Teaching and Language Corpora

Edited by
Anne Wichmann,
Steven Fligelstone,
Tony McEnery & Gerry Knowles

Applied Linguistics and Language Study

Teaching and Language Corpora

APPLIED LINGUISTICS AND LANGUAGE STUDY

GENERAL EDITOR

PROFESSOR CHRISTOPHER N. CANDLIN,

Macquarie University, Sydney

For a complete list of books in this series see pages v–vi

Teaching and Language Corpora

Edited by

Anne Wichmann
Steven Fligelstone
Tony McEnery
Gerry Knowles

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 1997 by Addison Wesley Longman Limited

Published 2013 by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN
52 Vandaerbilt Avenue, New York, NY 10017, USA

Routledge is an imprint of the Taylor & Francis Group, an informa business

ISBN 13: 978-0-582-27609-3 (pbk)

British Library Cataloguing-in-Publication Data
A catalogue record for this book is
available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog entry for this title is available from the
Library of Congress

APPLIED LINGUISTICS AND LANGUAGE STUDY

GENERAL EDITOR

PROFESSOR CHRISTOPHER N. CANDLIN,

Macquarie University, Sydney

Language and Development:
Teachers in a Changing World
BRIAN KENNY *and*
WILLIAM SAVAGE (EDS)

Autonomy and Independence in
Language Learning
PHIL BENSON *and* PETER VOLLER (EDS)

Literacy in Society
RUQAIYA HASAN *and*
GEOFFREY WILLIAMS (EDS)

Phonology in English Language
Teaching: An International Approach
MARTHA C. PENNINGTON

From Testing to Assessment:
English as an International Language
CLIFFORD HILL *and* KATE PARRY
(EDS)

Language as Discourse:
Perspectives for Language Teaching
MICHAEL MACCARTHY *and*
RONALD CARTER

Language and Discrimination:
A Study of Communication in
Multi-Ethnic Workplaces
CELIA ROBERTS, EVELYN DAVIES *and*
TOM JUPP

Translation and Translating:
Theory and Practice
ROBERT T. BELL

Language, Literature and the Learner:
Creative Classroom Practice
RONALD CARTER *and*
JOHN MCRAE (EDS)

Theory and Practice of Writing:
An Applied Linguistic Perspective
WILLIAM GRABE *and* ROBERT B. KAPLAN

Measuring Second Language
Performance
TIM MCNAMARA

Interaction in the Language
Curriculum: Awareness, Autonomy
and Authenticity
LEO VAN LIER

Second Language Learning:
Theoretical Foundations
MICHAEL SHARWOOD SMITH

Analysing Genre – Language Use in
Professional Settings
V.K. BHATIA

Rediscovering Interlanguage
LARRY SELINKER

Language Awareness in the Classroom
CARL JAMES *and* PETER GARRETT (EDS)

Process and Experience in the
Language Classroom
MICHAEL LEGUTKE *and*
HOWARD THOMAS

An Introduction to Second
Language Acquisition Research
DIANE LARSEN-FREEMAN *and*
MICHAEL H. LONG

Listening in Language Learning
MICHAEL ROST

The Classroom and the Language
Learner: Ethnography and Second-
Language Classroom Research
LEO VAN LIER

Second Language Grammar:
Learning and Teaching
WILLIAM E. RUTHERFORD

An Introduction to Discourse Analysis
Second Edition
MALCOLM COULTHARD

Learning to Write: First
Language/Second Language
AVIVA FREEDMAN, IAN PRINGLE
and JANICE YALDEN (EDS)

Stylistics and the Teaching of
Literature
HENRY WIDDOWSON

Listening to Spoken English
Second Edition
GILLIAN BROWN

Observation in the Language
Classroom
DICK ALLWRIGHT

Vocabulary and Language Teaching
RONALD CARTER *and*
MICHAEL MCCARTHY (EDS)

Bilingualism in Education: Aspects of
Theory, Research and Practice
MERRILL SWAIN *and* JIM CUMMINS

Reading in a Foreign Language
J. CHARLES ALDERSON *and*
A.H. URQUHART (EDS)

Language and Communication
JACK C. RICHARDS *and*
RICHARD W. SCHMIDT (EDS)

Error Analysis Perspectives on Second
Language Acquisition
JACK RICHARDS

Contents

This page intentionally left blank

Preface

Most of the chapters in this volume were first presented at the conference on Teaching and Language Corpora (TALC) at Lancaster University in the Spring of 1994. The idea for the conference, and hence the book, arose from discussions among members of ICAME (International Computer Archive of Modern English), and in particular as a result of a session initiated and led by Steve Fligelstone.

