Lancaster
University
April 1994

Dear Participant,

Welcome to Talc94, the First International Conference on Teaching and Language Corpora. We hope that you enjoy your stay here in Lancaster, and that you will find the conference stimulating and worthwhile.

This conference has arisen as the result of interest shown at the last two ICAME ("International Computer Archive of Modern English") conferences, in Nijmegen in 1992 and Zürich in 1993. ICAME has long been the favourite gathering of a substantial group of committed corpus linguists, and during these last two conferences two things were becoming increasingly apparent:

1. many colleagues are now spending a great deal of time endeavouring to bring their interest in corpus-based research into their teaching

2. the interest in corpora and teaching is spreading so quickly that even regular ICAME participants cannot keep up with innovations, and lack a focused forum in which to share ideas.

However, we were not prepared for the sheer range and depth of involvements which were brought to our attention in the course of arranging this conference. Those attending Talc94 include many long-standing ICAMErs, but the majority of participants are not from this relatively close-knit community. We hope that this gathering will inform you about what is going on, and enable you to develop new contacts.

We have organised Talc94 as a single event. You may think it would be a good idea to hold it as a regular event. This is something we will address in our plenary discussions.

Finally, we would like to thank you for coming, especially those of you who have undertaken to organise workshops. We wish you a very enjoyable conference.

Best wishes,

The organisers

Note: most information on Campus facilities is to be found in the grey pamphlet "Lancaster University Conference Centre". Further detailed information (including catering facilities for those extending their stay) may be found in the ring-binder in Bowland Senior Common Room.

• Incoming messages

Incoming messages will be posted to the conference notice board in Bowland Senior Common Room.

FAX messages should be sent to the Linguistics Department and marked Talc94. FAX number: +44 524 843085.

Telephone contact may be made in emergencies by telephoning the University switchboard on +44 524 65201. Providing the caller gives your name and the name of the conference, they will be able to trace you quickly.

The organisers offer no general facility for taking telephone messages or calling people to the telephone. For non-emergency use, we suggest that you find the number of the telephone extension nearest to your room, and give that to anyone who might want to call you.

• Communication at the conference

Conference delegates are encouraged to use the conference notice board which will be situated in Bowland Senior Common Room. Note also that many of the telephones you see around the campus are internal only and may be used free to make calls around the University.

• Useful telephone numbers

Taxis

Local:	32090
To Manchester Airport:	417234 (Airlink) 411097 (Transair)

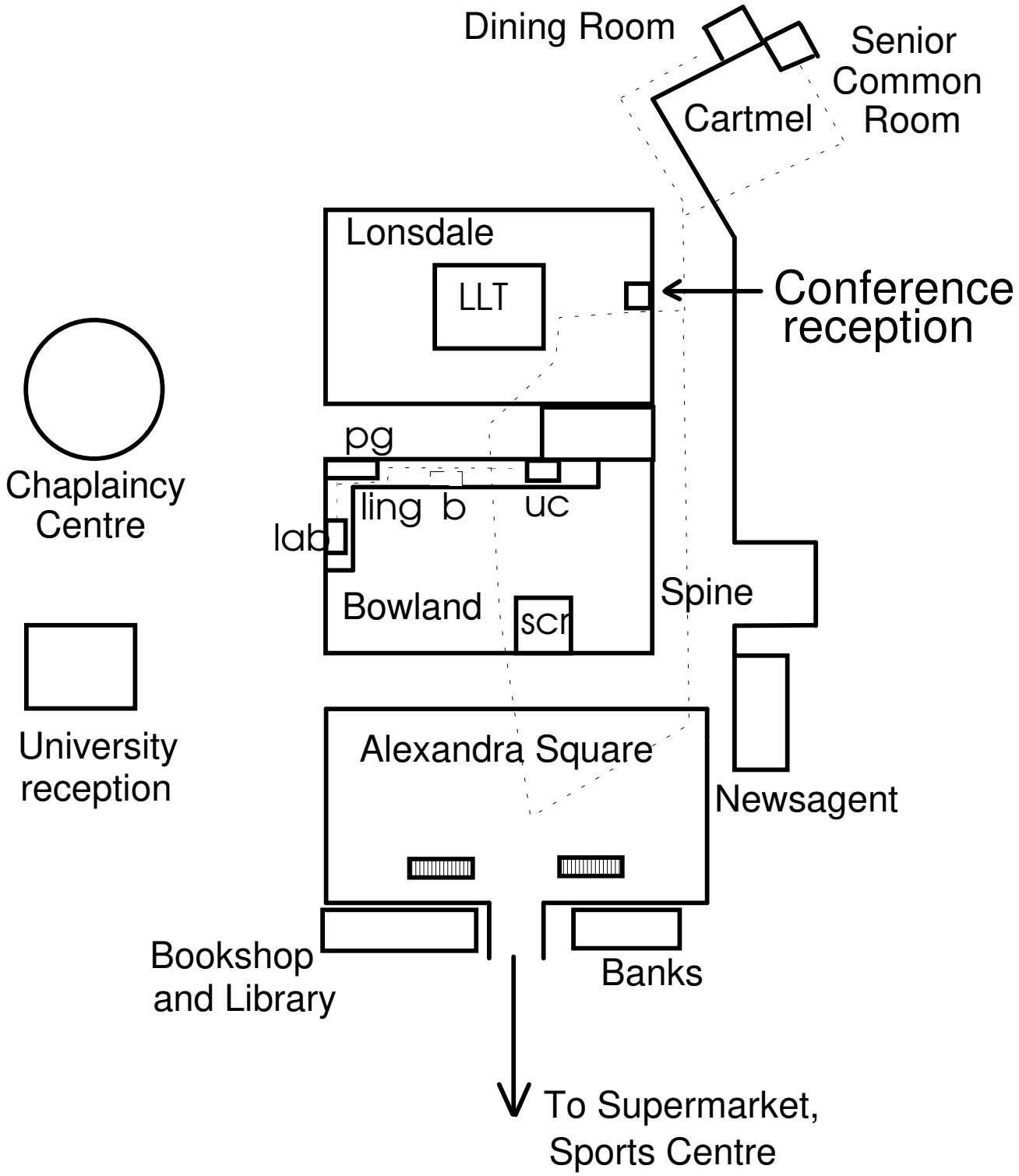
Train enquiries

Lancaster:	32333
Preston:	0772 59439

Doctor

University health centre:	Extension 4130/4134

Most of the places you will need to find lie on the circuit marked by the dotted line. The venues for the conference papers, refreshment breaks, demonstrations and workshops are indicated by the key at the bottom of the page.



LLT	Lonsdale Large Lecture Theatre
ling	Linguistics Department (1st floor)
lab	Linguistics Postgraduate Computing Lab (B71)
pg	Postgraduate Teaching Room (B80)
uc	UCREL Lab (B102)
b	Linguistics “Boardroom” A68 (Ground floor)
scr	Bowland Senior Common Room (Ground floor)

PROGRAMME FOR TALC94

SUNDAY 10TH APRIL – DAY 1

REGISTRATION AND TEA (4.00 – 6.30, Lonsdale College Foyer)

DINNER (6.30, Cartmel Dining Room) followed by (at about 8 p.m.)

INFORMAL RECEPTION (Cartmel Senior Common Room), kindly sponsored by
Longman Higher Education

MONDAY 11TH APRIL – DAY 2

LATE REGISTRATION (9.00 – 9.30, Bowland Senior Common Room)

INTRODUCTORY SESSION (9.30 – 10.15, Lonsdale Large Lecture Theatre):

Welcome by the organisers and introductory address:

Geoffrey Leech (Lancaster University):
Text Corpora in Education: the Grand Design

COFFEE (10.15 – 10.45, Bowland Senior Common Room)

SESSION ONE – LEARNERS INVESTIGATE I (10.45 – 12.15, Lonsdale Large Lecture Theatre)

Pam Peters (Macquarie University):
Micro- and macro-linguistics

David Minugh (Stockholm University):
All the language that's fit to print – using British and American newspaper CDs as corpora

Elena Tognini-Bonelli (Birmingham University):
Corpus work and linguistic issues: the corpus of modern Italian at Birmingham University

LUNCH (12.15 – 1.30, Cartmel Dining Room)

SESSION TWO – LEARNERS INVESTIGATE II (1.30 – 3.00, Lonsdale Large Lecture Theatre)

William Dodd (Birmingham University):
Teaching German with the Mannheim corpus of German newspapers

Guy Aston (University of Bologna):
Training learners to use corpora

John Kirk (Queen's University of Belfast):
Teaching and language corpora – the Queen's model

TEA, and **PUBLISHERS' EXHIBITION** (3.00 – 4.30, Bowland Senior Common Room)

SESSION THREE – SOFTWARE DESCRIPTIONS (4.30 – 6.00, Lonsdale Large Lecture Theatre)

Adam Kilgarriff (Longman Dictionaries):
Corpus use at Longman

Hans van Halteren (University of Nijmegen):
Syntactic databases in the classroom

Tim Johns (University of Birmingham):
CONTEXTS: A computer program for generating concordance-based exercises

CONFERENCE DINNER (7.00)
To be held in the Dalton Suite at the Lancaster House Hotel, within 10 minutes' walking distance of the conference venue

TUESDAY 12TH APRIL – DAY 3

SESSION FOUR – DEVELOPING CORPORA (9.00–11.00, Lonsdale Large Lecture Theatre)

Gordon Inkster (Lancaster University):
First catch your corpus: building LANFRANC, a French language text bank for teaching purposes
Philip King (Birmingham University):
Creation and processing of corpora in Greek and Cyrillic alphabets
Randall Jones (Brigham Young University):
Using a corpus of spoken German
Federico Zanettin (University of Bologna):
Retrieving and editing texts in different languages: corpus-based activities for language learning

COFFEE (11.00 – 11.30, Bowland Senior Common Room)

SESSION FIVE – CORPUS-BASED TEACHING AND LEARNING MATERIALS I (11.30–1.00, Lonsdale Large Lecture Theatre)

Dieter Mindt (Free University of Berlin):
Corpora and the teaching of English in Germany
Eve Wilson (Kent University):
Corpora as a source of CALL exercises
Lou Burnard (Oxford University Computing Services):
The British National Corpus: what, why and how

LUNCH (1.00 – 2.15, Cartmel Dining Room)

GROUP DISCUSSIONS AND WORKSHOPS (2.15 – 3.45)

See section 4 for details. Venues may be the subject of announcements during the conference

TEA (3.45 – 4.15, Bowland Senior Common Room)

POSTERS, DEMONSTRATIONS AND DISPLAYS (4.15 – 6.00, Bowland Senior Common Room)

details overleaf

DINNER (6.00, Cartmel Dining Room)

EVENING EVENT: PLENARY DISCUSSION (8.00 – 9.30, Lonsdale Large Lecture Theatre)

TALC94 POSTER PRESENTATIONS AND DEMONSTRATIONS
(TO BE DISPLAYED DURING THE AFTERNOON OF 12TH APRIL 1994)

POSTER PRESENTATIONS

Vladimir Benko and Eduard Kostolansky (Comenius University, Bratislava):
Corpus work at the Faculty of Education, Comenius University

Simon Botley (Longman Dictionaries):
SPOCk – A computer program for generating concordance-based exercises

Jill Burstein (Educational testing Service, Princeton):
Parsing sentence fragments in automated test scoring

Jürgen Esser (Aachen Technical University):
On the theoretical status of intonation in spoken corpora

