Consiglio Nazionale delle Ricerche
ISTITUTO DI LINGUISTICA COMPUTAZIONALE

NERC

SERIAL No 82

Network of European Reference Corpora

TITLE	**Workshop on Textual Corpora (January 1992)**
SOURCE	**NERC Consortium** **Pisa**

Work Package	
Annex No	

Produced as part of the Report
of EC Project: "Network of European Reference Corpora"

Pisa 24-26 January 1992

WORKSHOP ON TEXTUAL CORPORA

Report

J. 2.2.80

A. EXECUTIVE SUMMARY

The workshop was organised to launch, inform and alert a consortium of European institutions who had agreed to prepare a feasibility study on the co-ordinated provision of text corpora in European languages (The network of European Reference Corpora, or NERC). The European Commission, which is helping to fund the workshop and the NERC consortium, recognises the need for substantial corpora to be made available in Europe; the plurilingual nature of the European Community gives linguistic issues a unique importance both socially and economically. The questions posed to workshop and consortium concern the implementation of a network of corpora in Europe.

A major feature of the workshop was the substantial participation of scholars and sponsors from USA. This was a deliberate act of planning, for two reasons. Firstly, there has been a dynamic upsurge of activity in USA in the field of gathering and processing long texts;
the strategies and styles of research are markedly different from European practice, and so it was felt to be important that the benefits and drawbacks of each approach should be brought out and reconciled wherever possible.

The other reason for inviting policy makers from USA was the wish to explore the possibility to co-ordinate sponsorship, avoid duplication, and ensure efficient exchange of information across the Atlantic. The European participants were made aware of the abundance of work in the English language, the modest amount of French and the initial absence of work in other languages reported from USA.

The workshop unfortunately had to be postponed for one year, largely because of restrictions on Americans travelling to Europe in the first half of 1991. This allowed a certain historical perspective to enter into the proceedings, without distorting the original aim of the Workshop.
The list of the invited participants, the programme and even the timetable were not materially altered, and several of the papers required very little attention to bring them up to date.
The design of the workshop was that at least two experts in each field were asked to prepare position papers in advance. Each expert was allotted some twenty minutes in which to introduce the paper, and then there were various patterns of discussion. There was time, and the whole group was small enough, for fairly thorough discussion.
A total of 21 papers were distributed in advance.
The documentation was very efficient; all transparencies were copied and circulated. Rapporteurs at each session have summarised the presentations and discussions.

A.1. OPENING SESSION

The Mayor of Pisa, the President of the University, the Representative of the Italian National Research Council (CNR), and the Representative of the CEC-DGXIII-B, J. Soler, welcomed the participants.

In particular, J. Soler stressed the interest of the CEC for linguistic reusable resources, in the framework of the activities promoted, in various programmes, for the development of a European Linguistic Research and of Engineering and Language Industries, The availability of adequate corpora is recognized as a key issue, and the results of the Workshop, in the framework of the NERC project, are eagerly awaited.

A.2 SURVEY OF AMERICAN AND EUROPEAN PROJECTS.

M. Liberman described the general framework of the activities centred in North America, sponsored by DARPA and NSF, with particular regard to the DARPA spoken language Research, the ACL Data Collection Initiative, the Text Retrieval Conference, the Linguistic Data Consortium.

A. Zampolli, after a brief historical excursus, pointed out the current interest, in various European organizations (CEC, ESF, Council of Europe, EUREKA), in the problem of linguistic reusable resources. After a brief description of some major projects (ET7, MULTILEX, GENELEX, EUROLANG, ET-10-1/5, ACQUILEX, NERC, TEI, etc.) he pointed out the need to define a clear policy for the definition, creation, access and distribution of large reusable resources for NLP and speech, at the technical, organisational, juridical level.

R. Cencioni outlined the overall framework of the activities promoted by the EEC in the field, by sketching both the main lines of development and general organizational aspects.
In particular, he stressed the steps taken in the framework of the Linguistic and Research Engineering program and what is in store for the near future, with particular reference to those aspects which were felt relevant to the arguments touched on in the workshop. He confirmed the interest of the Commision in the issue of standardization, in particular within the field of linguistic resources, and agreed on the urgent need of reaching a consensus on functional specifications. He finally illustrated the procedural steps and a general organizational and budgetary framework within which prospective activities in the field could be funded.

A.3 USER NEEDS.

The needs of the Speech Community are substantially different from others. Some movement was made towards the provision of text reference material common to all researchers.

The needs of lexicographers and other language researchers are clear and detailed. However, language reference materials, such as dictionaries, grammars and thesauruses are not seen as end products of the language industries, but only as intermediate products.

The needs of the NLP (natural language processing) community were very varied and it was not possible to reach a clear position. However there is clearly a much greater priority placed on the provision of annotated text than that found among other language researchers, who often prefer to do their own annotations.

A.4 CRITERIA FOR CORPUS COMPOSITION

One body of opinion stressed the need for such concepts as balance and representativeness to be central to corpus design, and strategies for implementing such designs were discussed at length. Another body of opinion stressed the importance at the present time of making very large quantities of text material available to the public, without the delays and limitations caused by considerations of balance etc. Later developments suggested that most participants would accept the need for three different types of text provision, each with its own relevance to research.

(a) large and not-so-large sample corpora, carefully constructed; including both general corpora and specialised ones dealing with sub-languages and specialised varieties.

(b) Quickly collected, copyright free collections of texts to be made available cheaply and without fuss.

(c) Monitor corpora, combining some features of each of the others - a flow of language, controlled and balanced, passing continuously through computers, which retain a variety of linguistic evidence. The text material is not retained indefinitely.

A.5 TEXT REPRESENTATION/ENCODING

The question of standardisation of text reference material was central to the interests of the workshop. Several participants were heavily involved with the Text Encoding Initiative (TEI), a joint European and American project aimed at producing Guidelines and recommending detailed standards for representing language in text in computers.

There was general agreement that some level of standardisation was essential; universal approval of SGML (Standard Generalised Mark-up Language) as the appropriate set of conventions to be used, and a lot of detailed discussion. A substantial introduction to the special problems of representing spoken language in machine readable form was presented, and this was seen as a welcome development. The adoption of TEI Guidelines was judged to be highly desirable, and an interaction with the NERC partners for the evaluation of the part concerning corpora was recommended.

A.6 LINGUISTIC ANNOTATION

Some types of research using corpora are enhanced by the provision of annotations; usually morphosyntactic but with glimpses of semantic and discoursal annotations as well. There is a substantial tradition of tagging and parsing on small corpora, and more recent software which goes at speed through very large corpora.

A number of experienced researchers pooled their advice and drew attention to current problems. Chief of these is the heavy participation required by trained personnel, in order to bring machine annotations up to a standard dictated by the linguist's intuitions. It was also noted that there is a large range of variation in the task definitions - for example the number of word class tags can vary between c.30 and c.180, annotating the same text.

Attention was drawn to the frequency of structures such as names and addresses, which are not usually featured in linguistic analysis, but may be very important in certain applications.

A.7 METHODS AND TOOLS.

Another approach to the question of annotation and analysis is the provision of software tools which can be used when necessary on a corpus. As corpora get larger and demands on software get more sophisticated, it will become more and more difficult to provide ready-annotated text. However, the raw results of currently available machine analysis are not felt to be in a suitable state for distribution.

Analytic software capable of around 97% accuracy in applying a broad syntactic parse strategy was reported in the analysis of newswire data, giving hopes of general application. As well as conventional parsing, a wide range of lexical and syntactic tools was envisaged as being required shortly. The evidence gleaned from the study of corpora will be fed into the design of future annotation tools. Opinions differed on how fundamental would be the changes induced by careful attention to corpus analysis.

A.8 KNOWLEDGE ACQUISITION FROM CORPORA.

The statistical technique of mutual information is a powerful tool for uncovering certain types of regularity in corpora. Other statistical measures are appropriate to other types of regularities, and there is a wide range of tools available. Many have been well-used already in corpus linguistics. A sensible integration of data-driven and hypothesis-driven exploration of corpora can produce powerful insights. Subject matter can be processed, through proper names etc., as well as language.

Machine-readable dictionaries are a specialised form of corpus with particular strengths and weaknesses. The weaknesses can be in part compensated for by appeal to textual corpora. Different levels of prior annotation are appropriate to different knowledge acquisition goals. It was conceded that statistics is not the only route to the acquisition of "knowledge" from corpora, and it was pointed out that statistics as a means of presenting data is distinct from theories and models of the data. In many cases the statistical tools provided a first-stage sorting of evidence for the analyst to refine.

A.9 SUMMING UP AND RECOMMENDATIONS

Several important policy statements were made in the course of the Workshop. Participants were aware that they were exchanging research experience as part of a process which could lead to important decisions about future funding, especially in Europe. The likely development of EC support for this work was set out in some detail.

Participants were requested to make recommendations individually or collectively, and a considerable number were received. One collective recommendation proposed a tentative strategy for the development of corpora in Europe.

Although much of the detail was questioned, the suggestion of a stepwise approach was widely commended and passed to the consortium for implementation. A possible strategy can be outlined as follows:

Stage 0 The immediate collection and distribution of sample corpora in as many European languages as possible.
Stage 1 A one year project to improve and extend the provision of sample corpora, building on available material and achieving whatever standardisation of formats and comparability of selection is possible. No effort would be put into annotation.