ICAME meets annually to report on research based on computer corpora of English, and more recently other languages. Many of the members are also teachers, and it became increasingly clear in recent years that they were not only engaged in research into corpora, but were using corpora, or corpus-derived data, to inform their teaching and to drive the learning process.

This was clearly by no means a secondary activity, and many of them responded to our call for articles dealing with the pedagogical aspect of corpus exploitation. We were delighted at the interest shown and hope that, by bringing together in one volume some of these contributions, many more teachers will find inspiration and encouragement to embark on similar activities.

The Editors

Editors' acknowledgements

This book reflects the innovative work and commitment of a large number of teachers. Our thanks go first to them, and then to the many researchers, teachers and computer specialists whose work has provided the foundations on which these chapters build. In particular we would like to thank the members of ICAME who provide a supportive framework in which new ideas can be conceived, developed and, above all, shared.

Publishers' acknowledgements

We are grateful to the following for permission to reproduce copyright material:

Faber & Faber Ltd and Farrar, Straus & Giroux, Inc for extracts from the poems 'Send No Money' from *The Whitsun Weddings* (UK Title)/*Collected Poems* (US Title) by Philip Larkin. Copyright © 1988, 1989 by the Estate of Philip Larkin, and 'Elvis Presley' from *The Sense of Movement* (UK Title)/*Collected Poems* (US Title) by Thom Gunn. Copyright © 1994 by Thom Gunn; Newspaper Publishing Plc for the article 'Restrictions on young drivers considered' by Christian Wolmar from *The Independent* 26.1.93 and an extract from the article 'For church and state it's divorce – Italian style' by Michael Sheridan from *The Independent* 2.10.89; A.P. Watt Ltd on behalf of Michael Yeats and Simon & Schuster for the poem 'Memory' from *The Collected Poems of W.B. Yeats* (UK Title)/*The Collected Works of W.B. Yeats, Volume 1: The Poems,* (US Title) revised and edited by Richard J. Finneran (New York, Macmillan 1989).

This page intentionally left blank

Contributors

KHURSHID AHMAD
University of Surrey

GUY ASTON
University of Bologna

ANDREA DAVIES
Formerly University of Surrey

WILLIAM DODD
University of Birmingham

LAURA GAVIOLI
University of Bologna

GRAEME HUGHES
Lancaster University

GORDON INKSTER
Lancaster University

HOWARD JACKSON
University of Central England in Birmingham

TIM JOHNS
University of Birmingham

RANDALL L. JONES
Brigham Young University

BERNHARD KETTEMANN
University of Graz

PHILIP KING
University of Birmingham

GEOFFREY LEECH
Lancaster University

WILLIAM LOUW
University of Zimbabwe

DIETER MINDT
Freie Universität Berlin

DAVID MINUGH
University of Stockholm

PAMELA PETERS
Macquarie University

ANTOINETTE RENOUF
University of Liverpool

JOHN M. SINCLAIR
University of Birmingham

EVE WILSON
University of Kent

General Introduction

Why this book?

Corpora have long been used in research. Since many researchers are also teachers, it is possible that corpora, or at least corpus-derived data, have also been used for some time in the teaching and learning process. So what has changed?

Firstly, computers have become smaller, cheaper and thus more widely available, both to teachers and to learners. The data stored on them has become more readily accessible to the user. Above all, the amount of data available from the outset in machine-readable form is far greater than even a few years ago. In this way the practical prerequisites for corpus-based teaching and learning have improved dramatically.

Secondly, the current approach to language description is strongly oriented towards usage. There is a general need to accommodate the apparent unpredictability of real data. This need stems in part from developments in speech and language technology. Systems, such as part-of-speech taggers, developed to work on written text, must be robust enough to deal with texts as they are, rather than texts as we might like them to be. In language teaching, too, the preference for 'authentic' texts requires both learners and teachers to cope with language which the textbooks do not predict. And of course the end product of language teaching, the ability to communicate, must ultimately take place in the real world, and not in a linguistically contrived one.

Finally, some of the impetus for this book may lie in changing views of the role of teachers and learners. Students are increasingly encouraged to take charge of their own learning. While this

arises no doubt in part from economic pressures, it also reflects serious pedagogical concerns.