Laura Gavioli (University of Bologna):
Exploring texts through the concordancer: some notes on guiding students to process text corpora[1]

Kenji Kita, Yasuhiko Kato, Takashi Omoto, Yoneo Yano:
(Tokushima University and National Language Research Institute, Tokyo)
Automatically extracting collocations from corpora for language learning

John Milton (Hong Kong University of Science and Technology):
A corpus-based on-line grammar and writing tool for EFL learners[2]

Harris Papageorgiou, Lambros Cranias, and Stelios Piperidis
(Institute for Language and Speech Processing, Athens):
A new approach to automatic sentence alignment

Francine Roussel (University of Nancy):
Using mono- and bilingual concordances to foster the induction of some phonological rules of L2

The Lingua parallel concordancing project

Martin Wynne (Leeds University):
A Unix course for corpus linguists

Xu Xunfeng and Michael Oakes (Lancaster University):
The Hua Xia corpus of written Chinese and concordancing facilities

DEMONSTRATIONS:

Gavin Burnage (Oxford University Computing Services):
The British National Corpus

David Graddol (Open University):
!KwicTex – an interactive concordancer for the Acorn Risc machines

Hans van Halteren (Nijmegen University):
Linguistic databases in the classroom: the CLUES system

Adam Kilgarriff (Longman Dictionaries):
Concordancing software

Tim Johns (Birmingham University) **and Simon Murison-Bowie** (Oxford University Press):
MicroConcord – concordancing software by OUP

Tim Johns (Birmingham University):
CONTEXTS

Eve Wilson and Simon Lippmann (Kent University):
Corpus-based CALL software

[1] Laura Gavioli will be represented in her absence by Guy Aston of the University of Bologna
[2] John Milton will be represented in his absence by Gregory James of the Hong Kong University of Science and Technology

WEDNESDAY 13TH APRIL – DAY 4

SESSION SIX – CORPUS-BASED TEACHING AND LEARNING MATERIALS II (9.00 – 10.30, Lonsdale Large LT)

David Quinn (City University) **& Akiva Quinn** (University College London):
Linguistic modelling for a corpus-based CALL system

Hans Paulussen & Guy Deville (University of Namur):
Automating vocabulary learning material: a step by step experience

Jean-Marc Langé & Elizabeth Bonnet (IBM Scientific Center, France):
The multiple uses of parallel corpora

COFFEE (10.30 – 11.00, Bowland Senior Common Room)

SESSION SEVEN – THEORY AND PRACTICE I (11.00 – 12.30, Lonsdale Large Lecture Theatre)

John Sinclair (Birmingham University):
Rumours about English

Bill Louw (University of Zimbabwe):
The role of the classroom concordance in the revaluation of traditional literary critical terminology

Bernhard Kettemann (University of Graz):
How POW can help us decide between UG and CC

LUNCH (12.30 – 1.45, Cartmel Dining Room)

SESSION EIGHT– THEORY AND PRACTICE II (1.45 – 3.15, Lonsdale Large Lecture Theatre)

Andrea Davies & Khurshid Ahmad (University of Surrey):
The role of corpora in promoting and studying Welsh

Glyn Holmes (University of Western Ontario) **& Marilyn Kidd** (Huron College, London, Ontario)
Scanning electronic corpora to develop language skills

Gerry Knowles (Lancaster University):
Using corpora for the diachronic study of English: a new approach to historical linguistics

TEA and **PLENARY CLOSING SESSION** (3.15 – 4.00, venue to be announced)

Guy Aston
Scuola Superiore di Lingua Moderne per Interpreti e Traduttori, University of Bologna, Italy

Training Learners to Use Corpora

Corpora are potentially tools for the learner as well as the teacher, but like dictionary use, corpus use is something which needs to be taught. Over the last year we have been experimenting with the use of newspaper corpora (The Independent on CD-ROM, and the Oxford Micro-concord newspaper corpus) in a self-access context at the School for Interpreters and Translators of the University of Bologna. While not representative of the language as whole, these corpora have advantages for language learning: they derive from a single genre with which the learner may be relatively familiar, facilitating the understanding of exemplars and reducing their contextual variety. And while some recurrent patterns may be specific to the genre, they are ones which educated native-speakers are likely to know and which, therefore, most learners should probably learn.

Learners first practiced using these corpora in seminars to which they brought their own language and translation problems. Most of these concerned either a) the contexts and cotexts of use of particular expressions; or b) the lexis used in conjunction with particular topics and situations. Both these uses potentially concern typicality/unmarkedness (finding out what is recurrently done) as well as feasibility/appropriacy (finding out whether/how something can be done). However while information concerning appropriacy was immediately appreciated as supplementing that provided by dictionaries, information concerning typicality was less readily recognised and valued, and learners often had difficulty in identifying regularities and in discounting distracting instances. Using corpora as a guide to typicality may require a conceptual shift in the learner's view of language use - moving from a dictionary-and-grammar view of selecting lexical items and applying combinatory rules under felicity constraints, to include a more holistic, schematic one of reproducing or modifying situationally-associated idiomatic chunks. While experience of corpus use seems likely to encourage such a shift in perspective, in teaching it may be appropriate to initially propose the corpus as a source of model texts on particular topics and of sample sentences using particular expressions, only subsequently treating it as a guide to recurrent schemata.

Vladimir Benko and Eduard Kostolansky
Faculty of Education, Comenius University, Bratislava, Slovakia

Corpus Work at the Faculty of Education, Comenius University (poster)

Full-text retrieval systems and resources have been chosen as a “highlighted” direction in the computerization of humanities at the Faculty. Three reasons for such a decision are as follows: (1) interesting, relevant and ready-to-use data are available; (2) user-friendly, efficient and Slovak-localized software (WordCruncher) is at hand; (3) there are good experiences with this approach at our partner institution (L. Stur Linguistics Institute). The “positive contagion” approach has been adopted to overcome the natural resistance towards technology among humanists.

Six corpus-related projects that have been carried out at our Faculty will be presented: (a) a Slovak explanatory dictionary and (b) two encyclopaedic works processed in the form of a corpus; (c) a (traditional) Slovak 20 million word-form corpus; (d) a corpus of Slovak poetry; and (e) a corpus of students' theses, and (f) two corpora to support foreign-language teaching.

In our opinion, the creation and use of corpora in an educational institution can be useful for at least three reasons. First, they provide a "large-scale" resource of linguistic and extra-linguistic information; second, they provide feedback into the process of teaching; and third, it will contribute to information literacy in general. Students will learn how to specify a search task, how to structure data and what to expect from the use of computers.

Simon Botley
Dictionaries Department, Longman Group Ltd., UK

SPOCk: A Corpus-based Linguistics Learning Tool (poster)

This paper describes a pilot corpus-based linguistic analysis tool for students on courses in syntax and semantics. The system is known as SPOCk (for Subject Predicate Object Complement). The SPOCk system enables students to practise sentence analysis on pre-selected computer corpus texts, or on texts selected by the students. More specifically, SPOCk is designed to support the teaching and practise of Functional-syntactic analysis (Jackson, 1990). Such analysis can be carried out in laboratory-based classroom sessions, either individually by students, or under tutorial supervision.

The SPOCk system has been implemented in Microsoft C Version 6.0 on a 386SX PC platform, which is close to the target hardware configuration found in the computer laboratories at Lancaster University. The system reads and displays sentences from a corpus file on the screen for the student user to analyse using a set of alphabetic mark-up symbols, which form labelled brackets similar to those used in the work of Leech et al (See Black, Garside and Leech, 1993). These symbols can be entered rapidly by the user at the keyboard. All changes can be saved to an output file for further processing. As well as sentence analysis, the SPOCk system provides extensive on-line help, file browsing, and lexicon look-up facilities to enhance its pedagogical value.

As well as the above features, there is plenty of scope for the further development and expansion of the SPOCk system. One area of intended development is that of statistical analysis of student output files in order to effect user profiling which will enable a tutor to match text difficulty to student ability. The SPOCk system, therefore, provides a good starting point for further research into computer-aided linguistics learning methods, and serves to show how computer corpora may be used in the teaching of linguistics.

Lou Burnard
Oxford University Computing Services, UK

The British National Corpus : What, Why, and How

The British National Corpus is a co-operative research project, jointly funded by DTI and SERC under the JFIT programme. Partners in the venture are Oxford University Press, Longmans UK, Chambers, Lancaster University, the British Library and Oxford University Computing Services. Its chief deliverable will be a one-hundred million word corpus of Modern British English, automatically annotated with part of speech codes and tagging using a simple TEI-conformant mark up scheme. The corpus has been carefully designed to include samples of as many different kinds of English as possible -- including both spoken and written material, from pub conversations to radio shows and formal lectures, from paper bags to best sellers and technical manuals. The BNC will be freely available for academic research and teaching purposes later this year. The corpus will be distributed together with technical documentation and simple SGML-aware retrieval software, initially for the MS Windows and Motif environments.

Lou Burnard manages the team which has been handling corpus processing at OUCS since the project began in 1991. His presentation will describe what the corpus contains, how it is being constructed, and why its availability will change the face of linguistic studies in the nineties.

Jill C Burstein
Educational Testing Service, Princeton, New Jersey, USA

Parsing Sentence Fragments in Automated Test Scoring (poster)

INTRODUCTION
The text corpora generally used for computational linguistic research come from full-text sources, such as newspapers and academic journals. We are examining a different type of text corpora, specifically, "short-answer responses" to prototype test items. Examinees must give quick responses during the test in 5 - 15 words. Many of the responses are in sentence fragment form, e.g., "WEATHER TOO COLD," and "PARENTS DAY AT SCHOOL," which is acceptable since responses are scored for content rather than grammar.

AUTOMATED SCORING OF TEST ITEMS
We have developed an automated scoring system which handles "fragmented responses" (Kaplan and Bennett (1993)). We can successfully score these responses for content; that is, our system searches a response for semantic concepts according to criteria preset by human scorers. For example, an acceptable response might require the concepts "police" and "bullet-proof vests." The following sentence fragment response would be scored as correct since it contains both concepts.

(1) BULLET-PROOF VESTS FOR POLICE

IMPLICATIONS FOR TESTING

Testing in foreign language and higher education is incorporating more items which elicit sentence fragment responses. Therefore, our corpora of short-answer responses is increasing along with the need for advanced methods of automated scoring. We are augmenting an ATN parser to enhance our present scoring system. The parser will handle "sentence fragments." It will facilitate "lexical lookup" for responses by assigning syntactic labels to lexical items and entire sentence fragments. The parser will also be able to verify acceptable "fragment structures." This is necessary since fragment structure can alter the meaning of a response, e.g., "FOOD THAT CAN GROW IN BRINE," and "FOOD GROWN IN BRINE." Developing a grammar for non-standard text corpora is a crucial step in the evolution of natural language research, since until now "full-text" has received most of the attention in corpus-based analysis. The computational ability to parse fragments would open many possibilities for automated language processing in testing.

REFERENCE

Kaplan, R. and R. Bennett. (1993). "Using the Free-Response Scoring Tool to Automatically Score the Formulating-Hypothesis Item". Educational Testing Service, GRE No. 90-02b.

Andrea Davies and Khurshid Ahmad
Department of Linguistic and International Studies/Department of Mathematical and Computing Sciences, University of Surrey, UK

The Role of Corpora in Promoting and Studying Welsh

A corpus-based study of general and special language texts in Welsh is presented with a view to demonstrate the efficacy of corpus linguistic methods and techniques for teaching and learning the Welsh language.