There was some difference of opinion as to the balance of urgency against tidiness in Stages 0 and 1, and some participants suggested to merge them.

Stage 2. A project of perhaps 2+ years, following the NERC consortium recommendations to the European Commission. This would be a set of carefully designed, and planned corpora with good comparability and a sophisticated range of analytic software. As far as can be anticipated, the needs of all community users would be served, and the role of corpus provider would be defined.

Stage 3. It is anticipated that monitor corpora (see Corpus Design above) will develop, preserving the notions of selection and balance, and following the NERC technical and organizational recommendations. Large, relatively unstructured archives will also grow up. Specialised applications will derive from specialised corpora, which will be archived. The need for resources to promote this work is still waiting for further clarification.

B INDIVIDUAL SESSIONS

B.1 USER NEEDS

Chairman: Nicholas Ostler
Rapporteur: Wolfgang Teubert

Introductory Remarks by Nicholas Ostler

Within the field of corpus linguistics, the on-going discussion in Great Britain shows that major users of textual corpora (written or speech) are:

- the speech and the NLP (natural language processing) community
- lexicographers
- MT (machine translation) community
- language teachers/learners.

There are three targets for the construction of interfaces, answering the needs of those who build, annotate or actually use corpora. Standardization is a top priority to all three targets.

In the future, corpus evaluation under quantitative and statistical aspects will become more prominent. Also, more emphasis will be placed on non-linguistic or para-linguistic context features. Corpora will be understood within the framework of multimedia systems. But then corpus processing is too important to be left to linguistics alone.

Among the important issues of information processing, the following refer to corpus technology:

a) How can parallel corpora be correlated?
b) How can toy systems be extended into dependable robust systems?
c) How can statistics complement theory based systems?

R. K. MOORE

User Needs in Speech Research

Users of speech corpora are:
a) phoneticians, linguistics, psychologists (speech science)
b) research community, product developers, system integrators, manufacturers, customers (speech technology)

As to the available recorded speech material, there are different types (general purpose, analytic-diagnostic, task-specific), different styles (scripted/unscripted, monologue/dialogue, human/computer) and different contexts (psychological effects, physiological effects, noise environment, vibration & g, multi-modality, multi-tasking). The labelling of acoustic data consists of transcription (e.g. phonetic, orthographic, semantic, prosodic) and annotation (where there is a time alignment between data

and transcription). Annotated data specify explicit relationships and provide access, which in turn facilitates study, automatic model parameter estimation and assessment.

Initiatives to be mentioned in this area are:

a) national
- UK Speech Technology Assessment Group (STAG)
- IEEE Working Group
- French GRECO
- US DARPA

b) international
- NATO AC 243-Panel III-RSG 10
- ICASSP
- ESPRIT Speech Assessment Methods (SAM)

c) standardization organisations
- NPL (UK)
- NIST (US)
- AFNOR (France)
- BSi (UK)

d) workshops
- Noorwijkerhout, Netherlands (September 1989)
- Kobe, Japan (November 1990)
- Chiavari, Italy (September 1991)

User needs in speech research are:
- more annotated data
- richer annotation (levels of transcription)
- information about standard mark-up conventions and dictionary formats
- conventions to cover all talker generated sounds
- extensions to cover simultaneous acoustic/non-acoustic events
- conformance to agreed standards/formats
- consideration of the use of standards/formats
- consideration of the use of standard DBMSs
- annotations not embedded in the data.

DISCUSSION

H. Thompson: A very important aspect in data representation is the question of digitisation (to be seen in connection with the alignment issue).

Don Walker: The TEI provides a general mechanism to be used for text and speech corpora.

A. Zampolli: It has to be admitted that until now there is very little material of fully annotated speech. On the other hand, there are many transcripts not directly

aligned or connected with speech data.

R. K. Moore: Such corpora are very useful, too, e.g. for the construction of parsers and stochastic grammars.

J. MCNAUGHT

User Needs for Textual Corpora in Natural Language Processing

Textual corpora are of strategical importance in NLP. Today there are very few NLP practitioners left who would deny they need access to corpora. At the recent group of experts meeting in Luxemburg (December 91, organized by the CEC) there was a consensus for a growing dependence on large corpora.
This includes speech corpora as well, as there is a growing integration between NLP and speech research.

Reasons for the use of corpora in NLP: textual corpora provide good lexical coverage; they can be employed for testing different analysis models (rule-based, self-organising stochastic, hybrid); they can be used for building improved models.

The goals of NLP practitioners are wider than those of theoretical linguists; they are interested in unusual constructions including deviant language and in particular sublanguages.
Current NLP systems often perform badly because they are not based on tests with large corpora. Realistic requirements include:

- human support in corpus processing
- skeletal analysis often sufficient for some purposes
- tools for automatic skeletal analysis
- tools based on statistics and probability
- concentration on sublanguages
- exploitation of increase in authoring aids.

DISCUSSION

H. Thompson: It is still open to discussion if text really should be separated from annotation.

A. Zampolli: It is very important to define the concept of sublanguage more precisely. There are a number of different approaches.

H. Thompson: The TEAM-project at SRI should be viewed in this context.

J. M. SINCLAIR

Lexicographers' Needs

Experience has shown that any finite corpus, even of a size of 100 million words, will not be sufficient to fulfill the lexicographer's needs if the goal is to write a general purpose dictionary. Neology is an important aspect of vocabulary, implying not only new words, but also new compounds and new meanings for existing words. A new dynamic concept is needed for a corpus that is open-ended in size, thus reflecting the open-ended flow of language. There are two areas of special needs felt by lexicographers: a) safe parameters to define sublanguages by field, style etc.; b) design criteria for a well organized corpus. On the other hand, annotation should be free from unchecked linguistic hypotheses. Tools therefore should provide comparable results regardless of theoretical predilections.

DISCUSSION

A. Zampolli: It is open to discussion whether such a corpus conception could also account for the needs of the NLP community. Certainly it would not be feasible to fuse the Italian academic dictionary tradition with NLP requirements.

H. Thompson: The concept of a 'dynamic mode' would have to be clarified. Reliability in corpus design means there have to be certain proportions which have to be maintained.

W. Martin: As R. Moore has pointed out, the speech community is not interested in vast amounts of annotated data (other than orthographically transcribed). Apparently J. Sinclair agrees with this judgement also for textual corpora.

S. Warwick: Are parallel corpora of any use for lexicographers? Are they really reliable? The notion of 'comparable' is more useful for plurilingual corpora. Susan Atkins does not see much use in parallel corpora.

N. Ostler: Sometimes translated texts can be regarded as very reliable, e.g. the Bible.

K. Church: While there might not be theoretical consensus on parallel texts, from a pragmatic point of view they prove to be quite useful for a number of applications.

R. Moore: For annotation, whatever can be done automatically, should not be done by hand.

D. Walker: A full bibliographic reference is essential for any corpus text so that the user himself can judge if he wants to use it.

A. Zampolli: Certainly there are many useful applications for parallel texts.

H. S. THOMPSON

Unscripted Spoken Corpora: Resources for Real Language Systems.

To create real language systems, i.e. systems that are fit for casual users, there has

to be a revolution in (theoretical) linguistics, in many respects similar to what has changed the methodological approach to phonetics over the last 20 years. One has to move away from idealized data analysed by the introspective competence of an ideal linguist towards real life language which is more often than not deviant. Rule-based grammars are not able to cope with this kind of natural language unless it is strongly complemented by stochastic models. Large speech corpora are the source of information and the testbed for such models. They have to be available in orthographic transcriptions. Even with those developments in mind there will not be speech recognition system that really can interact with humans for a long time.

Whether linguistic annotations of corpora ever will be reusable is open to debate. This scepticism at least applies to all theory-based aspects of annotation. For all NLP systems, large corpora are needed. Collecting such corpora is very expensive, but institutions like the CEC will have to be convinced that this kind of language resources has to be provided as public domain. To ensure data interchange, standardization in necessary on all relevant levels (cf. TEI).

DISCUSSION

K. Church: Unfortunately until now industry does not know about the advantages of working with a corpus. There is a high potential for using corpora outside of lexicography and speech recognition, e.g. scanning for archival purporses; tools to navigate in machine readable libraries; information retrieval from large knowledge bases etc.

M. Liberman: This search for potential corpus users could easily turn out to be fruitless futurology.

J. Clear: At the British National Corpus, we find it extremely difficult to know what users really do require. E.g., are users really interested in marking up paragraph breaks?

D. Walker: Information seeking behaviour has to be analyzed. Susan Hockey has started research in this area, and it is worth looking into library science that has been dealing with this topic.

M. Marcus: There is an insatiable need for data; cf. DARPA's dogma: 'There is no data like more data'.

S. Warwick: Large corpora are essential for the development of tools within the area of machine translation, e.g. real bilingual dictionaries.

A. Zampolli: Perhaps the term 'user requirements'is more precise than 'user needs'. Important concepts are strategy, philosophy and criteria.

R. Cencioni: It is very important to present the corpus issue to politicians not as a means but as an end in itself. Since private companies are not willing to invest in this area, public funding is necessary. The CEC needs pragmatic short term goals like full

text data bases.