The contributors to this book are only a small number of those teachers around the world who exploit computerized language data, whether written or spoken, for teaching. Many, but not all, deal with the teaching of English; others have applied similar methods to the teaching of other languages. Some describe the use of corpora for helping learners to acquire a foreign language, with the main emphasis on increasing their proficiency. Others are more concerned with teaching *about* individual languages and language in general. Together they highlight just how many areas of language teaching and learning can profit from the use of corpora.

Anne Wichmann

This page intentionally left blank

1

Teaching and Language Corpora: a Convergence

GEOFFREY LEECH

There is every reason to believe that language corpora will have a role of growing importance in teaching. This book, and the workshop (TALC94) which gave birth to it, are testimonies to the richness of the interest and experience which are already being applied to the convergence of language teaching and language research, through the link of corpus-based methods.

1 Up to now

Until recently, teaching had little connection with the momentum behind the evolution of corpus-based methods in linguistics. There were other forces in play. But one of the functions of the TALC94 workshop, as the first-ever international (or even national) conference on corpora and teaching, was that it enabled us to learn, for the first time, about the whole range of largely unpublicized pedagogical activities making use of corpora.

The experience of ICAME (International Computer Archive of Modern English) has been especially indicative. For seventeen years, ICAME, with its annual conference and journal,[1] has spearheaded research developments in corpus linguistics, with particular reference to the English language. But it was not until 1992 that there was an item in the ICAME conference programme referring to the use of corpora in teaching. This was the paper by Steve Fligelstone, 'Some reflections on the question of teaching from a corpus linguistics perspective'. Fligelstone led a workshop at the 1992 conference in Nijmegen on the topic, and the paper was reworked for publication (Fligelstone 1993).

Future historians tracing the history of computer corpora in linguistics might easily assume that ICAME members had given no serious thought to the educational use of corpora up to that time. But this would be a false conclusion. Most of the members of ICAME were then, and still are, university teachers, and most of them will have increasingly been using their corpora and corpus-based techniques in teaching, as well as in research, for a number of years. In my own case, for example, I began using an incomplete prototype LOB (Lancaster–Oslo/Bergen) Corpus for postgraduate teaching as early as 1976, and this use of corpora in teaching has continued, and gradually been extended to new areas of the curriculum at Lancaster, ever since. The original 'trickle down' from research to teaching is now becoming a torrent!

The notion of 'trickle down' from research to teaching seems particularly appropriate to corpus linguistics. This is because the computer corpus, as a resource for finding out about language and texts, is totally neutral as to these two major interconnecting activities of universities. The corpus, purely as a resource, is rather like a shelf in a university library: it is there to be exploited, and the same resources are equally usable for research and teaching. The history of computer corpora, on the other hand, has been tied to the history of computer technology. Inevitably, while computers were limited to large mainframes available to the initiated few, computer corpora were largely restricted to research use. But as computers have grown smaller, cheaper, and massively more powerful, their use in teaching has grown immeasurably. It is natural that the movement from research to teaching has taken place in this way, as the information revolution in the use of computers has more and more extended itself from the laboratory to the classroom.

It is also evident that the corpus, as an information source, fits in very well with a dominant trend in university teaching philosophy over the past twenty years, which is the trend from *teaching as imparting knowledge* to *teaching as mediated learning* (cf. Laurillard 1993: 13–15). In this context, there is no longer a gulf between research and teaching (cf. Knowles 1990), since the student is placed in a position similar to that of a researcher, investigating and imaginatively making sense of the data available through observation of the corpus. As Tim Johns has said (quoted by Gavioli in this volume), [*Georges Clemenceau* (1841–1929): 'War is much too serious a thing to be left to the military.'] 'research is too serious

to be left to the researchers': teaching is a natural extension of research. The student-centred paradigm of 'discovery learning' – or what Johns has called 'data-driven learning' – can scarcely be better exemplified than through the use of the computer corpus. Almost uniquely, among the information resources of which students make use in education, a text corpus is of primary interest because of what it is. Other resources, such as databases, are of interest indirectly, because of what they are 'about'. But a corpus is, of itself, a rich resource of authentic data containing structures, patterns and predictable features that are waiting to be 'unlocked' by the human intelligence. Perhaps the nearest equivalent, in other disciplines, is in the direct confrontation with data that occurs in the scientific laboratory, or in fieldwork. It is this experiential confrontation with the material of study that can make corpus work so rewarding for the student. And it often happens that a student working on a relatively small corpus assignment comes up with original observations and discoveries which have probably never been brought to notice before, even in the most detailed dictionaries and grammars of a language.