The assessment and evaluation of end-user requirements, e.g. first- and second-language teachers and learners, is discussed. Aspects of the design and creation of a contemporary general language corpus of Welsh are presented together with the details of a 100,000 words plus corpus of Welsh. We show how this corpus can be used in the elaboration of entries in a typical dictionary and in highlighting the syntactic and semantic behaviour of Welsh verbs. A novel approach for text-based terminology management is presented that enables a terminologist to create term banks with the help of a specialist language corpus of domain-specific texts. A corpus of Middle Welsh has been analysed with a view of producing a 'dictionary' of that language. The identification of computer-based, language-informed and low-cost methods for producing teaching and learning resources for languages that are spoken by numerically smaller communities is the central theme of this paper: progress on this front comprises much of the paper.

Bill Dodd
Department of German Studies, University of Birmingham, UK

Teaching with the Mannheim Corpus

The Department of German Studies has recently begun to use part of the Mannheim corpus of written German (specifically, the 'Bonner Zeitungskorpus') for teaching purposes, using the principles of data-driven learning developed in the English for Overseas Students Unit at Birmingham.

This work falls into two main areas: (1) developing materials and methods for undergraduate language learning (using edited print-outs as well as direct 'hands-on' manipulation of the text-files). (2) Using concordancing techniques in the 'Modern German Language' module in our M.A. to stimulate linguistic debate and research amongst postgraduate students.

This paper will report on our work and present some of the materials which we have developed

Jürgen Esser
Institut für Anglistik, Rhein.-Westf. Technische Hochschule, Aachen, Germany

On the Theoretical Status of Intonation in Spoken Corpora (poster)

Corpus analysis in the field of grammar is concerned with types of linguistic units (e.g. word forms) or types of linguistic categorization (e.g. article, noun phrase). These types are independent of the medium of realization, i.e. they exist in the written and spoken form (medium-transferability) and show, according to Lyons, that language is an abstract structure.

Intonational features, as used in transcriptions of spoken corpora, have a different status. Their types (e.g. nuclei, falls, rises, high key, low key) may be regarded as elements of a given language that are tied to spoken realization. Their use in texts is not a matter of recurring patterns of grammatical correctness but of pragmatic appropriateness. Therefore, the ensemble of intonational features in a text is in the first place a configuration that is chosen by one individual (the speaker or reader) on one occasion. Thus the interpretation of intonational information in spoken corpora must take into account not only that the transcription depends on the theory behind the transcription model (Stockwell's old representation question), but also that other realizations by the same speaker and particularly by other speakers are possible in the case of reading texts, which form major parts of spoken corpora. Orally originated texts, on the other hand, do not have this indeterminacy. So the two kinds should be more clearly separated.

The intonation in spoken corpus data requires a kind of interpretation that is different from grammatical analysis: whereas the tagging of a word form with a grammatical tag is independent of the medium and the speaker and can be seen as an aspect of 'langue' (it is rule-governed), the tagging of a word form with an intonation tag depends heavily on the recorded speaker and his pragmatic achievement as a whole at a certain time. This cannot be directly regarded as rule-governed but rather as governed by pragmatic principles of textual rhetoric. It is these pragmatic principles that we are looking for in

intonation data. They can be traced in orally originated, spoken texts and in texts read by one person or several persons in parallel readings. This last aspect has so far been ignored in computerized spoken corpora.

Laura Gavioli
Scuola Superiore di Lingua Moderne per Interpreti e Traduttori, University of Bologna, Italy

Exploring Texts through the Concordancer: Some Notes on Guiding Students to Process Text Corpora (poster)

The use of computer text corpora in linguistics research has led to an increasing use of them also in language teaching. Exploitation of computer text corpora with text-retrieval and concordancing software has proved it a means to confirm or "stimulate" students' intuitions about foreign language use particularly from a lexico-grammatical (collocational) point of view (Tribble & Jones 1990, John & King 1991).

This paper proposes activities that guide students to study texts and text corpora through concordancing and text-retrieval software and help them interpret the data retrieved. Besides being familiar with the software, processing a corpus of authentic texts involves a range of levels of linguistic and meta-linguistic knowledge. So students who approach the concordancer for the first time are faced both with the difficulty of searching relevant patterns of language and of giving a correct interpretation to the results of their search. Students need to be guided to exploit search methods to obtain particular outputs and to be made aware of the fact that corpora with different characteristics may provide them with different outputs. This paper discusses possible activities which guide the students to see different interrelations existing between the theoretical criteria adopted in building up the corpus and the data obtained through computerized processing of that corpus.

In the curriculum of translators and interpreters, which I am particularly concerned with here, such activities may also help students to use text-retrieval software to create corpora or subcorpora *ad hoc* to satisfy individual linguistic or professional needs.

REFERENCES

Johns, T. & P. King 1991: 'Classroom concordancing', *ELR Journal,* Vol. 4 (special issue).
Tribble, C.T. & L.R. Jones 1990: *Concordances in the Classroom*, London: Longman.

Hans van Halteren
Department of Language and Speech, University of Nijmegen, The Netherlands

Syntactic Databases In The Classroom

For some years now, the English Department at Nijmegen University has used a computer training tool, the CLUES system, to enhance its syntax course. CLUES automatically selects appropriate utterances from a syntactically analysed corpus to present to the student in one of several exercise forms.

All exercises are based on suppressing information in a presented analysis tree of the utterance and letting the student supply what has been suppressed. Supported exercise types are:

- filling in function or category information for constituents
- answering questions about constituents
- pointing out specified types of constituents in the tree
- replacing removed connections or nodes in the tree

The individual exercises are not hardwired into the system. A teacher can design his own by way of Exercise Schemes, in which he can specify what kind of utterances must be selected and what exercise is desired. In fact, CLUES is an adaptation of the Linguistic DataBase system, designed for research on syntactically analysed corpora, and Exercise Schemes are an extension of the Linguistic DataBase system's Exploration Schemes.

The choice for automatically selected corpus utterances rather than hand-crafted examples has its advantages and disadvantages. A clear advantage is that a corpus, if large enough, contains a virtually infinite supply of examples. Furthermore, we think that the "real-life" quality of the examples is more stimulating to the students than the sometimes obvious artificiality of crafted ones. One the other hand, there is very little control over the parts of the utterance not directly related to the selection process and this may lead to examples which contain constructions as yet unknown and hence confusing to the students.

Our experience with the system, and the overall enthusiastic reaction by the students, lead us to believe that the advantages outweigh the disadvantages and that we should further develop the CLUES system.

Glyn Holmes[a] and Marilyn Kidd[b]
[a]Department of French, The University of Western Ontario, Canada
[b]Department of French, Huron College, London, Ontario, Canada

Scanning Electronic Corpora to Develop Language Skills

The increasing availability of electronic texts and software to manipulate those corpora offers interesting possibilities for language teachers and learners. In this paper the authors--one a primary school teacher, the other a university teacher of French--will explain and demonstrate how such electronic tools can be used, at all educational levels, by first-and second-language learners alike, to develop language skills.

The electronic tools will first be briefly described. TACT is a text analysis system developed at the University of Toronto and available at cost from the Center for Computing in the Humanities. This software is relatively sophisticated, but the applications described will involve quite simple and easily learned elements of the software (mainly word lists and

concordances). The corpora used include several English language children's stories, current English and French language newspaper articles, and a number of French literary texts.

The presenters' main objective will be to show how students can search a given corpus to retrieve specified datasets on the basis of which they can formulate and test hypotheses about linguistic phenomena: grammar, syntax, semantics, morphology, orthography, punctuation.

For example, learners of English often have difficulties with prepositions. Students can scan a corpus, retrieve all occurrences of the word in and formulate (or test) rules about the semantics of the term ("Does it always mean inside?") about its position in the sentence ("Does it ever end a sentence?") and about its syntagmatic relationships to other elements of the sentence ("What kinds of words follow it?"). A similar examination can be done of the multifaceted French prepositions *de* and *à*. Other examples will relate to verb forms, prefixes and suffixes, relative pronouns; to orthographical difficulties including homonyms, homographs, doubling of consonants, use of contracted forms (let's, isn't in English; au/aux/du/des, etc in French); and to questions of punctuation. In reality, the possibilities are myriad.

A final consideration will be given to the pedagogical approaches and applications which the above electronic tools can serve. Many language courses are still based on deductive and prescriptive methodologies and indeed our tools could well be use to find examples to fit the rules. However, the tools might best serve an inductive approach where the students can formulate their own descriptive hypotheses and rules. Moreover, there are numerous remedial applications, depending on the types of corpora used.

Randall L. Jones
College of Humanities, Brigham Young University, Utah, USA

Using a Corpus of Spoken German

My paper is based on a recently completed corpus of spoken German. During the past three years I have collected approximately 400 interviews with native speakers of German in sixty areas of Austria, Germany, and Switzerland. Each interview is about twelve minutes in length and deals with one of a variety of topics, including occupations, education, hobbies and other special interests, travel, politics, current events, reminiscences of the past, etc. The speakers reflect a cross section of the population in each country with regard to geography, age, education and gender. Each interview has been transcribed and proof-read, then indexed using the text retrieval software WordCruncher. Any word or phrase in the approximately 700,000 word corpus can now be accessed instantaneously.

The corpus is currently being used at Brigham Young University by faculty and students to investigate certain lexical and syntactical properties of modern colloquial (spoken) German. In spite of the limitations of a spoken corpus based on short interviews, the corpus does reflect well most of the features of the German language as spoken today. In the Lancaster

presentation examples will be illustrated using German particles, pronouns, conjunctions, prepositions, future tense, passive voice and subjunctive mood.

Gordon A Inkster
Department of Modern Languages, Lancaster University, UK

First Catch Your Corpus: Building LANFRANC, A French Language Text-Bank For Teaching Purposes

The Lancaster corpus of French texts was conceived as a source of empirical examples for teaching purposes. It was initially created by homogenizing the format of already available texts, mainly literary, drawn from several sources. Subsequent expansion of the collection, and its extension to include further varieties of language, has become the subject of a special undergraduate course. Students create and edit a machine-readable text prior to its exploitation and choose a simple mark-up convention as a function of the information they wish to extract. This creates an awareness of numerous features of French that are frequently overlooked elsewhere in their language learning.

The increasing range of network accessible French material makes it easy to add potentially ill-formed material and introduce those collating it to living processes of linguistic change.

Tim Johns
English for Overseas Students Unit, University of Birmingham, UK

CONTEXTS - A Computer Program for Generating Concordance-based Exercises

Over the past eight years, corpus-based work in the English for Overseas Students Unit at Birmingham University has focused on the development of a classroom methodology ("Data-driven learning") that makes central use of computer-generated concordances and concordance-based materials.

This paper will give an overview of the DDL approach as currently practised in the Unit, and will describe the construction and trialling of CONTEXTS, a computer-assisted language learning program that incorporates certain features of that approach.

Bernhard Kettemann
Department of English, University of Graz, Austria

How POW Can Help Decide Between UG and CC

This paper argues that the use of corpora in linguistics teaching is rewarding and necessary. It describes using the (POW) Polytechnic of Wales Corpus of child language in a course on linguistic theory in order to determine the importance of quantity and frequency factors in the acquisition sequence of grammatical morphemes in English, thus using corpus driven research to

decide on the respective explanatory power of e.g. universal grammar and parameter setting (UG) vs. cognitive constructivism and self-organization (CC).