J. Sinclair: Quantity and quality have to be assessed by the targets of investigation. There are problems that best can be dealt with by small high-quality corpora. Collections of a type called for by H. Thompson have been available for many years.

D. Walker: New technology will revolutionize the kind of information that one will be able to extract from corpora. There will be many new users apart from the classic users.

N. Ostler (summary): A new aspect that this session has shown is the need to correlate levels of representation, particularly the digitized speech record with the transcript. Another point is that there seems to be no obvious way from special sublanguage corpora to a general language corpus. One major problem that also needs to be addressed is the copyright question. Finally, the discussion pointed at the dichotomy between self-organizing techniques in resepct to annotation that allow us to do things now vs. theory-based techniques that presuppose a highly idealized and puristic view on language and will not yield satisfactory results for a long time.

B.2 CRITERIA FOR CORPUS COMPOSITION

Chairman: J. Sinclair
Rapporteur: T. Kruyt

D. BIBER

Representativeness in corpus design

Biber's main purpose is to argue for representative, stratified corpora. A representative corpus is a prerequisite for results being generalized to a larger population. The trustworthiness of generalizations (external validity) can be threatened by bias error (i.e. the sample is systematically different from the target population) and by random error (i.e. the sample is too small to accurately estimate the true population parameters). Random error can be minimized by increasing the sample size, bias error by a representative corpus only. Empirical research oncorpora and lexical data of various languages has proved that different texts varieties are markedly different in their linguistic characteristics. As languages are not monolithic, homogeneous wholes, external validity can only be reached by a representative corpus, not by a very large sample of a particular text variety.

DISCUSSION:

Church argues, on the basis of the AP Newswire, that language is changing very fast. Therefore, stable sampling in language is impossible and empirical results cannot be used to obtain a balanced corpus. Large homogeneous samples are preferred. Biber does not agree. The problem put by Martin about how to put in practice theoretically motivated selction criteria is not solved.

J. CLEAR

Corpus design.

Aspects of corpus building are discussed from the lexicographer's persepctive. A corpus is defined as a subset of an electronic text library (ETL) built according to explicit design criteria for a specific purpose. Rather than a once-only activity, corpus building should be an iterative process of collection of selceted texts on one hand, and corpus analysis, feed-back and refinement on the other. The copyright problem may be relevant not only to lexicographers but to other corpus users as well. A corpus for lexicographical purposes requires stratified sampling. The population to be sampled needs to be defined in terms of reception and production. Text selection should be based primarly on external evidence (context, function and use of texts) and secondarily on internal evidence (linguistic features). A preliminary proposal for corpus typology and text attributes is presented.

DISCUSSION

Leech argues that we should not denigrate the concept of monolithic language, as

automatic taggers satisfactorily operate on very different corpora. Walker, Clear and Biber support the use of both internal and external evidence. Church argues in favour of very large homogeneous samples rather than a collection of smaller subcorpora, in view of the robustness of parsers. Statistical relevance is required. Clear, on the contrary, argues that robustness is affected by language phenomena wich are usually not covered by a general language sample, but are found in specialized samples. Subcorpora may be large as well. The discussion, with contributions from, among others, Marcus, Zampolli, Thompson and Martin, focuses on the demand for very large homogeneous corpora of sublanguages vs. large scale broad representative corpora, depending on the purpose. Thompson, points to the fact that size is now always feasible. Liberman concludes with four points: 1) different linguistic registers have different properties, 2) broad sampling is desirable, 3) something is much better than nothing, 4) the rest being equal, more is better than less.

Second session

SINCLAIR: summary of the main points of the first session. The first point concerns the validity of corpora. The best and only solution to random error is more text. With respect to bias error, there are two opinions. One stand is: everything is representative of something, which is favoured by people more concerned with techniques, developments, tools etc. The other is: what really matters is linguistic validity , favoured by people more concerned with results and descriptions of patterns of language. A second point concerns balance. Given the need for balance, the question is how to relate internal and external criteria. Another question is whether sublanguages or subcorpora have independent validity. The third point concerns sponsors: who is paying for corpora. Corpora do not sell to the end user. In order to convince sponsors of the usefulness of corpora, targets may be: reporting reliability on a language, coping with new unknown material, and operating cost-effectively.

SUMMERS
Longman Lancaster English Language Corpus: Criteria and Design.

The Longman Lancaster English LAnguage Corpus (30 million written words) and the British National Corpus (100 million words of which 5-10 million are spoken) are described. Common design features: 1) selective and microcosmic halves, 2) topic (subject field) basis, not genre basis, 3) a distinction can be drawn between two types of features: a) selection features determine target percentages of informative/imaginative domain/subject field, medium, level and time; b) classification features are features that are recorded but do not contribute to determining target percentages. The BNC adopts a two-part approach to spoken corpus building: demographic and context-governed. Both corpora are motiviated by lexicographical needs.

Gibbon claims that the needs of the speech and language community cannot be separated. At present there is a data-bottleneck. A possible solution may consist in deriving subcorpora from corpora; furthermore the possibility can be envisaged of substantially enlarging small corpora. An overview of speech projects (SAM, EUROM, Pre-scribe) is presented. Speech and language communities need to meet

and to join forces in order to obtain mutlilingual flexible data.

FURTHER DISCUSSION:
Church proposes two alternatives to the balanced position: introduce weighting functions, and analyse samples separately. Internal and external arguments determine whether or not two samples are equal.
Thompson argues that the targets mentioned by Sinclair are not universal but specifically lexicographical. Disposition raises a discussion concerning whether the set of items is different between lexicographers and other corpus users. Among others, Thompson, Sinclair, Clear, Calzolari, Martin, Warwick, Biber, Marcus, Walker and Leech contribute to the discussion. Clear argues that the NLP community needs lexicographers (cf. use of LDOCE in NLP). Calzolari points out the methodological risk in laying too much stress on the dichotomy between lexicographers and linguists. Corpora are needed for better NLP-lexicons. The lexicographical approach should be incorporated in the NLP approach. There is a need for a new profile the linguist/lexicographer, in order to build sound NLP-lexicons. With resepct to the dichotomy between balanced and unbalanced, it is argued that an unbalanced subdomain approach is only a partial view. A flexible baanced corpus design enables users to make specific subdomain selections. However, sample size may be a problem (Warwick).

Liberman asks for which languages corpora are available for research. The results of a survey by Zampolli and Walker show that there are corpora of many European languages. They have a different status of availability. Moreover availability is complicated by different schemata. Liberman and Warwick argue in favour of a better access to European corpora.
At the end of the discussion, some consensus is reached: both unbalanced narrow scale and balanced broad scale corpora are needed. A general policy for corpus design has to be maximum utility. Zampolli asks for recommendations, including scientific and organisational aspects.

Chairman's report:
For some users, a corpus should attempt to be a reliable record of a state of a language, general or restricted. This means that its constitution should avoid both bias error and random error.
For other users, the only control needed is for random error. It is contended that a lot of activity can be simulated by quantities of text regardless of source.
For random error only, the quantity of material that can be acquired nowadays is thought sufficient in principle. For bias error, the difficulty of accumulating some types of data, e.g. conversation, should not be overlooked.
Some applications that are envisaged are too specialised for a corpus to be centrally provided in advance of need. Corpus providers should be explicit about criteria of corpus design, so that ad hoc corpora can be assembled.
For corpora which control for bias error, it is accepted that we do not yet know the critical parameters and proportions, nor the best sampling technique. A cyclical process of parameters and proportions over time is envisaged, progessively aligning external and internal evidence.

B.3 TEXT REPRESENTATION AND ENCODING

Chairman: Don Walker
Rapporteur: Danielle Candel.

M. SPERBERG MC-QUEEN

TEI and Text Representation.

The SGML standard, officially adopted by the US Government, has been recently accepted by the EEC as well. The Japanese are also highly interested in SGML and TEI. TEI can therefore be said to be truly international. The possibility of extending TEI recommendations to other fields (e.g., visual input, types of images and the like) is currently under evaluation.
All sorts of languages are taken into account. The Spanish version of the SGML manual is ready, the Japanese one is currently being translated. The first major requirements that came out of a recent user needs survey concerns the need of a so called "intralanguage", meaning a common, reusable notational convention for defining relevant (para)linguistic units in texts.
There is a large variety of opinions as to what should be annotated in a text and what should be left out. Some scholars are in favour of a skeleton annotation only, some others would like a much more detailed and rich annotation. Annotated material can be scaled up or down. What is important to stress at this juncture is SGML flexibility in handling all kind of mark-ups, notational conventions and related problems.

SGML syntax represents the text hierarchically, as nested elements specifiable in terms of attribute/value pairs. Multiple hierarchies are also possible. SGML has an open ended, extensible schema. Among the nice features of SGML there is the fact that SGML owns a fairly closed set of special characters (start-tags, end-tags, tag delimitators, delimitators of entity references). One does not need to use one's own software, one can use existing SGML software.

L. BURNARD

The Text Encoding Initiative: a progress report.