Having quoted Tim Johns, I should celebrate the nature of his contribution, as a prime example of a university teacher who has exploited the computer corpus mainly for teaching. Indeed, the above quotation 'research is too serious to be left to the researchers' reverses, by implication, the traditional donnish assumption that research or scholarship is the more important thing, and that teaching is just a spin-off from it, the vehicle whereby students are permitted to participate in the don's world of *recherché* knowledge. Tim Johns's earlier work in CALL (Computer-assisted Language Learning) – see Higgins and Johns (1984), Johns (1988) – naturally availed itself of the corpus-rich atmosphere of Birmingham, and he became among the first to advocate and to explore the use of corpora in teaching. Perhaps it is significant that Johns, being a teaching-oriented rather than research-oriented lecturer, never became a *habitué* of ICAME! And it is also significant that he was the first to insist that the use of computer corpora in teaching was itself a topic for research (Johns, this volume). (See also Johns 1991a, 1991b, 1993.)

Those, like Johns, who have been placing teaching with corpora at the forefront of our attention as a matter of primary interest, may well find my 'trickle down' metaphor unhelpful, or even offensive. 'Trickle down' implies that research is 'up there' as an élite

activity, and teaching is 'down here' in a lower, subservient role. But, in the experience of many, there is not a one-way dependence of this kind. One finds that 'trickle up' from teaching to research can be just as important.

The *convergence* mentioned in my title is a natural coming together of teaching with research from various points of view. This is natural whether we consider it from the 'trickle down' point of view, where the resources and techniques used in research progressively become available for teaching, or from the Johns 'trickle up' point of view, where the development of language-teaching techniques naturally appropriates to itself the resources available for research, and becomes a topic for research in its own right. The convergence is aided by the increasing similarity, in higher education, of the paradigm for research and the paradigm for teaching (using and analysing resources in a self-access mode).

Research with corpora, over the past twenty years, has become an amazingly fertile development. Whereas as recently as ten years ago corpus-based methodology was the fringe activity of a tiny minority of eccentrics, it has now become the mainstream of computational linguistics, and has increasingly established itself in mainstream linguistics. This in itself means that corpus linguistics is appearing more and more as a part of the university curriculum in linguistics, both at undergraduate and postgraduate levels.

But the thing to avoid, if we can, is treating the use of corpora in teaching as a bandwagon. Teaching bandwagons, if driven too far and too fast, can do much harm to those on the receiving end. Some will remember that ten years ago, when the new educational possibilities of CALL were very much in the ascendancy, many warned heavily against too great an enthusiasm for this new toy – the computer in the language classroom. False expectations of the powers of technology, it was recalled, had been raised by an earlier innovation, the introduction of language laboratories. The warning was salutory: since the computer entered the classroom, students have learned a lot about how to handle computers. But has their knowledge of languages taken a great leap forward? The educational benefits of technologies are still far from fully understood and acknowledged.

At that time, Higgins and Johns (1984: 12) were among those who warned against a revolutionary zeal for computers. As a conception of the role of the computer in the classroom, Higgins (1988: 12–15) preferred the metaphor of the 'pedagogue' (in ancient

Greece, the slave who accompanied the pupil to school) to that of the 'magister' (the Roman 'master' to whom the pupil submitted in obedience). Unlike the magister, the pedagogue was merely a humble facilitator of the learning process. This view of the computer certainly comes to the fore when we think of corpora. The computer is simply the device that gives access, the intermediary between the learner and the corpus as a fountain of knowledge and understanding. But we may go even one step further, and say that the *corpus* itself has no more than the facilitative 'pedagogue' role. It enables the learner/student to explore, to investigate, to generalize, to test hypotheses; but it does not itself initiate or direct the path of learning.

It is timely, none the less, to welcome the emergence of the computer corpus as a linguistic learning resource. The convergence of research and teaching is already taking place – it is a *fait accompli.* Our task is to make the best use of it, and to exchange ideas on how the computer corpus can be exploited to the best advantage in the future. This means, first, exchanging experience on how we have used corpora in teaching in the past, and how we are developing these techniques at the present time. The time is right for taking stock. My plan is to do this by surveying the activities connecting corpora and teaching, and their motivations.