The standard approach to the question of the influence of input frequency on the acquisition sequence is that there is no correlation between order of acquisition and input frequency and quantity. Brown (1973:362), Dulay/Burt/Krashen (1982:38) and Lightbown (1983), for example, see no relation between parental or tutorial frequency and the child's order of acquisition. Other findings seem to contradict this, though. For example Larsen-Freeman (1976) and Long (1981) found that there is a correlation between tutorial frequency and acquisition order.

Let us then have our students approach the problem on a very tentative but exemplary basis. This is an important and interesting question and relatively easily answered using the appropriate research tools.

They correlate the established acquisition order with the general frequency percentages of the respective morphemes in the corpus, e.g.

Acquisition Order vs. Mean Frequency in %

1. -ing (overall, not only Progr.)	?
2. -s (Plural)	?
3. Irregular Past	?
4. -'s (Possessive)	?
5. Regular Past	?
6. be + ing (full Progressive)	?
7. -s (3rd pers., in full verbs)	?

The frequency figures will be the result of a concordancing program like LMC, WCS, MicroConcord being run over the corpus extracting these morphemes.

Please note that whatever the result of this investigation, we are not saying that frequency is the only factor that influences acquisition order. What we are saying is, that corpus investigation has many possible fruitful applications in linguistics teaching: in the course of this small project the students have been exposed to a basic linguistic debate, to an interesting research question, have been introduced to methods and data etc.

Adam Kilgarriff
Longman Dictionaries, Longman Group Ltd., UK

Corpus Use At Longman

There are two things you need before you can do any work with corpora : a corpus, and software for looking at it. At Longman we have state-of-the-art versions of both.

The corpora are the Longman-Lancaster corpus, the Spoken English component of the BNC (as it becomes available), and the Longman Learners

Corpus. The Longman-Lancaster corpus is both large (26M words) and covers a wide range of text types, so has both the ingredients which Biber (1993) shows to be essential for corpus-based accounts of 'General English'. Each word is tagged for part-of-speech. The Spoken and Learners' corpora add to the variety of text types covered in the corpus database. Each document in each corpus has a 'header' stating genre, region, speaker/writer's age, gender, dialect and so forth.

The Corpus Workbench is a software package produced to Longman's specifications by Textware, a Danish software house specialising in linguistic applications. It provides very rapid access to concordances for words, lemmas, collocations and other phrasal units, and, where the corpus is part-of-speech tagged, to word-and-POS combinations. Its functions include sorting, sampling, and statistical reports, and all functions can be applied to subcorpora as defined by header information. The interface is windows-based and shows, at a keystroke, the complete text a concordance line comes from or the header for that text.

With this setup, many things are possible. For example, a comparison of the relative frequencies of the two forms of ditransitive verbs - GIVE PERSON THING vs. GIVE THING to PERSON - in written and spoken English can be made. This would exploit the facilities for searching-by-header-field (written vs. spoken), POS-tagging (to see, e.g., where a pronoun or name immediately followed the verb), multi-word searches (for the verb + 'to', where 'to' counts as a hit if it comes anywhere between two and six words after the verb), and statistical reports.

Ref: Biber, D. 1993: 'Using Register-Diversified Corpora for General Language Studies'; *Computational Linguistics* 19(4) pp 219-242

John Kirk
Queen's University of Belfast, Northern Ireland, UK

Teaching and Language Corpora: the Queen's Model

Corpora of English are used at the Queen's University of Belfast for the purpose of enabling students to learn about the structure of English, develop a descriptive and theoretical vocabulary, and cultivate a methodology for dealing analytically with, and writing effectively about, language.

The importance of corpora is their role as models. As models, corpora can be studied in terms of data design and their collection, transcription and a growing range of annotations and enhancements (such as grammatical tags and parses); they can be studied for the many existing qualitative and quantitative descriptive or taxonomic analyses based on their data as well as the results and conclusions drawn from these investigations; and they can also be used as case-studies for testing existing, often general, hypotheses about language, speakers or texts. As papers at ICAME, NWAVE and a growing number of one-off conferences (e.g. 1991 Nobel Symposium, 1993 GURT Round Table; 1993 Ohio Conference on Very Large Corpora) are confirming over and over again, the corpus paradigm - subject to and

successfully vindicative of all the standard constraints on scientific enquiry - is, for the study of language, the strongest there is. It is therefore right that students of English should be made aware of, and come to assimilate, these achievements.

For some tutors in Corpus Linguistics, it might be sufficient to get to this point. In the Queen's approach to Teaching and Language Corpora, it is no more than a very big starting point. Our goal is the replication of the methodology on an extremely reduced scale, in the form of a project. Students are encouraged to find their own data and create their own micro-micro-corpus. If at all possible, the data should be spoken, transcribed (minimally orthographically, but perhaps with some prosody if appropriate), and enhanced (e.g. tagged or parsed, if appropriate and the necessary software is available). If written, it should preferably be non-literary (e.g. from newspapers). A model should be selected (usually from the existing scholarly literature on main corpora), the range of possibilities is enormous, and innovation is also encouraged. The model is then applied to the data so that students come to analyse their very own data. The discovery of items and patterns, including frequencies, is often a source of great delight.

Some of the models chosen, such as the co-operative principle, or politeness, or models of discourse, go a long way towards characterising afresh any spoken data, and thus the exercise serves to test the students' knowledge of these models as much as their successful characterisation of the new data. Other models can be quite specific, such as variation between root and epistemic senses of modal auxiliary verbs or the distribution of modal verb exponents of obligation and necessity, tokens of which, if they don't occur in sufficient abundance in data envisaged by students, can be provided by the tutor in a concordanced format from unanalysed Scottish or Northern Irish material. The results should therefore include comparison of the students' data, for which there any no absolute comparisons anyway, with those of closest possible data available from anywhere else.

In these ways, students replicate the Corpus Linguistics method. They come to assimilate the paradigm and, in an elementary way, come to make it their own. The comparison of their results with others serves to validate their findings and further their experience of research methodology: if the sets of results are similar, what does this suggest? And if they are different, how can that be explained? Often the results are different, and this has to do with the nature of the data used not only by the students but those in the corpora used by other investigators. Thus it is essential that students are able to provide as an early section of the project a clear description of their data. Students are encouraged to extrapolate out from their data and results - the object of their investigation - and draw general conclusions about a general theme (such as the universality of discourse strategies or the homogeneity of English within the British Isles).

The final stage of this work is presentation: a project has to be written up, effectively. While guidance and a general structural format is laid down, individual flair is encouraged. Writing up usually proves harder to students than they thought. An early difficulty is a clear description of their data - it is no longer sufficient to assume it as 'given'. A further recurrent difficulty is the presentation of the chosen model in such terms as they can use beneficially

later in the project. Another is distinguishing first in their own minds then on paper between 'results about the data' and 'conclusions about the general subject under investigation'. In the end, after the production of a spiral-bound, word-processed, laser-printed document, there is almost universal satisfaction for, behind the physical object, the results about the students data reflect some genuine originality now proudly attributable to the name on the cover. Thus there is a sometime frustrating, but in the end usually rewarding, emotional element which is never far away in learning about language through corpora.

Philip King
English for Overseas Students Unit, The University of Birmingham, UK

Creation and Processing of Corpora in Greek and Cyrillic Alphabets

Particular problems are posed for corpus-builders and text exchange by non-Latin alphabets. They include the lack of standardisation for ASCII-mapping, problems with software basically designed for Latin-alphabet languages, alphabetical sorting and screens and printers. The author describes his own work with a small corpus of Greek and demonstration corpora of Russian, Macedonian and Ruthenian, and shows what can be achieved using commercially available software.

Kenji Kita, Takashi Omoto, Ryo Okamoto and Yoneo Yano
Dept. of Information Science, Tokushima University, Japan

Automatically Extracting Collocations from Corpora for Language Learning (poster)

In language learning, learners must pay more attention to how words are used rather than to individual words by themselves. Collocational knowledge indicates which words co-occur with other words and how they combine syntactically. Collocations are, however, difficult to produce for second language learners. Thus, it is important to provide learners with collocations of wide-coverage.

We will present a new method for automatically extracting collocations from corpora by introducing cost criteria. Cost criteria make it possible to measure quantitatively the extent to which processing will be reduced in NLP systems by considering a multi-word pattern (collocation) as one unit. Therefore, cost criteria is a kind of modelling of learners' work load and it naturally reflects learning efficiency.

Since space is very limited, we describe the core idea for cost criteria. Assume here that, in the corpus, there exists a word pattern ‘A' which is composed of ‘|A|’ words, and whose frequency of occurrences is ‘NUM(A)'. Here, we define a cost ‘COST(A)' for a multi-word pattern ‘A' as

$$COST(A) = (|A| - 1) * NUM(A).$$

Also assume that it takes a cost of 1 to process one word. Similarly, when processing ‘A' as one unit, its processing cost is 1. If a word pattern is processed one word at a time, it is reasonable to consider that the processing cost is proportional to the length of the word pattern. That is, the processing cost for ‘A' is equal to ‘|A|'. By considering ‘A' as one unit, the processing cost is reduced to ‘|A| - 1'. Since ‘A' appears ‘NUM(A)' times, we can conclude that the total cost becomes ‘(|A| - 1) * NUM(A)', which is the definition of ‘COST(A)'.

If two patterns ‘A' and ‘B' contains some words in common, cost definition must be modified. Consider the case where a word pattern ‘A' is a subpattern of ‘B'. Then, we have:

$$NUM(A) >= NUM(B)$$

As for the word pattern ‘A', ‘NUM(B)' times out of ‘NUM(A)' times is referred to as ‘B'. Thus, the actual cost for ‘A' is defined as:

$$COST(A) = (|A| - 1) * (NUM(A) - NUM(B))$$

In the presentation of the final paper, we will include some experimental results to show the effectiveness of our method.

Gerry Knowles
Department of Linguistics and Modern English Language, Lancaster University, UK

Using Corpora for the Diachronic Study of English: A New Approach to Historical Linguistics

'History of the language' courses have an image problem. The historical approach flourished a hundred years ago, and is today perceived as old fashioned. Secondly, scholars have long since worked out the general development of the language, and the student merely has to follow in their footsteps. This makes for extreme teacher-centred courses with lots of facts and little scope for the students to make any contribution of their own. Student response to a new corpus-based course being developed at Lancaster makes us confident that these problems can be overcome, and that this is an area which promises exciting developments in the next few years.

Using historical texts - including the Helsinki corpus - in WordCruncher, students can test generally accepted views for themselves, e.g. on the distribution of *thou/ye/you* in early Modern English. It is of course only in relation to the variety in the raw data that textbook generalisations can have any real meaning. How deeply do students who only read textbooks understand what has happened in the language? In the study of semantic change, examples retrieved in context usefully complement traditional accounts. To tell students in a lecture that *meat* has changed its meaning gives them an isolated fact: to give them a text which they can only understand if they first infer that *meat* means 'bread' employs - and perhaps develops - their text processing skills.

When working on corpus texts, students reveal interesting things about their own learning which rarely emerge in conventional seminars. It is easy, for example, to persuade students to take a descriptive approach to language as a matter of principle. In practice, however, the attitudes they express when confronted by old grammar and spellings are often prescriptive. Students consulting the on-line OED tend to reveal the limits of their ability to make use of the information contained in lexical entries. This is something they can conceal when they are working on the printed OED in the library. The study of corpus texts requires the application of information which in more conventional courses can be reproduced in its original form. Does Shakespeare come before or after Caxton, and before or after Samuel Johnson? After watching students learning, one can understand how they can take a conventional year's course on the history of the language, and learn nothing useful at all.