TEI goals are: make data exchanges easier, and provide guidelines to text encoders.
Guidelines concern: what to encode and how to do it.
Desiderata are: wide acceptance, simplicity, comprehensiveness, conformity to international standards and software, hardware and application independence.
Topics are:
1 - dealing with the text stream. Only a subset of ISO-646 character set can currently be relied on for data exchanges. By using entity references or transliteration schemes SGML allows for suitable extensions of a shareable character set. This is achieved thanks to formal Writing System Declarations. In this respect the TEI has set itself the promising goal of designing a machine independent support for all writing systems in all languages)

2 - text documentation: a special header contains information about file, source, tagging schemes for notes, names, cross references, hypertextual links, "crystals", etc.
3 - multiple parallel intepretation
4 - representing all kinds of linguistic annotation (e.g. part of speech tagging)
5 - tailoring TEI to user needs

The second draft of the TEI will provide:
- tutorial guides (theory and practice of markup)
- guidelines
- a case book of extended examples of TEI.

G. SAMPSON wonders about the possible neutrality of linguistic annotation (as to grammatical analysis, there is a contrast between those who believe that word-class tags should be seen as simple atomic symbols and those who believe that they represent clusters of separate logical features).

It is replied that the notational flexibility of SGML allows for any type of grammatical annotation to be encoded appropriately.

S. JOHANSSON

Representing speech in machine-readable form.

The proposals which have been put forward in the field are compatible with TEI's proposals. TEI forces one to be precise and address problems. Detailed information is provided in a recording statement, and in a list of participants, which contains demographic information as well. More effort has been put into transcription rather than acoustic/phonetic representation. There are different types of events going on in spoken situations; some localisations are not lexical, such as hesitation marks and the like. They can be aptly encoded as coordinating elements which point from the text transcription to the recording timeline.

DISCUSSION

D. GIBBON: the datastructures are quite complicated: how is it possible to organize different and possibly mutually incompatible syntactic structures, anaphoric structures, prosodic structures, paralinguistic structures, which are not nested in a tidy tree structure?

M. SPERBERG MC-QUEEN: different annotations are kept separate from the text and put in different places in the file; such places can be pointed to from the text line. This is analogous to what is done in transcriptions where pointers refer to the timeline from the textline.

D. GIBBON: How is the timeline represented?

L.BURNARD: Timeline can be very flexible. A minimum of alignment information can be worded in this way:" this occurs before this". You can have a part of the text

which is very densely annotated and another part which has no synchronization at all. A more complex instance of alignment information is the following: "not only this point follows this point, but it follows by a given time span". You can give the locations relative to any preceding point or to the base.
R. MOORE would like to have tools for relating acoustic signals to the markup (for phonetic annotations they already exist).

Don WALKER

We recognize the need for providing software to facilitate the use of TEI results. We intend to submit the results to ISO. The EC and the US have committed to SGML, and Japan, being interested in SGML, has a strong program in the same direction. It has been planned to consider the relationship to database technology so that the TEI tagging can be applied to database elements on one hand, and the database tachnology may contribute a framework that would enhance the TEI program on the other hand.

B.4 LINGUISTIC ANNOTATION
Chairman: W. Martin
Rapporteur: V. Pirrelli

G. LEECH
Corpus annotation schemes.

The dichotomy between so called "raw" corpora and annotated corpora is more a matter of degree rather than a clear-cut, principled divide. Some of the added linguistic features which have been described as basic in transcribed corpora of spoken language (for example by R. Moore), can be easily said to belong, when looked at from another perspective, to annotated corpora in a broader sense.
slide 1: *Maxims for annotation*
1) raw corpus and annotation should be easy to separate.
2) annotation schema should be automatically available and accompanied by a detailed guide concering criteria for annotation.
3) annotators must be trusted; user should be informed as to who made annotation.
4) "caveat emptor" principle; annotation must be under quality control, and users should be told about degrees of reliability.
5) as far as possible, we should make use of **consensual** categories. Dictionaries are, in this respect, being informative to all specialists and non specialists as well, can be taken as a model of repository of well-known, widely understood and relatively uncontroversial categories.
6) standardization is something to be hoped for. That a scientific committee could eventually reach a unanimous conclusion on a fixed nomenclature, is hardly likely. Rather, it is hoped that some sort of de facto standard will emerge in the near future.

DISCUSSION on slide 1.

Textual material made available through CD ROM in the framework of an ACL-DCI sponsored initiative, reflects by and large Leech's recommendations. However, annotation criteria are not automatically available, and records are not kept as to who did what. (Marcus)
Appeal to ordinary book dictionaries as sources for standardization is felt questionable by some scholars, since there does not appear to be the wide classificatory agreement which is hoped for. (Church)
An error in tagging is just what does not comply with standards as they have been laid down according to one's annotation criteria. In particular, classes of particularly frequent errors have been spotted, so that notorious erroneous cases can be easily tracked down automatically and consistently checked through human intervention. For most controversial cases, users should be given unresolved ambiguities. (Sinclair)
Manual grammatical tagging is inconsistent. Fully automatic tagging is, so far, impossible. At the moment, post-editing of automatic tagging is the only realistic solution. (Leech)
Researchers must be aware of the circularity of this practice, where both the notion of "correct tagging" and "erroneous tagging" are fraught with a number of a-priori,

intuition-based assumptions. (Sinclair)
slide2: Levels of linguistic annotation: the state of the art.
The following list is not exhaustive but provides a picture of work already done in the field:
1) phonetic
2) part of speech (tagging): fairly well known canonical form of annotation. It is not easy but definitely easier than others.
3) skeletal syntactic parsing, usually in the form of labelled bracketing: there is no reliable automatic strategy, though (see infra).
3) semantic annotation: word sense resolution, mainly for lexicographic purposes. No total sense representation is provided; only sufficient information to distinguish among multiple readings.
4) discoursal annotation: across sentence boundaries are made explicit; anaphoric links are represented to provide a first stage input to NLP systems aimed at pronoun resolution.
The most successful processing methodologies adopted so far can be synthetically referred to as follows:
HAMA: Human aided Machine Annotation, and
MAHA: Machine aided Human Annotation.
Clearly, there is a continuum between the two. There has been much effort in trying to develop fast human interfaces to speed up annotation work; previously encoded levels of annotation are a welcome help in this respect.
Comments on slide 2.
There is a problem of consistency, not only within each level of annotation, but between different levels of linguistic annotation as well (Lafon).
As to the aspect of internal consistency and wide acceptability of annotation schemata at the level of part of speech tagging, two different stands have emerged:

a) consensual schemata are an attainable goal, since there appears to be a fairly large amount of de facto agreement in this respect; differences of opinion come down to different uses; most tag sets can be mapped onto each other (Marcus, Warwick); there is an observational level of linguistic analysis at which most people seem to agree (Martin).

b) large consensus is due to the fact that most scholars do not face real linguistic data as attested in corpora. It would be very unwise to rely on categories. Work on corpora tells us that grammatical generalizations must be grounded on lexical evidence (Sinclair).

slide 3: Desiderata on both sides.
DESIDERATA FOR ANNOTATORS:
1) Accuracy 2) Consistency 3) Speed
DESIDERATA FOR USERS:
1) Useability for applications
2) Variable delicacy of analysis
3) Descriptive categories
DESIDERATA FOR BOTH: linguistic validity

G. SAMPSON

Needed: a Grammatical Stocktaking.

Experience on carefully, richly annotated texts has brought out the fact that there exists a number of large, yet unexplored areas of linguistic phenomena which are still waiting to be explicitly focused on and properly investigated. Moreover, even in more common areas of standard grammatical description, terms are used in different ways by different scholars. Setting categories is not enough, if one does not fix boundaries between categories very carefully. All in all, there seems to be an urgent need for scientific agreement on standardized criteria of linguistic classification as a basis for moving forward. A similar kind of scientific urge led Linnaeus to devising a finely tuned nomenclature assigning one name to one particular biological specimen. To suspend agreement until a full level of understanding of the field were achieved would have hampered progress in biology.
A list of relevant dichotomies is then presented:
1) relatively small & richly annotated corpora vs big corpora with skeletal annotation; insufficient attention has been paid so far to carefully annotated small corpora.
2) we should not presuppose similarity among languages; each language must be investigated in its own terms.
3) we should impose categories rather than discover them; we cannot do both.
4) neglected areas of linguistic descriptions such as names, weights, addresses and the like, must ve grammatically described.
5) biological taxonomy should be superimposed on logical analysis.

DISCUSSION

The crucial problem in annotation practices apparently rests not on the need for imposing standards, but on how to design standards: a) what sort of delicacy you have in mind; b) what discovery procedures you use.(Marcus, Sinclair)
Different languages are much less different than they are thought to be. Tagging schemes can be naturally transferred from one langauge to another. Skeletal parsing is already a broadly shared, consensual area for linguistic coding.(Calzolari)
Linnaeus taxonomy was claimed to be grounded on some principle of structure reduction, which is exactly what is disclaimed here. Linneus seemed to be much more committed to finding the right set of primitives.(Wilks)
Another attempt in this same direction apparently failed: Roget's Tesaurus proves that linguistic units are not observable entities in the way creatures are. (Summers)

E. EJERHED

The Swedish Annotated Corpus.

Main goal: delivery of a balanced annotated corpus of 1M tokens. The BROWN corpus structure has been taken as a model. Collection of a Raw Super Corpus is also under way. Annotation includes part of speech tagging, lemmatisation, inflectional features. An effort has been put in getting at a fairly subtle level of tagging; distinctions have been made which are not familiar to Swedish grammatical tradition. At such a level of accuracy, having achieved a success rate of 77% in automatic tagging is remarkably good.
Tagging output is used as an input to automatic, rule-based syntactic processing:

results are manually checked. The overall enterprise is not devoid of germane scientific relevance.