2 And now?

Like many fields of endeavour, the corpus-aided language teaching field can be thought of as containing a core – a central or focal area – and an expanding periphery. The core, which can be seen as the main concern of this book, is the direct use of corpora as resources for teaching. The periphery can be seen as a set of corpus applications which indirectly contribute to teaching. Both the core and the periphery are important for the way we think about the field, and for realizing its potential. Yet a further set of activities, more peripheral still, takes the form of teaching-oriented corpus developments. The follow list summarizes the activities I have in mind:

- Direct use of corpora in teaching:
  Teaching about
  Teaching to exploit
  Exploiting to teach

- Use of corpora indirectly applied to teaching:
  Reference publishing
  Materials development
  Language testing
- Further teaching-oriented corpus development:
  LSP corpora
  L1 and L2 developmental corpora
  Bilingual/multilingual corpora.

The three main headings above can be viewed as three concentric circles, starting with the innermost one, which will occupy our attention in the following section.

2.1 Direct use of corpora in teaching

What is the nature of the interaction between corpora and teaching? The three different ways in which corpora may be used in teaching, as listed above, are distinguished by Fligelstone (1993): *teaching about [corpora], teaching to exploit [corpora] and exploiting [corpora] to teach.*

Teaching about

The first of these is probably the least interesting or innovative. As I have already said, corpus linguistics, seen as a subdiscipline within linguistics, has now come of age (see Svartvik 1992), and is beginning to find its way into curricula, both postgraduate and undergraduate. One symptom of the 'arrival' of corpus linguistics is that introductory textbooks on the subject are already being written: I am aware of more than one such publication under preparation at present.

What does teaching corpus linguistics mean? Just as a student studying for a linguistics degree (or for that matter any language-related degree – say, in English or Italian) takes courses in such subjects as phonetics, syntax, sociolinguistics, or discourse analysis, there are now beginning to be courses on corpus linguistics, or courses containing corpus linguistics as a substantial component. I recently received a letter from an East European university asking me for a basic reading list on corpus linguistics, and some information about corpora and the software available. The letter enclosed an outline of a syllabus for a corpus linguistics course

the author was planning to introduce. This was of interest to me, in showing that, for an academic who had not reached corpus linguistics directly through research involvement, the teaching of corpus linguistics was nevertheless becoming an important part of the curriculum.

As with other courses, the curriculum will tend to cover main areas of the subdiscipline: say, its history, its data and subject-matter, its methods of investigation, the models or theories it employs. In the case of corpus linguistics, inevitably important topics are: (a) what corpora exist?, (b) can they be accessed, analysed or exploited?, (c) what software can be used for this purpose?, (d) what are the applications of corpus linguistics? And arising out of these is a more philosophical or theoretical question: (e) what view of language and methods and goals of linguistic study is presented through corpus linguistics? How does this view compare with other views? (Here the Chomskian distinction between 'internalized' and 'externalized' language comes to the fore: corpus linguistics very much identifying its domain as the latter – see Chomsky 1988.)

In principle, of course, corpus linguistics could be taught as a purely academic subject, in which the students never get their hands on a computer, or gain access to a corpus. But, I would strongly suggest, almost more than any other branch of linguistics, corpus linguistics requires that students have 'hands on' experience of the subject: of the use and exploration of corpora. A course which did not provide access in this way would be like an astronomy course in which the students were never allowed access to a telescope: it would be a dull course indeed. Only through *using* corpora can one gain a first-hand sense of their potential. For example, in using a grammatically tagged corpus, one starts asking intelligent questions about how it is possible to build an automatic tagger. And one also considers questions of linguistic content such as: What sort of information does grammatical tagging give, or fail to give, about language? How can it help other aspects of language processing? Hence, 'Teaching about' naturally shades into Fligelstone's second teaching category, which is 'Teaching to exploit'.

Teaching to exploit

Since the main rationale of corpora in teaching is their immediate availability for students' use, it is important that the students

should be able to acquire the necessary 'hands on' know-how, so that they can explore corpora for their own purposes. This activity of self-access exploitation can be to a greater or lesser degree manipulated by the teacher for the student's benefit; but however 'interventionist' behind the scenes, the teacher still remains cast in the role of adviser and facilitator, rather than the authoritative source of knowledge. (In fact, occasional bouts of ignorance on the part of the teacher can facilitate the process of learning enormously.) In the course of learning how to manipulate corpus searches, the student will need, at least initially, to be supplied with sample tasks or exercises, and with feedback on those exercises. To help there are published textbooks, such as Tribble and Jones (1990), exploring the use of the Longman Mini-Concordancer (LMC) package – authored by Brian Chandler.