And yes, I really have been asked why "s" used to be written "f"!

Jean-Marc Langé and Elizabeth Bonnet
IBM France Scientific Center, Fontenay-Sous-Bois, France

The Multiple Uses Of Parallel Corpora

INTRODUCTION

Collecting and using multilingual corpora (MLACs) has become an important activity of our MT group. Initially focussed on the statistical approaches to MT, we have since then been developing methods for bilingual terminology extraction. It turns out that MLACs, the basic fuel for such methods, has much more possible uses. We will describe some uses we have made of MLCs in our group, and other potential uses.

ALIGNED CORPORA

Most of our methods require that the bi- or multi-lingual corpora are aligned: each sentence in one language is linked to its translation. Some researchers also refer to such resources as "parallel" corpora. The methods are now well-documented. We have improved on some of them (e.g. Brown 91, Gaussier 92) and adapted ideas from others (Simard 92).

THE USE WE MAKE OF MLACS

We use MLACs in three areas:

- Our initial research on bilingual terminology extraction (Gaussier 92), which led us to undertake a European R&D project on this subject (Langé 92). Parallel corpora are essential since the sentence-to-sentence alignment greatly reduces the space of possibilities for word-to-word or term-to-term alignment.

- In the field of MAT, IBM has recently announced a translator's workstation, Translation Manager/2 (TM/2) which has a facility known as "Translation Memory" or "Repetition File", which consists in storing all translated sentence pairs and retrieving the translations when identical or similar sentences are to be translated. We have suggested that the

alignment methods could be applied to convert to Translation Memory all translation archives, provided that both source and target text are available. Thus, a new customer will be able to profit from his existing translations and enjoy productivity savings when starting using TM/2. This approach has been experimented with success in our own Translation Center, where thanks to a good archiving policy we have been able to increase the machine-aided translator's productivity by 50%.

- In the field of MT, we are permanently updating lexicons for our English-French LMT prototype. We recently facilitated the bilingual lexicographers work by proposing them on-line access to a bilingual aligned corpus. The user enters a word, and the system returns and pages through all bilingual sentence pairs containing the word. Such a process ensures that the adaptation of the lexicon to a specific domain indeed sticks to this domain lexical specificities.

OTHER APPLICATIONS OF ALIGNED CORPORA

Other applications of MLACs come to our mind, which are more or less developed. Most apply to M(A)T, one to Machine Aided Learning, and one to general Natural Language Processing:

- Statistical MT (Brown 90)
- Monolingual terminology extraction (Gaussier 93)
- Example-based MT (Sumita 91)
- Computer-aided teaching of Foreign Languages
- An MAT facility, whereby the translator could consult in context
- The existing bilingual base
- Other disambiguation methods in NLP (Dagan 92)

REFERENCES

Brown et al.: "A Statistical Approach to MT", Comp. Ling. 16, 1990

Brown et al.: "Aligning Sentences in Parallel Corpora", Proc. ACL 1991

Dagan et al.: "Two Languages Are More Informative Than One", Proc. ACL 1991

Gaussier et al.: "Towards Bilingual Terminology", Proc. ALLC/ACH, Oxford 1992

Gaussier et al.: "Monolingual Word Associations Through Bilingual Information", submitted to 3e journees scient. LTT 1993

Langé et al.: "A Mundus Novus in NLP: Corpus-based Complements to Rule-Based Methods", Proc. 5th Intl. Symp. on Know. Eng., Sevilla 1992

Sumita et al.: "Experiments and Prospects of Example-based MT", Proc. ACL 1991

Simard et al.: "Using Cognates to Align Sentences in Bilingual Corpora", Proc. TMI, Montreal 1992

Geoffrey Leech
Department of Linguistics and Modern English Language, Lancaster University, UK

INTRODUCTORY ADDRESS:

Text Corpora in Education: The Grand Design

In this introductory survey, I will try to map out the field of corpus-driven or corpus-aided teaching, taking account of present limitations and future possibilities.

I will characterize the advantages of corpus-aided teaching (?TALC) as follows:

(a) Computer corpora tend to promote a **learner-centred** approach to language learning and language education.
(b) **Open-ended** and uncensored supply of language data encourages an exploration and discovery approach to learning.
(c) The corpus enables the learning process to be **tailored** to the individual's needs and wishes.

I will then address the following issues, as a means of exploring and structuring the domain of TALC:

(1) What is the nature of the interaction between corpora and teaching? Steve Fligelstone (1992) distinguishes between:

a. Teaching about [corpora]
b. Teaching to exploit [corpora]
c. Exploiting [corpora] to teach

(2 a) What languages are involved?
(2 b) And are we thinking of native language, second language, or foreign language teaching?

(3) What kind of corpora (a) exist?
(b) are needed? e.g.
(i) ESP corpora
(ii) Developmental corpora
(iii) Corpora of learners' language
(iv) Bi-/Multilingual corpora

(4) Are these involved (a) directly, or
(b) indirectly in the learning process?

(5) What are (some of) the direct applications of corpora to teaching?
(a) Divergent methodology
(b) Convergent methodology

(6) What are (some of) the indirect applications of corpora to teaching?
(a) Reference publishing
(b) Development of language-learning materials
(c) Language testing

(d) Grading of texts and language-learning materials
(e) Simplification of dictionaries, texts and materials

(7) Where do we go from here?

Bill Louw
Department of English, University of Zimbabwe, Harare

The Role of Classroom Concordance in the Revaluation of Traditional Literary Critical Terminology

The use of corpora in the teaching of English literature will mean increasingly in the future that literary critical judgements are likely to rely less on introspective or intuitive evidence. The move away from intuition will initially take the form of verification of critical judgements against corpora of increasing magnitude, many of which are already available in the classroom. Definitions of style which hinged on the extent of deviation between a literary text and the norms of the standard language were long considered acceptable in theory but impossible to demonstrate in practice (e.g. the work of Bernard Bloch). Data-driven literary criticism is now in a position not only to verify and possibly vindicate such definitions, but is on the brink of providing entry points to criticism which are offered by the machine rather than the intuition of the reader. This goal was long cherished by stylisticians (e.g. Leo Spitzer), who even today have difficulty justifying their discipline as a 'respectable' form of linguistic enquiry because of its apparently ineluctable commitment to a subjective point of entry.

This paper will argue that because computers and corpora offer more autonomy to pupils in the literature classroom than in the past, we can expect the impetus towards data-driven criticism to come from the pupils rather than from the teacher. Furthermore, because the new technology will find easier acceptance among the young, teachers will need to adapt rapidly. New facts about the language and the nature of literariness are emerging almost with every key stroke. The paper will deal with a number of these insofar as they are likely to bring pressure to bear on teachers and on the conservative producers of dictionaries of literary terms to adjust traditional or working definitions of literary terms. The paper will examine the terms METAPHOR, IMAGERY and IRONY closely and point to ways in which other terms are ripe for revaluation in the light of the findings of data-driven literary criticism.

John Milton
Language Centre, Hong Kong University of Science and Technology

A Corpus-based On-line Grammar and Writing Tool for EFL Learners
(poster)

Although fluency in English is demanded by Hong Kong employers and entry to university requires minimal levels of performance, language classes account for only a fraction of the timetable in secondary schools and little time is given for remedial English courses at the tertiary level. Instructors in all content areas report that students' writing is unsatisfactory, which most

equate with 'bad grammar'. The persistence of errors in the writing of Hong Kong learners' English that the learners are, in the main, able to identify in discrete practice and tests suggests that students are able to address many of their errors - surface and otherwise - if they have available reliable advice during the writing process and if they are guided through systematic composing and proof-reading strategies.

This presentation describes the design and rationale of an electronic grammar and writing tool based on the analysis of a corpus of English learners' writing collected in Hong Kong. The project is underway at the Hong Kong University of Science and Technology and at the Department of Linguistics and Modern English Language, Lancaster University.

Dieter Mindt
Free University of Berlin, Germany

Corpora and the Teaching of English in Germany

English is the most important foreign language in German secondary schools. There is a vast range of textbooks for teaching English in Germany. These textbooks are generally based on a grammatical syllabus.

The foundations of this syllabus are, however, of questionable value. There are, at present, no grammars which are designed for the purposes of foreign language teaching from the very beginning.

Current research in Berlin aims at developing grammatical descriptions of this kind. There are three fundamental steps in the construction of such a grammar:

- the compilation of a corpus of language data
- the construction of a (descriptive) didactic grammar from this corpus
- the derivation of (prescriptive) pedagogical grammars from the didactic grammar

The validity of the approach, which relies on machine-readable corpora and computational methods of analysis, is exemplified by recent findings in three areas of English grammar.

(1) The teaching of ***any***
A study of *any* has shown that its most frequent use is generally not taught in German schools. The new findings suggest a different gradation and presentation of the functions of *any*.

(2) The description of ***will*** and ***would***
Traditionally *will* and *would* are treated in the same chapter of a grammar. *Would* is generally considered as the past tense form of *will*. The joint description of *will* and *would* does not take into account that the two modal verbs have significantly different semantic profiles.

(3) The treatment of irregular verbs
Irregular verbs are normally presented in alphabetical order. A recently developed learning list of irregular verbs presents the verbs in an order that reflects their importance in actual language use.

David Minugh
Dept of English, University of Stockholm, Sweden

All the Language that's Fit to Print: Using British and American Newspaper CDs as Corpora

Two interconnected technological advances — the increasing availability of English-language newspapers on CD-ROM, and the advent of CD-ROM readers as standard PC equipment — now provide us with a relatively sophisticated tool for the exploration of a number of areas in contemporary vocabulary and syntax studies. The simplicity and accessibility of this procedure open up corpus studies for the first time to language students at relatively elementary levels, without major investments in hardware or long apprenticeships in computer software management, while simultaneously providing opportunities for language investigation that will reward even the most advanced student. For the EFL teacher and student, it provides a wealth of information on collocation and usage far beyond what handbooks can provide.

Harry Papageorgiou, Lambros Cranias and Stelios Piperidis
Institute for Language and Speech Processing (ILSP), Athens, Greece

A New Approach to Automatic Sentence Alignment (poster)

The automatic sentence alignment techniques reported so far in the literature are purely statistical and produce quite reasonable results. The contributions towards the error rate, however, are mainly due to ill-formed alignments (two to one sentence, one to none...). Clearly, to deal with these cases, more linguistic knowledge should be incorporated in the alignment technique.

Based on the fact that the principal requirement in any translation exercise is meaning preservation across the languages of the translation pair, we extract from the corpus those phrase elements that mainly participate in any meaning preservation attempt. To this end, the corpus is POS tagged and patterns of tags of content words are extracted. Content words are taken to be verbs, nouns, adjectives and adverbs of the languages of the pair. Morphological ambiguity problems pertaining to content words are treated by constructing ambiguity classes leading to a generalised set of content tags. The complexity of transfer in translation imposes the consideration of the number in which content tags appear in a language pattern; thus the morphological derivation procedures observed across languages are taken into account.

Two probabilistic models are presented which define a distance measure between an SL and a TL sentence. This probabilistic distance is being

applied in the framework of a Dynamic Programming scheme which deals with already aligned paragraphs of text (based on paragraph markers) and produces the optimum alignment of sentences within these paragraphs.

The first model uses Multiple Linear Regression on a training corpus to estimate the sum of content tags appearing in the TL sentence based upon the contents of the different content tags in the SL sentence. The error of estimation is normally distributed producing a measure of distance between the SL and TL sentence.