CONCLUDING DISCUSSION

There is a large range of variation in the task definitions (number of word-class tags can vary between 30 and 180), time and costs (small, richly annotated corpora tend to be more expensive than large corpora with skeletal annotation). Experiences in gathering material and coping with problems of copyright and restrictions imposed by publishing houses have shown a similar broad range of variation. Generally speaking, costs should be weighed up against aspects such as marketability of the final product and pay-off returns. Exchange of information and experiences and coordinated efforts are a key issue in this context.

B.5 METHODS AND TOOLS

Chairman: Y. Wilks
Rapporteur: E. Tognini Bonelli

M. P. MARCUS and B. SANTORINI

Building very large natural corpora: the Penn Treebank

There is a growing consensus that significant, rapid progress can be made in both text understanding and spoken language understanding by investigating those phenomena that occur most centrally in naturally occurring unconstrained material and by attempting to automatically extract information about language from large corpora of natural language. Such data bases are valuable for enterprises as diverse as the automatic construction of statistical models for the grammar of both the written and colloquial spoken language, the development of explicit formal theories of the differing grammars of writing and speech, the investigation of prosodic phenomena in speech and the evaluation of the adequacy of parsing models, the various formal syntactic theories embedded in those parsers, and the particular grammars of English encoded within those theories.

Ultimately, a corpus of at least 100 million words of annotated text is desirable for such research. As a first major step towards this goal, the "Penn Treebank Project" has built a corpus of over 4 million words of unrestricted text and has developed and tested techniques for annotating linguistic structure. Todate, we have annotated the corpus for part-of-speech (POS) information (with over 3 million words freely available, and the remainder available to those with a license for the Brown Corpus), and skeletal syntactic structure (with over 1.2 million words now annotated.)

Our sponsors and the research community we are part of are primarily interested at the moment in using the materials that we provide as the basis for training large statistical models of language which are expected to be of a fairly low degree of delicacy, as compared to many linguistic descriptions. As a result, key concerns in the design of our annotation schemes for both the POS tagging and the syntactic bracketing tasks were that annotators should be able to add information quickly and consistently and tat they should be sure of the information they add, without being subject to the pressure of having to make arbitrary decisions.

The Penn Treebank tagset is based on that of the Brown Corpus. However, the stochastic orientation of the Penn Treebank Project and the resultant concern with sparse data has led us to modify the original Brown Corpus tagset by paring it down considerably rather than extending it. A major consideration in the tagset reduction was recoverability. A prime example concerns the tags for the verbs have, do and be. The Brown Corpus distinguishes five different verb forms for main verbs. The base form is tagged VB, and forms with overt endings are indicated by adding D for past tense, G for present participle/gerund, N for past participle and Z for third person singular. The Penn Treebank tagset eliminates the distinction between HV and VB as

well as their suffixed variants, since the distinction can easily be recovered with reference to the small set of the forms of have if necessary.

Our strategy for POS tagging has been to use PARTS, a highly accurate (3% error) stochastic algorithm developed at AT&T Labs (Church 1988), to provide an initial POS assignment. The output of PARTS, which uses a modified version of the Brown Corpus tagset close to our own, is then automatically modified (4% error) to be in accordance with the Penn Treebank tagset. The learning curve for the POS tagging tasks takes under a month; annotation speeds after a month exceed 3,000 words per hour.

Our methodology for bracketing is parallel to our POS tagging methodology. We use FIDDITCH, a deterministic parser developed by Donald Hindle first at the University of Pennsylvania and now at AT&T Bell Labs, to provide an initial parse of the material (Hindle 1983), (Hindle 1989), and then edit the output of the parser by hand. Annotators do not need to rebracket much of the parser's output - an especially time-consuming editing process. Rather, their main task consist of "gluing" together the syntactic chunks produced by the parser. This is done using a primarily mouse-drive process, by which an annotator moves an unattached chunk of structure under the node to which it should be attached.

After early concerns about productivity, we investigated a range of methods for syntactic annotation with respect to annotator speed, for annotators postediting the output of FIDDITCH. The key results are as follows:

- Annotators take substantially longer to learn the bracketing task than the POS tagging task, with substantial increases in speed occurring even after two months of training.

- Annotators can postedit the full structure provided by FIDDITCH at an average speed of 375 words per hour after three weeks, and 475 words per hour after six weeks.

- Reducing the output from the full structure to a more skeletal representation similar to that used by the Lancaster UCREL Treebank Project increases average speed to approx. 700 words per hour. At this speed, a team of five part-time annotators working three hours a day should maintain an output of approx. 2.5 million words a year of "treebanked" sentence, with each sentence postedited once.

- Performance varies more by annotator in the bracketing task than in the tagging task, with speeds after two months ranging from approx. 400 words per hour to over 1,000 words per hour.

3. QUESTIONS AND DISCUSSION

S. Hockey started the discussion by asking whether the output of parsing in M. Marcus texts was of use in other texts.
M. Marcus commented that this kind of statistical method had reduced the error rate

quite drastically in parsing over the past few years.

W. Martin asked (1) whether M. Marcus had anything to say about lemmatizing, (2) whether the figures could be scaled up, from a methodological point of view, (3) considering that after the automatic analysis there is a residue which has to be hand-corrected, whether the annotators had to go through the full text.

K. Church answered question (1) by saying that he only used morphology "when desperate".

M. Marcus said that there had not been a great deal of interest in lemmatization. Concerning question (2), Marcus answered thatthere is a limit to improving the accuracy: 98% is as far as you can go. Concerning question (3), Marcus said that there are techniques which are being developed to pinpoint likely errors.

G. Samson queried the validity of merging the distinction between auxiliaries and other verbs in the case of contractions.

M. Marcus answered by saying that they did not collapse anything that they could not recover, for example it had been discovered in practice that the word "to" required only one tag.

S. Hockey asked again whether the parsing strategy had been tested against other text types. Some inconclusive discussion followed.

D. Biber referred to the persistent tagging errors on the words male and female and suggested that different text types might show different patterns of error.

J. Clear pointed out that the distinction between male (NN) and male (JJ) is still at present a problem for the Penn Treebank tagger and not a usable contrast.

S. Hockey asked whether M. Marcus was planning to use the parsing system on other materials other than newspapers.

M. Marcus replied that it would be advantageous to train the parser on million-word corpora of a range of text types and hoped that sponsorship would be found, as well as someone to do the job.

M. Liberman said that tagging is cheap and that rather than doing studies of what is involved in analyzing different text types it would be cheaper simply to do them as required.

S. Hockey questioned whether this policy would ultimately be the cheapest.

J. M. SINCLAIR

The Automatic Analysis of Corpora

J. Sinclair started by stating that annotation tools and the process of annotating corpora are very strongly linked and that if one has suitable tools the corpus can be analyzed in real time. The principal argument against this being that it can't be done at present to an acceptable standard.

Concerning the goals of annotation, Sinclair noted the pre-existence of a wide range of standards and goals which did not seem to be adequately stated. He divided goals into interim goals, such as parsing, which could be seen as stepping stones to other goals: real world goals such as understanding language, speech recognition. He said that he did not recognize any open-ended goal. He insisted on the fact that goals should be very clearly stated.

Looking at the results of annotation, again he noted the wide range of results according to different types and methods. He pointed out that one should be very careful when it came to evaluating them: they should not be evaluated in their own terms, but against real world goals.
Sinclair mentioned that in the written version of the paper presented for the workshop he had listed different tools which were not of the type so far exemplified, and that he wanted to offer a rather different stance which required us to change and adapt our methods of description. He said that the study of language had moved from a state of being starved of data to actually being submerged by it. As a consequence the stance of the linguist must change and pre-corpus theory and categories should not be taken for granted. These, he said, had failed with respect to data - and this had been admitted - and therefore should not be nurtured and brought in a pristine condition to bear on material for which they were, in principle, irrelevant.
To ignore this would lead us into making another "serious and expensive error".

Sinclair noted that we had come to accept pre-corpus beliefs (PCB) as given and unquestionable: correctness meant consistency with PCBs; consensus was the set of PCBs. No change had taken place. He advocated a pose of humility and self-criticism in the face of evidence of a type, quality and quantity that our techniques had not been adjusted to and invited linguists to adapt their methods of description accordingly.

Looking at policy, Sinclair noted that there was not a lot of disagreement at the level of strategies: everybody wants speed, robustness, precision, etc. It was rather the stance with respect to data which may create problems. He suggested that, as a matter of policy on the administrative side, the corpus-providers should simply provide the corpus and whatever automatic tools are available. He insisted that the crucial point from a policy point of view was that the tools should be fully automatic, as, in his experience, no two users were likely to want the same type of annotation; furthermore, automatic tools were the only ones to provide consistency and therefore validity to the analysis. "Hand-finishing", he said, was a private matter, and should not be funded as a service with public resources.