Works such as Tribble and Jones (1990) are particularly tied to the use of pedagogical concordance programs. In spite of their accessibility to the learner, these imply certain limitations – such as the size of the corpus that can be accessed. However, more generally, it might be desirable to give students a broader sense of what corpora are capable of. At Lancaster, a postgraduate course in how to exploit corpora (designed mainly by Tony McEnery) enables students to progress from the simpler and more restricted packages, such as the LMC, to the more challenging ones, such as WordCruncher, which allow instant access to large 'standard' corpora of a million or so words, such as the LOB Corpus, the Brown Corpus, and the Helsinki Historical English Corpus. A still greater range of corpus searching is then provided for by the in-house 'CONCORD' facility (developed by Fligelstone), enabling students to search a wide range of tagged and syntactically annotated corpora, and enabling them to make use of annotations of grammatical categories (for example, in searches for adjective sequences, or passives, or phrasal verbs). A still vaster command of data will be possible in the near future, when we begin to make serious educational use of the custom-built SARA package for searching the 100-million-word BNC (British National Corpus). So, for our students, learning to exploit corpora in all their vastness and variety is a stage-by-stage process, graduating from simpler to more abstract or sophisticated tasks. This gradual progression is important, as the gentle 'nursery slopes' of the LMC are a means of wooing beginners to the use of the computer, while the more computer-literate will want to go further faster (see Renouf, this volume).

However, a course which teaches students to exploit corpora has a larger educational content than simply acquiring the know-how to use software – although this in itself can be important. By learning to interact with the corpora, students find themselves learning a great deal about language, and how to study language. They learn about the kinds of questions that can be usefully asked and answered by reference to a corpus of data. In using an annotated corpus – for example, the tagged LOB Corpus or the BNC – they also learn what the linguistic categories used in tagging mean, and the difficulties with defining linguistic categories in this or that way. So the process of getting to grips with the software invariably shades into getting to grips with the techniques of linguistic analysis.

Exploiting to teach

What this last activity means, for me, is making selective use of corpora in the teaching of language or linguistic courses which are not intrinsically computational and would traditionally be taught by non-corpus methods. All thorough-going corpus linguists believe that the study of corpora can illuminate virtually all areas of linguistic study. The merit of the corpus is simply to enable data to be delivered in a convenient form for the investigator, whatever area of linguistics he or she is concerned with. This certainly applies in the research sphere – where corpora are used for syntactic, lexical, semantic, sociolinguistic, stylistic, psycholinguistic and historical language studies – to mention just a few subdisciplines. But what applies to the research sphere also applies to the sphere of teaching: in virtually all branches of linguistics or language learning, confrontation with relevant data can be illuminating.

Again, if I can draw upon the experience of my own department in recent years: corpus-based work has been an integral component in a number of different courses, varying from first-year undergraduate to postgraduate levels. The areas chiefly involved are present-day English language, syntax, semantics, pedagogical grammar, and the historical study of English. Typically, students undertake assignments in which they select their own topic (let us say the progressive aspect in modern English), and are provided with contextualized corpus examples of sufficient variety and scope for the study of that topic. If it is felt that the corpus-based paradigm is being overworked through its recurrence in different courses,

I would emphasize the advantages of the great variety of corpora, datasets, and techniques that may be employed with this paradigm. I would also stress the need for balance between different educational tasks at university level. The critical and argumentative type of essay assignment, which is more or less standard in many areas of university study, should certainly not be abandoned, but should be balanced with the type of assignment (often, but not necessarily, using computer corpora) which invites the student to obtain, organize, and study real-language data according to individual choice. This latter type of task gives the student the realistic expectation of breaking new ground as a 'researcher', doing something which is a unique and individual contribution, rather than a reworking and evaluation of the research of others.

In a way, 'exploiting to teach' implies 'teaching to exploit'. How can a student gain access to the corpus for the purpose just described, without being taught the means of access? Well, there is a simple way, if so desired, of avoiding the student's confrontation with the computer, which is that the students be supplied with data in the form of print-outs from the corpus. This is the method we have used, to avoid the necessary 'nuts and bolts' type of instruction students have to go through if they are to obtain their own data from the corpus. Some colleagues will regard this as 'the easy way out', and very much a second best to the 'teaching to exploit' learning sessions which produce a better long-term result. But (again, in my experience) 'the easy way' ensures that the maximum number of students are able and willing to participate in this kind of learning experience. Neither technophobes nor those who wish to acquire the necessary corpus-searching skill are discouraged from going ahead in their own self-access mode.

