Corpus Survey". Technical report, IDS Mannheim, December 1992. NERC-WP1-136.

[136] N. Ruimy. "The argument structure in EUROTRA general principles and applications". Working paper, ILC Pisa, December 1992. NERC-WP8-65.

137] A. Saba, D. Ratti, M.N. Catarsi, and G. Cappelli. "MORFSIN". Working paper, ILC Pisa, 1991. NERC-WP9-22.

138] S. Scheiter. "German spoken language corpora and their text representation schemes - an overview". Working paper, IDS Mannheim, August 1992. NERC-WP4-43.

139] S. Scheiter. "Morphosyntactic annotation schemes in German language corpora". Working paper, IDS Mannheim, December 1992. NERC-WP8-124.

140] S. Scheiter. "Text representation and annotation schemes in German language corpora". Technical report, IDS Mannheim, December 1992. NERC-WP8-135.

141] J. Sinclair. "Lexicographer Needs". Working paper, University of Birmingham, 1991. NERC-WP2-6.

142] J. Sinclair. "The automatic analysis of corpora". Working paper, University of Birmingham, August 1991. NERC-WP9-19.

143] J. Sinclair. "Coherence in Text". Working paper, University of Birmingham, December 1992. NERC-WP8-133.

144] J. Sinclair. "Estimate for the provision of parsed text in English". Working paper, University of Birmingham, August 1992. NERC-WP8-71.

145] J. Sinclair. "Spoken language encoding - evaluation for English". Working paper, University of Birmingham, October 1992. NERC-WP4-46.

146] J. Sinclair. "The Analysis of Topic". Working paper, University of Birmingham, December 1992. NERC-WP8-134.

47] J. Sinclair. "Annotation beyond the syntactic". Technical report, University of Birmingham, February 1993. NERC-WP8-174.

48] J. Sinclair. "Trust the text". In M. Davies and L.J. Ravelli, editors, *Advances in Systemic Linguistics: Recent Theory and Practice*, London, forthcoming. Frances Pinter. NERC-WP8-16.

[149] S. Smith. "Peter French's Transcription Conventions and Sample Text". Working paper, University of Birmingham, September 1992. NERC-WP4-39.

[150] A. Spanu. "Semantic annotation in text corpora". Working paper, ILC Pisa, September 1992. NERC-WP8-66.

[151] F. Surdel. "Consignes de saisie des textes de français moderne". Working paper, INALF Paris, March 1990. NERC-WP3-7.

[152] TEI. "TEI AI 1W2 List of common morphological features for inclusion in TEI starter set of grammatical-annotation tags". Working paper, TEI, June 1991. NERC-WP8-14.

[153] W. Teubert. "Report 29th annual meeting of the ACL Conference". Working paper, IDS Mannheim, June 1991. NERC-WP2-3.

[154] W. Teubert. "Interviews with representative corpus & corpus technology suppliers and (potential) users". Interim technical report, IDS Mannheim, September 1992. NERC-WP2-27.

[155] W. Teubert. "Report on trip to the USA". Working paper, IDS Mannheim, December 1992. NERC-WP1/WP2-123.

[156] W. Teubert. "Interviews with Representative Corpus and Corpus Technology Experts, Suppliers and Users". Working paper, IDS Mannheim, February 1993. NERC-WP2-169.

[157] W. Teubert. "More Interviews with Corpus Experts". Working paper, IDS Mannheim, February 1993. NERC-WP2-146.

[158] W. Teubert. "Phonetic/Phonemic and Prosodic Annotation". Technical report, IDS Mannheim, February 1993. NERC-WP8-171.

[159] W. Teubert. "Workpackage 2: User Needs". Technical report, IDS Mannheim, February 1993. NERC-WP2-170.

[160] W. Teubert. "Worpackage 1: Survey of Textual Data". Technical report, IDS Mannheim, February 1993. NERC-WP1-172.

[161] E. Tognini-Bonelli. "All I'm saying is ...: The correlation of form and function in pseudo-cleft sentences". 1991. NERC-WP8-18.

[162] E. Tognini-Bonelli. "Uses and functions of the adjective "real"". Working paper, University of Birmingham, December 1992. NERC-WP8-148.

[163] P. van der Kamp. "Proposals for standard to access corpora". Working paper, INL Leiden, October 1992. NERC-WP5-85.

[164] P. van der Kamp, A. Boom, and J. van der Voort-van der Kleij. "Overview of and experiences with hard- and software used at the Institute for Dutch Lexicology". Working paper, INL Leiden, September 1992. NERC-WP5-51.

[165] W. Vercouteren and M. Meijer. "Facts and figures on data collection". Working paper, INL Leiden, October 1992. NERC-WP6-86.

[166] W. Vercouteren, M. Meijer, and J. Grinwis. "Supply and demand on the linguistic market". Working paper, INL Leiden, January 1992. NERC-WP6-52.

[167] J.A. Villena-Ponsoda. "Representational Procedures and Schemes for Spanish oral Corpus of University of Malaga". Working paper, University of Malaga, December 1992. NERC-WP4-141.

[168] D.J.G. Visser and D.W.F. Verkade. "Corpora - Questionnnaire Juridique". Working paper, University of Leiden, February 1993. NERC-WP11-173.