The second model treats the count of different content tags as entries of a vector and calculates the multivariate probability density function of these vectors based on a training corpus. Then, the distance between two sentences is calculated as the probability that this pair of sentences emanates from the existing distribution.

The results indicate that while the already high success rate in dealing with one to one sentence alignments is further boosted, the technique can also effectively deal with ill-formed alignments. We are currently working on the adaptation of this technique to the problem of automatic subsentence alignment

---

Hans Paulussen and Guy Deville
ELV, University of Namur, Belgium

Automating Vocabulary Teaching Material: A Step by Step Experience

Since the introduction of the computer into educational environments, a considerable amount of information has been spread via the electronic medium, but not always in a very efficient way. Word processors are still too often used as typewriters and the reusability of texts and formats is not always evident. Aware of the need to exploit the electronic information efficiently, our foreign language teaching department tried at a very early stage to prepare all didactic material in electronic form. One of the first steps consisted in entering original foreign language text material and the related contextualised vocabulary in a very strict yet easy to manipulate format. The input phase evolved from the simple typewriter philosophy to the use of efficient and powerful word processing facilities, where style sheets assume an important role. This consistent policy has enabled us to produce complementary material, such as lab exercises and specialised glossaries, using simple electronic editing facilities. Thanks to our adhering to a formal input protocol during the preliminary stages, we have compiled a uniform corpus of general and specialised texts which is now being converted into a text database. The development of such a database and the additional linguistic tools will serve the needs of both teachers and students. A grammatically and semantically tagged lexical database will produce some general statistics regarding the quantities and the kinds of words the students have to master, thus enabling us to improve text and vocabulary selection. At a later stage, we intend to develop a number of programs which can generate automatically a battery of vocabulary exercises and tests. This paper gives a step by step overview of how we structured our corpora and a description of the linguistic database.

Pam Peters
School of English and Linguistics, Macquarie University, Australia

Micro- and Macro-linguistics

A new course in corpus-based linguistics for final year students at Macquarie University is designed to complement those of the previous two years (which are conceived in terms of either structural or systemic-functional approaches to language), and to demonstrate the testing of theoretically based hypotheses on corpus data. Using a digitised database of Australian speech, questions about the nature of the Australian accent can be examined, and questions about how (written) Australian English relates to British and American English can be explored through parallel databases (Brown, LOB and the Australian ACE). Questions about the lexical and grammatical correlates of particular genres are also on the agenda. The variety of corpora, reflected in their varying size, constituency and extent of annotation is discussed, as a basis for maximising their use -- while recognising their limitations. Ultimately the course highlights the nature of linguistic inquiry, but it also shows the numerous applications of corpora in speech technology and in the writing of grammars and dictionaries.

David Quinn[a] **and Akiva Quinn**[b]
[a]Department of Business Computing, City University, London, UK
[b]Survey of English Usage, University College London, UK[3]

Linguistic Modelling for a Corpus-Based CALL System

The examples and annotations in a corpus may be exploited in computer assisted language learning (CALL). Our CORpus-based Teaching EXpert (CORTEX) design and prototype, overviewed elsewhere, includes a model of this linguistic information, using the International Corpus of English. Here we further detail and extend the model.

In contrast to grammatical annotations, we examine the information-rich word level annotations (tags). Tags categorise words according to shared morphological, syntactic, or semantic properties. The linguistic model is built around a hierarchy of word classes, some corresponding to tags, others generalised from more specific classes. This hierarchy represents the relations implicit in the tags' component features.

Morphological classes (for example, singular nouns) are linked to corresponding word classes, using morphological information in the tags. Morphological rules are used to generate the forms of a lemma (word), and to characterise morphological classes not identified by the tags (for example case, for pronouns).

Syntactic information, such as verb transitivity, is also available from the word classes. Simple syntactic rules (involving word adjacency) are added,

[3]Now at Neuronics, 2nd Floor, 180 Albert Road, South Melbourne, Victoria 3205, Australia

where they complement word level phenomena (for example, periphrastic comparison, using 'more' and 'most').

The student is shown examples containing words from particular classes. These may be: A single class (such as intransitive verbs); a class based on hierarchical position (for example, comparative word forms, or adverbial comparative forms only); related classes for contrast (such as the various transitivity classes); or a specific lemma (for example, the inflections of a specific verb).

The linguistic model describes a structured information space. Topics, such as adjectival comparison (inflectional and periphrastic), define coherent sub-spaces suitable for exploration learning, or for use in an ordered teaching plan.

We consider linguistic modelling an essential component of corpus-based CALL systems, for teaching and exploration learning.

Francine Roussel
Faculté des Lettres, Nancy II, France

Using Mono- and Bilingual Concordances to Foster the Induction of some Phonological Rules of L2 (poster)

Some phonological features of English are not “heard” by most French speakers because they do not have their counterpart in French. This is the case, for example, of word stress patterns linked to suffixes; and of contrastive tonic placement on auxiliaries and on other form words.

Exposure to the spoken language is in such cases of little avail, especially as occurrences of these features in conversation are usually too few and far between to facilitate the induction of underlying rules. Teaching these rules in a traditional way is ineffective as well because:

a) the learner is in a passive and boring role
b) (s)he is presented with decontextualised lists of words or utterances;

As a result, little transfer takes place in the real situation.

Materials based on mono- and bilingual concordances have therefore been developed to help students of English come to terms respectively with word stress and contrastive tonic placement rules. These materials, in the form of game-like activities, challenge the learners to find these rules for themselves from the observation of authentic, contextualised data.

These materials are part of a double-entry volume to be published by the University of Nancy (France). According to their respective learning strategies, the learners will either start by working on the script of short one-act plays illustrating given phonological points, then listen to the accompanying tapes and report to the concordance-based exercises whenever necessary, or they will do these activities in the reverse order.

Francine Roussel
Faculté des Lettres, Nancy II, France

The Lingua Parallel Concordancing Project (poster)

The project, which started in January 1994 and is due to extend over a maximum period of two years, aims at

a) developing a parallel concordancer (first for the PC, then for the Macintosh)
b) collecting parallel corpora of well-translated texts in six (eventually eight or nine) European languages and corresponding to a range of genres and linguistic levels
c) producing teaching materials, whether in electronic or paper form, based on parallel concordances obtained from this data

The idea is to offer a useful CALL and CAT tool, accessible from within a word processor so as to throw light on obscure language points as one is writing a text, or producing a translation, in a foreign language. Blanking and other facilities will also enable it to produce pedagogical materials, at any level, with little effort on the part of the teacher.

The University of Nancy II (France) is co-ordinator of the project. The development of the alignment program and of the user interface is the responsibility of the Centre de Recherche en Informatique de Nancy, with the valuable contribution of Tim JOHNS and David WOOLLS of the University of Birmingham. The data entry and insertion of cross-references in the parallel versions of texts are done by the Universities of Nancy, Birmingham, Hull, East Anglia, Turin (Italy), Wuppertal (Germany), Aarhus (Denmark) and Patras (Greece).

John Sinclair
School of English Corpus Linguistics, University of Birmingham, UK

Rumours about English

What happens when the information about a language that we get from a corpus does not match with the information that is part of the language teaching consensus? What are the implications for teachers? for learners? for writers and publishers?

This paper will start with some observations about fairly common, but largely unrecognised, patterns of everyday English, and explore the implications for language teaching.

Elena Tognini-Bonelli
Department of Italian, University of Birmingham, UK

Corpus Work and Linguistic Issues: The Corpus of Modern Italian at the University of Birmingham

The Department of Italian at the University of Birmingham has recently acquired a 3.5 million corpus of modern Italian Language from the Istituto di Linguistica Computazionale in Pisa. This forms the basis of a course for IVth year students aiming to explore linguistic issues related to the Italian Language and stemming from the direct and guided exposure to concordances.

This paper will briefly outline the issues raised by corpus work in language theory: from the definition of a unit of meaning, to querying the boundary between semantics and pragmatics and bringing ever closer the distinction between "*langue*" and "*parole*". Observations on syntactic patterning and collocation will point to different cases of de-lexicalisation and idiomaticity as well as at more general issues of gender and culture.

Examples from both the Italian and the English Corpora (COBUILD) will be shown and discussed.

Eve Wilson
Computing Laboratory, The University of Kent at Canterbury, UK

Corpora as a Source of CALL Exercises

One of the most onerous tasks faced by teachers of English as a Foreign Language is the compilation of original exercises to suit all levels of ability from diverse cultural backgrounds. The availability of electronic corpora can help us to provide a solution to this problem by facilitating the automatic generation of exercises for Computer Assisted Language Learning (CALL) Packages.

Linguistic exercises for students can be divided into three broad categories:

lexical,
syntactical, and
discourse.

Lexical exercises are concerned with the formation and precise use of words: such exercises require an authoritative source which lists attributes such as part of speech, irregular plurals, derived forms, gives example sentences for all the different senses of a word, and explains idiomatic usage.

Syntactical exercises are concerned with simple sentence structure: the use of articles and pronouns, the agreement of subject and verb, tense mood and voice in the use of verb, the use of prepositions in clauses of time, place and manner, prepositions as verb particles. Source material for this kind of

exercise should ideally be parsed sentences i.e. sentences which have been annotated to show the syntactic function of each word within the sentence.

Discourse exercises are concerned with clause complexes i.e. passages of text longer than a simple sentence. The devices needed to ensure coherent and cohesive discourse are subtle and some are not yet easily amenable to automatic analysis by program. However, others, such as the use of conjunctions and simple linking frameworks can be retrieved from suitable text sources and used in exercise generation.

This paper considers methods of retrieving suitable examples from Corpora such as dictionaries, in particular the Oxford Advanced Learner's Dictionary, annotated corpora, such as the Susanne Corpus, a particularly well annotated subset of the Brown Corpus, and other free text sources without annotation such as literature and technical papers.

Martin Wynne
Department of Linguistics, Leeds University

A Unix Course for Corpus Linguists (poster)

For the past three years I have taught 'Introduction to Unix' to MA Linguistics and Information Processing and MA Speech Sciences students. The aim of the course is twofold: (i) to teach the students to be competent general purpose users of Unix (familiar with the principles of file hierarchies, file access, use of commands, input and output, *email*, interactive *C shell* features and *Bourne shell* programming), and (ii) to teach them to be able to develop their own applications for processing machine readable corpora of language data, using standard Unix tools combined in *Bourne shell* scripts. For the latter purpose they are introduced to commands such as *grep*, *sed*, *cut,* etc.

In order to get to the advanced appreciation of regular expressions that is necessary for developing language processing applications, a thorough introduction to the *ex* line editor is included in the course. *Awk*, *perl* and *C shell* programming are not covered.

The aim of the course is to enable to students to be capable of writing their own simple applications without needing to know a programming language. Students have by the end of the course been able to write projects in which they carry out studies of phonological data and extract lexical and syntactic information from machine readable corpora.