From the point of view of policy on the research side Sinclair urged researchers and linguists to examine the language avoiding the pre-conceptions of one's own intuitions as there was ample evidence that intuitions only provide a partial view of language structure. The computer is able to retrieve and deliver the subliminal and unconscious patterns of language, and for this reason it is crucial to have tools of investigation capable of capturing these patterns and prioritising the prominent ones. From this first set of observations, one can provide explanations leading to hypotheses and ultimately to theory. This approach, Sinclair said, was largely "bottom-up", although there was scope for insights which didn't derive directly from the routine. However, he insisted that in particular in the areas of parsing and the larger scale and more subtle types of analysis it was important to restrict oneself to actual language patterns.

Concerning the type of tools that are going to be important in the coming years, Sinclair noted that at the moment we have a wide range of grammatical tools and a growing number of lexical tools such as collocators; however, it was clear that, in order to have a comprehensive type of analysis, these had to come together and operate simultaneously both with grammatical and lexical evidence.

Secondly, he pointed out the importance of modularization of tasks, so that a number of constituent partial parsers, which did some very specific tasks, would operate simultaneously in a UNIX-style type of approach, allowing for flexibility and adjustments.

Thirdly, he insisted on the need for a range of professional tools which would be very much devoted to interim goals, and therefore would be mainly aimed at the computational linguists and at facilitating the use of analytical tools.

Lastly, he mentioned the importance of multi-lingual tools that would operate in a similar way in more than one language, comparable tools that will go with comparable corpora.

3. QUESTIONS AND DISCUSSION

Y. Wilks started the discussion by saying that preconceptions do exist and wondered whether it would be possible to break out of such preconceptions.

J. Sinclair replied by saying that if you set aside the assumption that there are major word-classes like noun and verb, for example, then you do not in the first instance find them in a corpus. If one is not prepared to set aside such assumptions, there is no need to use a corpus.

W. Teubert noted that the corpus is useful for checking and correcting a view of language, but it cannot replace a theory of language.

J. Sinclair agreed and said that since we now had ample evidence of language structure we should not be overinfluenced by previous beliefs.

M. Marcus mentioned firstly that the demand is now for more heavily annotated text

and that high funding levels depended on paying for annotation. Secondly he felt that there was not time to wait and see if 'the right' theory would emerge form corpus study.

J. Sinclair said he hoped the funds would be diverted to support fully automatic annotation tools. He objected to paying for correcting mistakes originating in theories which were not data-sensitive.

N. Calzolari firstly pointed out that Sinclair's position on 'pre-conceptions' was too extreme since these were already present in the choices one made when looking at the data in the corpus. Secondly, she said that although she found both the ideas of completely automatic annotation and of the monitor corpus very appealing, the problem seemed to be that the fully automatic tools that were available performed only the early stages of an adequate linguistic analysis and at the moment, in order to have something sound one had to rely on human intervention.

J. Sinclair agreed and said that he did not mean to say that one should not have a theory at all, but that one should bring out hypotheses as they interacted with the data. He also agreed that at the moment we did not have the ability of merging all the automatic tools available and to make a comprehensive analysis. At the present time in order to produce annotated material one would have to do it partially by hand. However the issue was who should pay for manual finishing. In his opinion that should not be the corpus provider on public funds.

Y. Wilks said that there are many full automatic processes, the only issue was whether one was prepared to live with the error rate that followed.

V. Della Pietra pointed out what he perceived as a contradiction in Sinclair's advocating automatic tools on the one hand, but on the other hand being against any hand-annotated data. In his opinion one needed hand-annotated data in order to produce an automatic annotator. In annotation, he said, one is just trying to match what humans do. Making another point, Della Pietra agreed with Sinclair on the need for goals, but he said that automatic analysis of many different kinds could be co-ordinated to provide information for mechanical translation, with the human being the ultimate decider.

J. Sinclair answered that if all you want to do is replicate what humans do you do not need a corpus for that; what the corpus could do is provide information that is not available to the human. Concerning machine-translation, Sinclair said that, in his opinion, it needed to have a much sounder theory than just the merging of several sets of statistical information.

G. Leech said that one could not ignore the two thousand years of linguistic scholarship. Language was also a mental phenomena, and that in analyzing corpora Sinclair was trying to dispense with the human mental aspect. He said that when we parse sentences by machine we need to make use of the human intelligence which can judge whether we have parsed it correctly according to our intuitions or not.

J. Sinclair replied that he did not refuse the experience of linguistic scholarship; on the contrary he believed in approaching a corpus with a long training in data-oriented linguistics, allowing for one's own knowledge and intuitions. On the other hand it was important to accept that the corpus might have something new to show, patterns which had not been noticed because one's intuition and intelligence are limited by what is consciously available; it was clear from observation that one's actual linguistic behaviour is different from one's understanding of it.

B.6 KNOWLEDGE ACQUISITION FROM CORPORA

Chairman: H. Thompson
Rapporteur: N. Ostler

K. CHURCH

Using Statistics in Lexical Analysis

The session began with an introductory presentation from Ken Church of AT&T, giving a very brief exposition of the Noisy Channel Model of Information Theory, adverting to its applications in recognition, translation, text compression, cryptography, information retrieval and parsing.

V. DELLA PIETRA

Class-based n-gram models of Natural Language

V. Della Pietra of IBM then took up the tale, on Statistical Methods for Finding Word Classes. He began with a glance at IBM's statistical work in machine translation, based on the Canadian Hansard bilingual corpus; but then focussed on the main substance of his talk: a qualitative overview of the strengths and weaknesses of grouping words into classes based on statistical similarity of their context, and the use of these classes in simple probabilistic language models.

He distinguished sticky pairs ("Humpty Dumpty", "jiggery pokery") from semantically sticky clusters ({question, questions, asking, answer},), both identified using mutual information, but on different contextual windows. He then moved to consider various sub-optimal classification methods to generate a set of semantically similar classes that would include all words in a corpus.

Stochastic language models were the next topic. The number of unknown probabilities needed as input to these models could be minimized by using word-classes, and their multual probabilities, as proxies for those of the words themselves. If low perplexity was accepted as a criterion of a good language-model (with respect to a given body of data), it could be shown that the choice of classes which gave the best 2-gram model would be the same as the choice which maximized mutual information between adjacent words.

DISCUSSION

Although the flavour of Della Pietra's talk was mostly an exposition of statistical methods, he agreed with the next speaker, Ken Church, in holding that statistical methods were useful principally as a set of tools for testing linguistic hypotheses, rather than as a impartial mechanism to generate useful models without human prejudice or intervention.

Church proceeded to demonstrate the points in which statistical analysis could be useful in lexical analysis. He felt that statistical techniques could help to overcome the faults of other (quite legitimate) approaches to examining lexical data: intuition gave unstable results; citation indexes over-emphasized the special and untypical; and concordances would not necessarily be organized in a way that made significant co-occurrences salient.

As against proponents of "self-organizing" systems, he suggested as indispensable three points for human judgement in the use of statistics:

- to choose an appropriate statistic (e.g. mutual information, t-score)
- to pre-process the corpus to highlight properties of interest (e.g. tagging or parsing it)
- to select an appropriate unit of text in which to look for generalization (whether bigrams, SVO triples, or whole discourse).

As against opponents of statistical methods altogether, he suggested that they provided an objective methods to focus on the central and typical facts of a language, rather than to attempt characterization of its rules as a whole.

He also showed that certain types of linguistic corpora (e.g. news media) could have features which would never "settle down" whatever the size of corpus examined: certain topics (and the language distictive of them) occurred sporadically, with unpredictable peaks.

N. CALZOLARI

Lexical Knowledge Acquisition from Textual Corpora.

Nicoletta Calzolari of Pisa Institute of Computational Linguistics spoke last, siting her remarks with reference to those Ken Church.
In general she confirmed his English results in her work on Italian, but went on to emphasize the frequent need for balanced corpora, in order to gain linguistically significant results. However, grammatial information, and extraction of compounds were less sensitive to this requirement, the former because it would emerge from any type of text, the latter because it could be best undertaken on the basis of corpora reflecting particular sub-languages.

She also showed how these corpus-based techniques complemented those useful in the processing of machine-readable dictionaries. In particular, corpora enable the analyst partially to overcome the decontextualization of meaning in lexica, sometimes suggesting objective criteria for subtle and salient senses. Together, corpus-based knowledge acquisition and exploitation of the form of machine-readable dictionary entries would provide the foundation on which to develop lexical databases.

DISCUSSION

In the discussion, Geoffrey Leech of Lancaster University asked about the specific advantages of annotated corpora. All the speakers accepted that there were useful

roles for annotated corpora in all their statistical analysis techniques, but that the utility of annotations had to be judged case by case. In practice, annotated corpora were not always available, but significant progress was possible even without them. Jeremy Clear (Oxford UP) pointed out that statistical techniques such as those used at IBM served to extract generalizations rather than knowledge, e.g. rough facts about classes, rather than sharper particular truths about words. This meant that the information became more tractable for statistical manipulation, but at the cost of reliability and accuracy. Generality is paid for in lost granularity.
The chairman (Henry Thompson - Edinburgh University) offered a challenging overview: that statistical results were of interest in the absence of valid theory -- class-based generalizations in default of a grammar, word-based ones in defaullt of a semantics. Since the latter need was actually more strongly felt, word-based generalizations were the more interesting.