2.1.1 The advantages of using the computer

It may be wondered, at this point, what are the advantages of the computer in setting up learning tasks? Here I will mention four particular benefits:

1 *Automatic searching, sorting, scoring.* The computer has immense speed and accuracy in carrying out certain low-level tasks, and can therefore deliver data in a form valuable to the human learner. Concordances and frequency lists are obvious examples.

2 *Promoting a learner-centred approach.* The computer brings flexibility of time and place, and adaptability to the student's need and motivation.
3 *Open-ended supply of language data.* The computer thus encourages an exploratory or discovery approach to learning.
4 *Enabling the learning process to be tailored.* The computer can customize the learning task to the individual's needs and wishes, rather than simply providing a standard set of examples or data.

2.1.2 Divergent and convergent paradigms of learning

The above advantages can be realized in two contrasting paradigms of computer-aided learning, which may be termed *divergent* and *convergent.* The concordance-based task exploited by Johns and others is already familiar enough, and may be characterized as 'divergent' in the sense that different learners, given the same data and the same set of task instructions, will end up with very different results and interpretations, all of which may be valid in their own way. The evaluation of how well a student has performed will be itself an open-ended task, in which the assessor has to exercise judgement about the student's powers of observation, analysis, inference, organization, presentation, and (last but not least) imagination.

Contrast this with another kind of learning task, which is 'convergent', in the sense that all students given the same task will, to the extent that they are successful, tend to converge on the same answer. On the face of it, this type of task does not have the merits of being 'open-ended and tailored to the learner's need' claimed for computer-aided learning above. One thinks of the well-tried formula of the multiple-choice test, in which learners simply have to select the correct answer from a closed range of possibilities. Many CALL packages have taken this form, and although they can to a degree exploit three of the four computer strengths, those of being 'automatic', 'learner-centred' (= self-access, etc.) and 'tailored' (= adaptive), they conspicuously lack the advantage of 'open-endedness' which is essential to the exploration-and-discovery learning approach.

The 'learner-centred' and 'tailored' qualities are fully realized only where the program is fully adaptable to the learner's individual needs and preferences. To go further, 'open-endedness' is achieved only where the learner has an ability to select from an

unrestrictive range of responses, or even to come up with responses not envisaged by the teacher. These advantages of the computer corpus-based paradigm do not appear to apply, seemingly, where the goal of learning is precisely to attain the basic level of competence of distinguishing a correct from an incorrect analysis of some data. For such 'convergent' tasks, the well-tried formulae of multiple-choice, gap-filling, and other computer-delivered activities, though uncreative, would seem to be ideal.

Even here, though, a corpus-based approach can provide new advantages. Let us take, as an example, the need for students to acquire basic grammatical skills such as part-of-speech recognition and simple parsing. In a project sponsored by Lancaster University under the IHE (Innovation in Higher Education) scheme, we have developed and tested prototype programs for a self-access grammar tutor.[2] The software makes use of annotated (grammatically tagged and/or parsed) corpus data, and is in principle open-ended in the sense that the data can be selected from a very large data bank of text. The type of text, similarly, can be automatically varied – ringing the changes of spoken and written material, of different kinds of scientific materials, and so forth. From the student's point of view, the annotations are invisible on the screen, so that one very simple version of the task is to undertake part-of-speech identification by typing in labels which are then checked automatically by the computer against masked corpus annotations. Two experiments making use of this simple activity have so far indicated that the computer corpus-based technique of learning is more successful and also more congenial to students than more traditional techniques of learning grammar – such as being tutored by a grammar specialist, or using grammar textbooks. 'More successful' here means that (a) students on the whole achieve a higher accuracy rate, (b) they also achieve a greater quantity of analysis (more data being covered in the same time period), and (c) inter-student variation is smaller, perhaps indicating that the 'mental block' that afflicts many students in the study of grammar does not apply to the same degree in the computer-driven mode. One can only speculate about the reasons behind this promising result (preliminary as it is), but it appears likely that the following factors are in play. (1) The computer is not, as human teachers may be felt to be, judgemental about students' lack of grammatical knowledge. (2) The task gives students unlimited scope to try the same test repeatedly, to monitor their improvements in accuracy,

to test their capabilities on realistic samples, selected (for example) randomly or according to text-type. This self-testing, trial-and-error learning has something in common with the experience of computer games, and appears to be more stimulating and challenging than one which relies on experts' explanations and rules (although in the computer-based mode, a 'Help' facility can also provide that kind of explicit guidance).