Xu Xunfeng and Michael Oakes
Department of Linguistics and Modern English Language, Lancaster University

The Hua Xia Corpus of Written Chinese and Concordancing Facilities (poster)

While there are a number of well-established English computer text corpora and concordancing facilities available for use in teaching and

research, Chinese corpora are still not easy to find and analysis utilities for Chinese corpora are even fewer. Since the launch of a Chinese studies programme at Lancaster University, the collection and analysis of a corpus of 2,000,000 Chinese characters have been initiated. The raw corpus is composed of nearly two hundred weekly issues of Hua Xia Wen Zhai (CND-CM), an electronic magazine published by the Chinese overseas community since 1991. All the texts are in GB binary code and the preliminary annotation is carried out with a Chinese word processing package called *NJStar*. We have produced concordancing facilities for the analysis of the corpus to run on the IBM-PC. The following information will be presented:

1. Representation of Chinese characters by GB binary code.
2. Excerpts from the Hua Xia Corpus of Written Chinese.
3. Excerpts from the KWIC concordancing of the Chinese texts.
4. Architecture of our concordancing system.
5. Applicability of the concordancing facilities for the teaching and analysis of the Chinese language.

443 主席的葬礼结束了，柯西金飞到了 中国 。因为，他的图１０４型客机必须飞
415 月九日，基辛格非常秘密地飞到了 中国 。 第一天晚上，叶剑英和黄
417 黄华，似乎不着边际地说了句：“ 中国 菜真让我肠胃功能紊乱。”
174 ，即使是在当前商品意识尚薄弱的 中国 大陆，也能有幸被列入独创之举了！
38 。 ★ 一九九四年一月一日， 中国 大陆报刊杂志大多涨价将近一倍。例
34 据国家信息中心预测，一九九四年 中国 大陆的人口将超过十二亿。 ★
42 一律涨价５２%。 ★ 美国驻 中国 大使芮孝俭在接受《纽约时报》专访
429 道的美国和西德货，但烹制方法是 中国 的，中西合璧。博士先生一定会感兴
227 应有的收入；我们学会了服从，在 中国 的土地上服从着一些并不适合中国国
469 他们的宣传品中有很大的篇幅介绍 中国 的文化大革命，仔细看看，他们竟然

Figure: *Sample of the concordanced text for the two-character sequence (Zhongguo) sorted to the right.*

Federico Zanettin
Scuola Superiore di Lingua Moderne per Interpreti e Traduttori, University of Bologna, Italy

Retrieving and Editing Texts in Different Languages: Corpus-based Activities for Language Learning

In curricula involving the contrastive study of languages (i.e. courses for translators and interpreters) it may be particularly useful to work with parallel corpora. Newspaper corpora in different languages are now easily available on CD-ROM (for example, the Independent, Le Monde, Il Sole 24 Ore) and can be used as a source for the automated retrieval of texts for classroom and self-access activities involving gap filling, reassembling and rewriting, which help students develop learning skills like predicting, guessing from context, and using coherence and cohesion devices.

Newspapers supply instances of actual language use and show recurring patterns which learners can use to observe regularities of the written language and to distinguish among genres in newspapers, drawing on their knowledge of what conventions are at work in a newspaper text.

This paper deals with the use of retrieval and word processing software as a means for customising texts taken from newspaper corpora. Starting from a main corpus, different sub-corpora are obtained by selecting items and downloading them into PC readable files. Desired codification (genre, topic, newspaper section, sequence, place, author) is inserted by editing texts through a word processor, and these tagged mini-corpora are subsequently converted to database format for use with TACT or DBT. Sub-corpora are created according to time span (e.g. all the items published in one week), keyword (e.g. all the items containing the word "immigrant/s"), or a combination of the two, thus becoming more manageable and providing a quick access to large amounts of texts possessing common characteristics. Selection procedures employed to generate activities from database files can then be used with different sub-corpora by both students and teachers.

Activities can be created which guide the students to compare selected items (i.e. concordance of particular words, texts on similar topics, texts of the same genre) from corpora in two or three languages, providing them with a source of linguistic information of a contrastive kind. In this sense computer processing of plurilingual corpora may reveal itself a useful tool not only for foreign language learning but also for assisting professional translators and interpreters.

A GUIDE TO THE TALC94 WORKSHOPS

(all workshops commence at 2.15 p.m. on Tuesday 12 April)

1. **Demonstration/Tutorial Workshop: Using the ICAME CD-ROM**. This will be an intermediate level workshop for those with an interest in the ICAME CD-ROM, covering its use in PC, Mac and UNIX environments, and dealing with the broad range of data and software items contained on the disk. The workshop will be led by **Knut Hofland**, of the Norwegian Computing Centre for the Humanities, where ICAME is based. This workshop will run for about 2½ hours, with a refreshment break.
   PROVISIONAL VENUE: Lonsdale Large Lecture Theatre

   Note: all other workshops are scheduled for 90 minutes, followed by refreshments.

2. **Practical Workshop: Elementary Concordancing for Teaching Purposes**. Aimed at relative newcomers to the field, this workshop will introduce people to the use of two of the more accessible PC-based concordance programs, Longman Mini Concordancer and OUP's MicroConcord. There will also be an introductory survey of other concordancing software available to the teaching community. This workshop will be led by **Michael Barlow** of Rice University, Houston.
   PROVISIONAL VENUE: Linguistics Postgraduate Computing Lab, B71

3. **Discussion Workshop: Methodological Issues**. Aimed at people with some experience of using corpora in teaching situations, this workshop is intended to stimulate exchange of experience and ideas for effective teaching methods. The workshop will be run by **Tim Johns**, Birmingham University. On arrival at the conference, people wishing to take part in this workshop will be asked to list the topics they would most like to discuss. This feedback will be used in the final planning of the event.
   PROVISIONAL VENUE: Linguistics Department "Boardroom", A68

4. **Discussion Workshop: Multilingual Corpora**. Aimed at people with varying degrees of experience, this workshop will encourage discussion of the situation regarding the availability of multi-lingual corpora, and likely trends in their development, as well as examining ways in which they may be used for teaching purposes. **Roger Garside** and **Tony McEnery**, University of Lancaster , and **Jean-Marc Langé**, IBM Scientific Center, France, will lead this workshop.
   PROVISIONAL VENUE: Linguistics Postgrad. Teaching Room, B80

5. **Tutorial/Discussion Workshop: Creating a Corpus**. Aimed primarily at people who feel they are "on the brink" of setting out to create a corpus, the workshop will also benefit from and, we hope, be of interest to, people with some experience and expertise in this area. This workshop will be led by **Hans Paulussen**, University of Namur, Belgium. As with Workshop 3, participants will be invited to give advance notice of the topics they would like to cover in the workshop.
   PROVISIONAL VENUE: UCREL Laboratory, Room B103, Linguistics Department

Dr. Khurshid Ahmad (Senior Lecturer) The Artificial Intelligence Group Department of Mathematical and Computing Sciences University of Surrey GB-Guildford GU2 5XH **k.ahmad@mcs.surrey.ac.uk**	J. C. Alderson (Professor) Department of Linguistics & Modern English Language Bowland College GB-Lancaster LA1 4YT **eia019@lancaster.ac.uk**	Quentin Allan (Postgraduate) CELS - School of English (Westmere) University of Birmingham Edgbaston GB-Birmingham B15 2TT **allanqg@birmingham.ac.uk**
Joan Allwright (Tutor in Academic Support Programme) Institute for English Language Education George Fox Building Lancaster University GB-Lancaster LA1 4YJ	Poul Anderson (Project Officer, European Commission) DG XIII-E-5 - Jean Monet Building B4/114 Plateau de Kirchberg L-2920 Luxembourg	Guy Aston (Professor) Scuola Superiore di Lingue Moderne per Interpreti e Traduttori University of Bologna Corso della Repubblica 136 I-47100 Forlì Italy **vk1a@icineca.earn**
Michael Barlow (Visiting Assistant Professor) Department of Linguistcs Rice University PO Box 1892 Houston TX 77251 USA **barlow@crl.ucsd.edu**	Vladimir Benko (Research Worker) Computational Linguistics Laboratory Comenius University Faculty of Education Moskovska 3 SK-81334 Bratislava Slovakia **jazybenk@savba.savba.sk**	Gunnar Bergh (Lecturer/Researcher) Department of English University of Göteborg Renströmsparken S-41298 Göteborg Sweden **bergh@eng.gu.se**
Mrs. Susan Blake (Lecturer in German) S L E S University of Wolverhampton Stafford Street GB-Wolverhampton WV1 1SB **le1963@ccub.wlv.ac.uk**	Philip Bolt (Senior Lecturer) Department of English Hong Kong Polytechnic Hung Hom Kowloon Hong Kong **egphbolt@hkpcc.hkp.hk**	Simon P. Botley (Research Assistant) Longman Dictionaries Longman House Burnt Mill GB-Harlow CM20 2JE
Gavin Burnage (British National Corpus) Oxford University Computing Services 13 Banbury Road GB-Oxford OX2 6NN **gavin@natcorp.ox.ac.uk**	Lou Burnard Oxford University Computing Services 13 Banbury Road GB-Oxford OX2 6NN **lou@vax.ox.ac.uk**	Dr. Jill C. Burstein (Associate Development Scientist) Educational Testing Service Rosedale Road 11-R Princeton NJ 08541 USA **jburstein@rosedale.org**

Peter Christian (Documentation and Training Officer) Computing Service Goldsmith's College New Cross GB-London SE14 6NW **peter@uk.ac.gold**	Lambros Cranias (Researcher) Institute for Language and Speech Processing 22 Margari Str. 115 25 Athens Greece **stelios.piperidis@eurokom.ie**	Jonathon Culpeper (Teaching Fellow) Department of Linguistics & Modern English Language Bowland College Lancaster University GB-Lancaster LA1 4YT **eia039@lancaster.ac.uk**
Andrea E. Davies (Research Officer) The Artificial Intelligence Group Department of Mathematical and Computing Sciences University of Surrey GB-Guildford GU2 5XH **a.davies@mcs.surrey.ac.uk**	Dr. William J. Dodd (Lecturer) Department of German Studies University of Birmingham Edgbaston GB-Birmingham B15 2TT **w.j.dodd@bham.ac.uk**	Jürgen Esser (Professor) Institut für Anglistik RWTH Aachen Karmanstraße 17/19 D-52062 Aachen Germany **juergen.esser@rwth-aachen.de**
Liz Eyes (Research Associate) Unit for Computer Research on the English Language SECAMS University of Lancaster GB-Lancaster LA1 **eia025@lancaster.ac.uk**	Steve Fligelstone (Research Associate) Department of Linguistics and Modern English Language Bowland College University of Lancaster GB-Lancaster LA1 4YT **eia002@lancaster.ac.uk**	Gwyneth Fox (Editorial Director) Cobuild University of Birmingham Westmere 50 Edgbaston Road GB-Birmingham B15 2RX **gwyneth@cobuild.collins.co.uk**
Roger Garside (Senior Lecturer) Computing Department SECAMS University of Lancaster GB-Lancaster LA1 **rgg@lancs.comp**	Laura Gavioli (Researcher) Scuola Superiore di Lingue Moderne per Interpreti e Traduttori University of Bologna Corso della Repubblica 136 I-47100 Forlì Italy **gavioli@alma03.cineca.it**	David Graddol (Lecturer) School of Education Open University Walton Hall GB-Milton Keynes MK7 6AA **d.j.graddol@open.ac.uk**
Louise Grainger (Product Manager - COBUILD) International Division HarperCollins Publishers 77-85 Fulham Palace Road Hammersmith GB-London W6 8JB	Christopher Greaves (Senior Lecturer) Department of Communication and Languages Luleå University S 941 87 Luleå Sweden **chgr@lh.luth.se**	Hans van Halteren (Researcher) Department of Language and Speech University of Nijmegen P.O. Box 9103 NL-6500 HD Nijmegen The Netherlands **hvh@lett.kun.nl**