Pierre Lafon (CNRS-INaLF) pointed out that the statistical techniques that had been expounded were in fact very old, and had in previous research shown themselves very brittle, producing results which were very sensitive to the particular corpus used, and to the specific annotations that might be imposed on it. Consequently, it was essential that lexicographers should do their corpus-analysis, so as to be competely au fait with the properties of the data.

Roger Moore (Speech Research Unit, Malvern) emphasized the importance of concentrating on the details of the models, rather than the statistical techniques and probability theory themselves. Della Pietra heartily concurred.

B.7 CONCLUDING SESSION and remarks

Chairman: M. Liberman, A. Zampolli
Rapporteur: N. Calzolari

The concluding session was articulated in two parts:
1) reports by the chairmen of each session, highlighting the main issues both of the presentations and of the discussion
2) a series of recommendations to be submitted by the workshop to the NERC Consortium.

B7.1 REPORTS by the sessions' chairmen

INTRODUCTION

D. Walker

There is a critical need for international cooperation that is recognized explicitly in the convening of this workshop. Respectful of European leadership, DARPA and NSF have begun US investment to establish a new infrastructure for research and development.

US activities are exemplified by the DARPA spoken language project, by the ACL/DCI, by the Text Retrieval Conferences (TREC) and the Message Understanding Conferences (MUC), and by the Linguistic Data Consortium. All stressed massive resource acquisition.

The European efforts have a longer explicit history and emphasize the demands of a multilingual community supported by a number of different transnational structures where there has been a cumulative effect reflecting a concern with feasibility testing, reusability, multifunctionality, and polytheoreticity.

The prospects for coordination of US and European efforts - with those of Japan as well - are excellent.

USER NEEDS

N. Ostler

4 classes of User:

(i) industrial IT product builders
(spell checkers, text compacters, text DB searchers, etc.)
(ii) language theorists

preferring raw, or skeletally/automatically analyzed data

(iii) language theorists
preferring 'reliable'/carefully analyzed data

(iv) subject-matter analysts
psychologists, sociologists, humanists, economists, education

Points:

1. Probably require increasing complexity of annotation/classification of mass data.
2. Constraints or availability of data are **not** only technical, but also legal.

3. **Market pull/justification for support** is likely to come:
   - in the short term from (i) - difficult for Europe?
   - in the (medium/long) term from (iv)
     . professional information (law, trade, etc.)
     . multilingualism, language teaching
     . cultural access, multi-media.

Leech added "education" as one of the points which should be emphasized.

CRITERIA FOR CORPUS COMPOSITION

J. Sinclair

For some users, a corpus should attempt to be a reliable record of a state of a language, general or restricted.
This means that its constitution should avoid both bias error and random error.

For other users, the only control needed is for random error. It is contended that a lot of activity can be stimulated by quantities of text regardless of source.

For random error only, the quantity of material that can be acquired nowadays is thought sufficient in principle.

For bias error, the difficulty of accumulating some types of data, e.g. conversation, should not be overlooked.

Some applications that are envisaged are too specialized for a corpus to be centrally provided in advance of need. Corpus providers should be explicit about principles of corpus design, so that ad hoc corpora can be assembled.

For corpora which for bias error, it is accepted that we do not yet know the critical

parameters and proportions, nor the best sampling technique. A cyclical process of improvement over time is envisaged, progressively aligning external and internal evidence.

Church stressed that for certain categories of users the available stuff seems adequate.

TEXT REPRESENTATION AND ENCODING

D. Walker

Speakers: M. Sperberg-McQueen (MSM), L. Burnard (LB), S. Johansson (SJ)

The Text Encoding Initiative (TEI) is a major international project sponsored by ALLC, ACH and ACL to develop guidelines and standards for encoding and interchange of machine-readable texts. It has substantial support from CEC and from NEH and the Mellon Foundation is the USA.

MSM noted that the TEI provides a basic comprehensibility which is crucial to the reusability of data. It allows for multiple annotations of the same material with an SGML-based markup language, that can be scaled up or down and is open-ended and extensible. A critical feature is that the tag sets come from the communities that work on a class of problems and reflect their consensus.

LB gave an overview of the TEI's implementation of SGML, in particular noting the revised proposals for linguistic analysis and the recommendations for documenting corpora. The forthcoming second draft of the TEI guidelines will consist of a reference specification, a case book of examples and a series of community-specific tutorials.

SJ outlined the TEI's initial proposals for spoken texts which seem to be able to deal with points raised at the end by R. Moore and D. Gibbon. In response to another question, it was noted that the TEI's character set recommendations can handle all major European languages and many others.

The TEI Steering Committee is now formulating more detailed plans for the TEI beyond 1992. The Japanese are very actively seeking funding to join the project. An interim year of dissemination, evaluation and testing is planned during which areas for further work (e.g. in linguistic resources and spoken texts) will be detailed. Software requirements for the widespread dissemination and use of the TEI guidelines are also being defined.

LINGUISTIC ANNOTATION

W. Martin

Two main presentations were given, summarized as maxims, viz.:
(Leech)

1. The raw corpus and the annotations should be easy to separate.
2. The annotation scheme should be automatically available to the end-user.
3. The user should also be informed how, and by whom, annotations were applied.
4. "Caveat emptor" principle - but users should be told of degrees of reliability.
5. As far as possible, consensual categories.
6. No standardization de iure (but perhaps de facto).

(Sampson)

7. Small and rich vs. big and skeletal.
8. Each language in its own terms.
9. Impose consistent categories vs. discover categories.
10. Names/addresses/weights/measures, etc. are as important as NP, relative clause, etc.
11. Biological taxonomy vs. logical analysis.

As a result of the discussion the following arose:

a. Dichotomy such as expressed in 7, 9 and 11 is not necessarily there, it could be interesting to see/find out how to combine e.g. small (+ rich) with large (- rich). In other words to rely on extensibility/enrichment procedures/tools.
b. The agreement relationships between data and theory based/inspired annotations should not be taken for granted.
c. Re-usability of annotations should be of primary concern.
d. Testable criteria therefore should accompany the annotation scheme.
e. The minimum consensus could be what has been called in ET-7 "minimal observational data" theories implicitly or explicitly assume/start from.
f. Annotations for separate levels should not abstract away from the others.
g. Assembling and annotating should also be viewed from a cost-effectiveness point-of-view, therefore development of flexible tools is important.

METHODS AND TOOLS

Y. Wilks

Marcus showed the general acceptability and applicability of a form of Church's tagging scheme to corpora such as Brown, Muc-3, etc.
There were low error rates on the order of 3-7% where "error" was defined against the hand-tagging of what one could call "privileged taggers". The test of error was always against some form of refined human intuition.

On this base, the FIDDICH parser produced partial parses (normally of noun phrases) and some heuristics due to Hindle were used to give higher level structure such as PP dependencies. The customers seemed to like the annotation scheme and it has been used as á basis for stochastic parsers as well.

Sinclair argued that we should recognize the a priori element in annotation schemes and the circularity of the notions of "error" and "confirmation" that follow from it. He urged that corpora suggest intuitively natural categories, but that these may not be traditional **linguistic** categories, e.g. noun/verb is not well supported by corpus statistics for English.

The problem seems to be whether annotation schemes are to be defended by linguistic intuition (Leech supported this), by discovery procedure (Sinclair's view) and whether the schemes are unique or potentially many (Della Pietra supported the latter).The final test must surely be whether any scheme, whatever its origin, leads to a testable outcome such as a translation. That would be a position compatible with standard hypothetico-deductive science.

KNOWLEDGE ACQUISITION FROM CORPORA

H. Thompson

1) Mutual information is a powerful tool for uncovering certain types of regularity in corpora.
2) Other measures are appropriate to other types of regularities.
3) A sensible integration of data-driven and hypothesis-driven exploration of corpora can produce powerful insights.
4) Machine-readable dictionaries are a specialized form of corpus with particular strengths and weaknesses. The weaknesses can be in part compensated for by appeal to textual corpora.
5) Different levels of prior annotation are appropriate to different knowledge acquisition goals.
6) Statistics is not the only route to the acquisition of "knowledge" from corpora.
7) Statistics as a means of presenting data is distinct from theories and models of the data.

B7.2 DISCUSSION AND RECOMMENDATIONS

The recommendations which came out at the end of the meeting will be channelled through NERC, which will take them into consideration for the study.
The recommendations can be divided in two different types:
1) Corpus building actions and National plans,
2) Specific issues.

B7.2.1 Corpus building actions

Liberman presented a proposal drafted by H. Thompson, which is enclosed (Appendix A), with a couple of preliminary remarks: 1) it is only a proposal for discussion, not something which can be legislated immediately, 2) section 0. of the proposal, after discussion with some NERC representatives, could be seen in the framework of cooperation between NERC and the Linguistic Data Consortium.
The proposal comes from a compromise between people wanting to produce something concrete and immediate and as big as possible, and people wanting something best designed and balanced.

After discussion on many issues related to the proposal - e.g. problems of uniformity, levels of coding and annotation, TEI conformity, documentation, standardization, copyright constraints, ownership of data collected, eventual profits - the proposal was approved only as far as point 0. is concerned.
The rest of the proposal can be accepted only as pointing at a very generic methodology subdividing the follow up effort in two subsequent steps (a) a rather modest collection of texts, b) collection of very large quantity of data) The various indications on different aspects (timing, languages, structure, computation, etc.) can be regarded only as pointer to problems area to be solved, not as proposals for concrete choices.
This staged approach is also in agreement with the planning of the CEC.