Dr. Noel John Heather (Lecturer in Arts Computing) Department of English Royal Holloway and Bedford New College University of London Egham Hill GB-Egham TW20 OEX **n.heather@uk.ac.rhbnc**	Mary Hodges (Research Assistant) BNC Laboratory, A15 SECAMS University of Lancaster GB-Lancaster LA1 **eia020@lancaster.ac.uk**	Knut Hofland (Senior Executive Officer) Norwegian Computing Centre for the Humanities University of Bergen Harald Haarfagres gt. 31 N-5007 Bergen Norway **knut.hofland@hd.uib.no**
Glyn Holmes (Associate Professor) Department of French The University of Western Ontario London Ontario N6A 3K7 Canada **gholmes@uwovax.uwo.ca**	Graeme Hughes (Systems Support Manager) Department of Linguistics and Modern English Language Bowland College University of Lancaster GB-Lancaster LA1 4YT **eia028@lancaster.ac.uk**	Irmgard Hüppe (Educational Software Consultant) 8 Church Way Iffley GB-Oxford OX4 4DY
Gordon Inkster (Lecturer) French Studies Department of Modern Languages University of Lancaster GB-Lancaster LA1 4YN **mla002@lancaster.ac.uk**	Dr. Gregory C. A. James (Director) Language Centre The Hong Kong University of Science and Technology Clearwater Bay Kowloon Hong Kong **lcgjames@usthk.bitnet**	Tim Johns (Lecturer) English for Overseas Students Unit University of Birmingham Edgbaston GB-Birmingham B15 2TT
Randall L. Jones (Professor of German, Dean of Humanities) College of Humanities 2054 JKHB Brigham Young University Provo Utah 84602 USA **jonesr@byu.edu**	Yasuhiko Kato (Researcher) The National Language Research Institute Kita-Ku, Nishigaoka 3-9-14 Tokyo 115 Japan **kateaux@tansei.cc.u-tokyo.ac.jp**	Bernhard Kettemann (Professor) Department of English University of Graz A-8010 Graz Austria **kettemann@bkfug.kfunigraz.ac.at**
Adam Kilgarriff (Computational Linguist) Longman Dictionaries Longman House Burnt Mill Harlow Essex CM20 2JE **100347.76@compuserve.com**	Mary Kinane Unit for Computer Research on the English Language SECAMS University of Lancaster GB-Lancaster LA1 **eia041@lancaster.ac.uk**	Philip King (Lecturer) English for Overseas Students Unit University of Birmingham Edgbaston GB-Birmingham B15 2TT **p.b.king@birmingham.ac.uk**

John M. Kirk (Lecturer) School of English The Queen's University of Belfast GB-Belfast BT7 1NN **eng0997@qub.ac.uk**	Dr. Kenji Kita (Associate Professor) Department of Information Science Faculty of Engineering Tokushima University 2-1 Minamijosanjima Tokushima 770 Japan **kita@is.tokushima-u.ac.jp**	Göran Kjellmer (Senior Lecturer) Department of English University of Göteborg S-41298 Göteborg Sweden **kjellmer@eng.gu.se**
Gerry Knowles (Senior Lecturer) Department of Linguistics and Modern English Language Bowland College University of Lancaster GB-Lancaster LA1 4YT **eia008@lancaster.ac.uk**	Jean-Marc Langé IBM France Scientific Center 54, rue R. Salengro F-94126 Fontenay-sous-Bois CEDEX France **jml@vnet.ibm.com**	Geoffrey Leech (Professor) Department of Linguistics and Modern English Language University of Lancaster GB-Lancaster LA1 4YT **eia014@lancaster.ac.uk**
Simon Lippmann (Postgraduate, University of Kent) 211 Wincheap GB-Canterbury CT1 3TP **sfl@ukc.ac.uk**	Paul Livesey (Lecturer) Department of Cultural Studies University of Central Lancashire GB-Preston PR1 2HE **livesey_pm@p1.uclan.ac.uk**	W. E. Louw (Senior Lecturer) Department of English University of Zimbabwe P.O. Box MP 167 Mount Pleasant Harare Zimbabwe **blouw@zimbix.uz.zw**
Hazel Marshall (Senior Lecturer) College of St. Mark and St. John Derriford Rd. GB-Plymouth PL6 8BH	Bill Mascull 4 Rue Froment F-75011 Paris France	Tony McEnery (Lecturer) Department of Linguistics and Modern English Language Bowland College University of Lancaster GB-Lancaster LA1 4YT **mcenery@lancs.comp**
John Milton Language Centre The Hong Kong University of Science and Technology Clearwater Bay Kowloon Hong Kong **lcjohn@usthk.ust.hk**	Dieter Mindt (Professor) Freie Universität Berlin Zentralinstitut für Fachdidaktiken Habelschwerdter Allee 45 D -14195 Berlin Germany **mindt@engdidakt.fu-berlin.de**	Ilka Mindt (Student, University of Würzburg) Ilsensteinweg 54 D-14129 Berlin Germany **mindt@vax.rz.uni-wuerzburg.d400.de**

David Minugh (Foreign Lecturer) Department of English University of Stockholm S-106 91 Stockholm Sweden **minughd@engelska.su.se**	Simon Murison-Bowie (Director, Electronic Publishing & Development) Oxford University Press Walton Street GB-Oxford OX2 6DP **murisons@oup.co.uk**	Michael Oakes (Research Associate) Department of Linguistics and Modern English Language University of Lancaster GB-Lancaster LA1 4YT **eia038@lancaster.ac.uk**
Takashi Omoto (Graduate Student) Faculty of Engineering Tokushima University 2-1 Minamijosanjima Tokushima 770 Japan **omoto@n88.is.tokushima-u.ac.jp**	Agoritsa Papaghatzi (Student) Block 1, Flat 7 Ashcroft Pritchatts Rd., Edgbaston Birmingham B15 2QU	Hans Paulussen (Lecturer) ELV Université de Namur Rue de Bruxelles 61 B-5000 Namur Belgium **hpaulus@cc.fundp.ac.be**
Phil Payne (Professor) German Studies Department of Modern Languages University of Lancaster GB-Lancaster LA1 4YN	Pamela Peters (Senior Lecturer) School of English and Linguistics Macquarie University New South Wales 2109 Australia **ppeters@srsuna.shlrc.mq.edu.au**	Luis Quereda (Senior Lecturer) Departamento de Filología Inglesa Facultad de Filosofía y Letras Campus Universitario de Cartuja s/n 18071 Granada Spain **lquereda@ugr.es**
David Quinn (Research Assistant) 55 Normandy Road GB-St. Albans AL3 5PR **d.d.quinn@city.ac.uk**	Paul Rayson (Research Associate) Unit for Computer Research on the English Language SECAMS University of Lancaster GB-Lancaster LA1 **paul@comp.lancs.ac.uk**	Dr. Daniel Robertson (Research Fellow) Department of Applied Linguistics, University of Edinburgh, 14 Buccleuch Place, Edinburgh EH8 9LN. **Daniel.Robertson@edinburgh.ac.uk**
Peter Robinson (Lecturer) English Language Unit Cornwallis West Univeristy of Kent GB-Canterbury CT2 7NF **par1@ukc.ac.uk**	Francine Roussel (Senior Lecturer, University of Nancy) Faculté des Lettres 23 Bd Albert 1 F-54000 Nancy France	Dr. Raphael Salkie (Principal Lecturer) Language Centre University of Brighton Falmer GB-Brighton BN1 9PH **rms3@vms.brighton.ac.uk**

Juan Santana (Lecturer) Departamento de Filología Inglesa Facultad de Filosofía y Letras Campus Universitario de Cartuja s/n 18071 Granada Spain **jsantana@ugr.es**	Bob Seatter (Postgraduate) SEMLS University of Kent GB-Canterbury CT2 **rds6@ukc.ac.uk**	Elena Semino (Lecturer) Department of Linguistics & Modern English Language Bowland College Lancaster University GB-Lancaster LA1 4YT **eia007@lancaster.ac.uk**
John M. Sinclair (Professor of Modern English Language) School of English — Corpus Linguistics University of Birmingham Westmere Mews Edgbaston GB-Birmingham B15 2TT	Nick Smith (Research Assistant) Unit for Computer Research on the English Language Computing Department SECAMS University of Lancaster GB-Lancaster LA1 **nick@comp.lancs.ac.uk**	Anna-Brita Stenström (Professor) Department of English University of Bergen HF-bygget, Sydnesplass 9 N-5007 Bergen Norway **stenstroem@hf.uib.no**
Izumi Tanaka (Research Student) Department of Linguistics and Modern English Language Bowland College University of Lancaster GB-Lancaster LA1 4YT **eib050@lancaster.ac.uk**	Dr. Jenny Thomas (Lecturer) Department of Linguistics and Modern English Language Bowland College University of Lancaster GB-Lancaster LA1 4YT **eia013@lancaster.ac.uk**	Jennifer Thurstun (Associate Lecturer) NCELTR Macquarie University NSW 2109 Australia **jthurstu@ocs1.ocs.mq.edu.au**
Elena Tognini-Bonelli (Lecturer) School of Modern Languages University of Birmingham Edgbaston GB-Birmingham B15 2TT **togninie@birmingham.ac.uk**	Alexander H. Urquhart (Senior Lecturer) The College of St. Mark and St. John Derriford Rd. GB-Plymouth PL6 8 BH	Dr. Tamás Váradi (Lecturer, School of Slavonic and East European Studies) 237 Mitcham Lane Streatham GB-London SW16 6PY **eia029@lancaster.ac.uk**
Martin Warren (Director of Studies) Department of Communication and Languages Luleå University S 941 87 Luleå Sweden **martin.warren@lh.luth.se**	Anne Wichmann (Senior Lecturer) Department of Cultural studies University of Central Lancashire GB-Preston PR1 2HE **wichmann_a@p1.uclan.ac.uk**	Jane Willis (Teaching Fellow, Aston University) 21 Oakfield Rd., GB-Birmingham B29 7HH **j.r.willis@aston.ac.uk**

J.D.Willis (Lecturer) University of Birmigham Westmere 50 Edgbaston Park Road Edgbaston GB-Birmingham B15 2TT	Eve Wilson (Senior Lecturer) Computing Laboratory The University Canterbury GB-Kent CT2 7NF **ew@ukc.ac.uk**	David Woolls (Researcher) School of English University of Birmingham Edgbaston GB-Birmingham B15 2TT
Martin Wynne (Research Student, University of Leeds) 5 Woodland Grove GB-Leeds 7 **eib024@lancaster.ac.uk**	Xu Xunfeng (Research Student) Department of Linguistics and Modern English Language Bowland College University of Lancaster GB-Lancaster LA1 4YT **eib035@lancaster.ac.uk**	Colin Yallop (Associate Professor) School of English & Linguistics Macquarie University NSW 2109 Australia **cyallop@pip.engl.mq.edu.au**
Mari Yanai (Postgraduate) 4E Tudor Court Parkwood University of Kent GB-Canterbury CT2 7UX **my7@ukc.ac.uk**	Mei-shu Yu (Postgraduate, University of Kent, UK) 5D Bishopden Court Parkwood University of Kent GB-Canterbury CT2 7UY **msy@ukc.ac.uk**	Federico Zanettin (Researcher) Scuola Superiore di Lingue Moderne per Interpreti e Traduttori University of Bologna Corso della Repubblica 136 47100 Forlì Italy **zanettin@alma03.cineca.it**