Suggestions and plans were presented for Italy (Zampolli) and France (Candel, Dendien, Lafon), strongly supporting actions in corpus building in the medium term and correlated to what is going on in the two countries.

In Italy there is a long tradition of work in this area. We have a Reference Corpus (about 20 million words) which is going to be enlarged (also with spoken texts).

We recognize the necessity of designing and building a stratified corpus which will constitute the core of a large shareable textual resource.
Part of it should be immediately annotated, for training purposes.
Further enlargement will be done for subsets whose priorities will be determined by previous experience and by the specifications coming out from NERC.

We stress the importance that a European action will be based on a set of agreed criteria, conventions, and principled design.
The guidelines of the TEI, as evaluated by the NERC, and strengthened by the contribution of the European expert groups, together with the outcome of the NERC should form the common starting point.

1. France has available the largest textual base in the world (FRANTEXT, 150 million occurrences), mainly composed of literary texts of the period 1789-1964. This data base can be interrogated from workstations in France and abroad. The United States had been provided with a duplicate of the corpus

situated in Chicago.
The texts of the data base can be interrogated through a query system, but they cannot be either copied or freely processed (except for particular agreements).

2. In order to complete this process, it would appear advisable to have available a very large French corpus including different types of contemporary texts. It would be necessary to proceed towards a well balanced choice through a range of different socio-linguistic variables, also including oral and multilingual data. Particular emphasis will be placed on the scientific and technical domains, in order to cover the most different needs, both public and private, of the sector of research or of the industry: linguistics and lexicography, terminology, speech recognition, multilingual lexicons. With regard to the annotating level, we think it should only be limited to those elements which can be obtained by automatic procedures. As a matter of fact, the annotation levels requiring human intervention, obviously excluding essential factual data, (bibliographical references, etc.), would lead to high production costs. This would reduce the volume of the recorded corpus, while, at this very moment, new and rich sources of textual data will realistically be available in the near future.

B7.2.2 Specific Issues

What follows is the literal text of other recommendations related to the specific issues.

D. Walker

- Address intellectual property rights issue from the standpoint of authors, publishers, and governmental agencies, to secure cooperation for research access and to motivate the development of markets for electronic distribution.
- Corpus developers should provide proper citations of the contents to give credit to their sources.

Hockey / Walker

- Explicit cooperation should be sought with the library and information science communities to ensure that corpus developments both can take advantage of library practice and can influence library practice to better reflect the needs for electronic documents.

R. Moore / D. Gibbon

- Contact should be established with the speech community:

  i) at the strategic level by means of formal links with the 'International Working

Group on Speech Corpora' chaired by Dave Pallet @ NIST.

ii) at the tactical level by exploiting the communication network provided by ELSNET, and

iii) at the practical level by establishing a working relationship with the ESPRIT SAM project (2589).

- Where possible, spoken material should be collected and handled in accordance with existing protocols and guidelines agreed with the international speech community (e.g. SAM procedure for recording, digitisation, storage, phonetic annotation, etc.).

G. Sampson

I have been hearing more than I want to about very large corpora of English. Obviously for some purposes size is important, but I feel there is also an element of boastfulness - "our corpus is bigger than your corpus" - and to me it is striking that even in 1992 an enormous amount of good work is still being done with Brown and LOB, which are fairly small and quite old but **available**.

If NLP is to advance with respect to European languages, it needs to have similarly standard, available corpus resources for these languages soon, and talking in terms 1000's of millions of words guarantees that won't happen.

My recommendation would be that we choose a size and specification of coverage for which it is realistic to produce matched corpora for all EC languages within 3-5 years - this might be as limited as Brown/LOB, 1 million words prior of published text per language, though possibility it might now be realistic to go somewhat bigger and broader - and propose to produce such corpora for each language (including English) to common transcription standards (which would certainly be more advanced than Brown/LOB) and make them widely available to stimulate cross-European NLP activity.

N. Calzolari

- Interrelations between Corpus projects and Lexical projects, at the levels of:
  - acquisition
  - description
  - representation
  - evaluation
  - standardization

- Eliminate, at least partially, the dichotomy between linguists and lexicographers. Need for a new profile: the "linguist/lexicographer" for a corpus based and linguistically sound methodology of building extensive computational lexicons for

NLP systems.

N. Ostler

- stressed the importance of corpus use in language teaching; attention should be given by the NERC consortium to the various prospective needs and technical requirements.

S. Johansson

I take it of that the idea of establishing a network of reference corpora may have two goals:

1) in general, striving towards some comparability as regards the sorts of texts collected and the type of encoding/annotation;

2) in particular, providing data for bilingual lexicography, contrastive analysis, and translation studies.

I will make some recommendations with special reference to the second goal.

A. Establish translation corpora (unless this has already been done). For each language, identify a set of texts that have been translated to all the other languages in the network. Create corpora of original texts plus translations.

B. Establish comparable corpora. Investigate whenever existing bilingual/multilingual corpora (such as the Danish-English-French corpora in contract law) can be usefully extended to cover all the languages in the network. Also, establish comparable corpora matching the translation corpora (making possible a study of"translationese").

C. In general, identify some general targets as regards particular types of texts to be collected and encoding conventions. Agree on a system of classification .... for texts (à la TEI), such that subcorpora can easily be extracted for particular purposes (perhaps for language comparison).

B7.3 Concluding Remarks.

The workshop has certainly reached its major goals.
A first overall view of the state of the art in the field, focusing on the crucial scientific and technical issues, was provided. The participants agreed that the Proceedings should be published. The deadline for submission of the final version of the papers is June 1992, and all participants agreed to contribute.
A comparison between the American and the European point of views was carried out in a very constructive way. US participants pointed out that the highest priority should be given to the task of collecting as much data as possible. European participants laid stress on the need of adopting a set of parameters and design criteria for the process of data collection to be properly guided.
Crucial points of discussion and disagreement were identified and put in perspective. It was felt that, in some cases, experimental data do not provide conclusive evidence yet, both for lack of data and difficulty of interpreting them. In general, and in particular where controversial decisions must be taken, it was contended that the choice between alternative solutions should take into account the point of view of perspective users. To be more concrete, consider the ongoing debate on the usefulness of providing linguistically annotated corpora. Linguists interested in verifying theoretical assumptions of some kind, may find it useless to deal with a corpus annotated according to a different theoretical framework. On the other hand, NLP system developers may be not interested in the process and nature of annotation as such, and rather wish to work on already annotated corpora.

The need to reach consensus on several aspects of corpora creation and processing (criteria for corpus composition, text typology, text representation, software etc.) was taken to be a pre-requisite for a multilingual, harmonized set of reusable corpora to be made possible.

The TEI was recognized as a major factor of standardization on both sides of the Atlantic, and all participants agreed on the opportunity of adopting the proposed TEI Guidelines. The tutorial on TEI organized in the framework of the workshop laid down a basis for useful interaction between the TEI and some of the major European Centres for corpus creation.

Contacts between different countries, communities and schools were promoted and established. NERC partners will highly benefit from this situation, and are now in a position to carry on their work on the basis of a constructive continuous consultation with both European and American actors in the field. As an example, an agreement was reached during the workshop on the production, through the channels of NERC and DARPA, of a CD-ROM containing textual samples in various languages, provided by both American and European purveyors and distributed at a production cost. The goal of this initiative is to provide easy access to textual material in MRF for researchers and developers, who would otherwise find it very difficult to gather data of this kind. These samples would give the opportunity of experimenting and evaluating the techniques of corpus Linguistics to a large community of prospective users, students included, and will strengthen a general awareness of the possibilities

offered by methods based on textual data, with particular reference to the creation of data-driven NLP components, but also to other emerging applications such as language learning and teaching, assessment of common word meanings in commercial and legal documents, new technology to aid library work, electronic publishing etc.

New trends have clearly emerged, which give strong indications as to priorities in research and development. Some examples follow. The need to refine and verify non trivial statistical methods for extraction of information concerning real language use, information which cannot possibly be tapped from other sources used in the past, such as: introspection, informant interviews, and the like. The urgency of deepening our knowledge about the nature, composition and general features of sublanguages, through which it is felt possible to tune NLP components to specific applicative tasks, by identifying and assessing distinctive intralinguistic parameters. The need to investigate and develop robust software to process and access large corpora. The necessity of designing filters to look, in the continuous flow of new, varying texts in MRF, for new uses and new linguistic phenomena to be taken into account in the development of NLP systems.

Mainly, it clearly emerged the need to identify and investigate different requirements from different types of corpus users, and to relate these requirements to the issues of corpus design and structure.

The participation, alongside DG-XIII-B representatives, of official representatives of the major governmental American Agencies (DARPA and NSF, see enclosed letter) is taken to be of extreme importance. They clearly expressed their interest in the creation of reusable linguistic resources, as infrastructural basic tools for research and development, and their intention to explore the possibility of an international coordinated effort in the field, including policy and strategy issues. A formal cooperation was established between the newly created DARPA linguistic data consortium and NERC. Information will be exchanged on a regular basis, and the possibility of joint efforts to study and experiment new methods and techniques for corpus processing will be explored.