This grammar tutoring is just one example, illustrating how the strengths of corpus-assisted teaching can carry across from the divergent to the convergent paradigm of learning and evaluation. Further sophistications are clearly needful, such as the tailoring of corpus sentences to different levels of difficulty. For this purpose, it should be possible to develop a program which selects, from a corpus, sample sentences which fulfil certain criteria, such as brevity, complexity, the presence or absence of certain lexical or syntactic categories, combinations of categories, and so on (see Wilson, this volume). In this way, the program can provide a grading of examples for students to work on at different levels of attainment.

2.2 *Use of corpora indirectly applied to teaching*

In keeping with the movement from central to peripheral areas on my subject, I will deal with the remaining topics more briefly.

Reference publishing

Probably, most English language teachers who are familiar with the idea of a text corpus first acquired that familiarity through dictionaries. In 1987, something of a breakthrough in this field was achieved by the publication of the *Collins Cobuild English Language Dictionary*, with John Sinclair as its editor-in-chief. The Cobuild dictionary was the first English language dictionary to be based on a computer corpus, the Birmingham Collection of English texts. Other major British English language dictionary publishers followed suit – indeed, in the case of one publisher, a longer corpus involvement can be claimed, since Longman had supported the Survey of English Usage Corpus and the LOB Corpus projects since the 1960s and 1970s. Longman have used a 'corpus network' extensively in their dictionaries, in particular developing the Longman-Lancaster English Language Corpus. Oxford University

Press became corpus-oriented slightly later, and took the lead in developing the British National Corpus (Leech 1993), to which Longman and Chambers also contributed. Cambridge University Press has also joined the corpus club, with its Cambridge Language Survey. Dictionaries of other languages and publishers in other countries are also benefiting from the 'corpus revolution'.

This involvement of dictionary publishers with corpora is not entirely motivated by educational goals. But there is a close and obvious link between dictionary publishing and educational publishing, as is manifest in the particular contribution that corpora have made to educational dictionaries (particularly advanced EFL dictionaries, such as the *Oxford Advanced Learner's Dictionary of Current English* (4th edn, Hornby 1989) and the *Longman Dictionary of Contemporary English* (3rd edn, 1995). Among the advantages of corpus-based lexicography are that computer corpora can be searched quickly and exhaustively, can provide frequency data, can be easily processed to provide updated lists of words, can provide authentic examples for citation, and can readily be used by lexicographical teams (especially through the mediation of computer generated concordances) for updating and verifying other levels of description such as dictionary definitions. In fact, the arguments for corpus-based lexicography (automatic processing, authentic data, etc.) are very similar to the arguments for use of corpora in language teaching. But, through dictionary publishers, the resources are concentrated in one place, and are filtered through dictionary publication, and increasingly through other media, to the educational market. In this sense, then, educational users of improved dictionaries have benefited at one remove from the corpus revolution, without having to have access to computer resources.

Under the heading of 'Reference publishing', although we naturally think first about printed dictionaries, we should also be aware that conjunction of corpora and language reference resources is expanding in new directions. One direction is towards electronic modes of publication: interactive dictionary resources on CD-ROM, for example, are now becoming widespread.[3] Another direction is towards different kinds of reference work. As early as the 1970s (Quirk *et al.* 1972, 1985) reference grammars of English were drawing heavily on corpus materials, and a new degree of corpus-dependence was reached with the *Collins Cobuild English Grammar* (Sinclair 1990), for which it was claimed that all cited examples were drawn from the Birmingham corpus materials. Probably we

still await the publication of the first thorough-going corpus-based grammar, in which all rules, generalizations, structures, etc., are derived from or tested against the evidence of 'real language' found in corpora. Another direction in which corpora are beginning to make an impact on reference publications is in the provision of frequency information, which in many respects is becoming available for the first time with the publication of corpus-based reference works. But there is a long way to go before the necessary manual and/or automatic processing provides us with 'push button' frequency information for word senses, grammatical structures, etc., in a form which the language teacher would find immediately useful. It is easy, however, to foresee that in the mid-term future, the corpus-based developments in reference publishing that I have mentioned will be readily available in combination: so that, for example, computer-delivered reference works of new kinds (frequency dictionaries, frequency thesauri, corpus grammars, and the like, interlinked with corpora) will be available in new and exciting forms, such as interactive access through multimedia and hypermedia.

Materials development

In the publishing world, there is a 'trickle down' process from large, scholarly works to smaller, non-specialist or educationally oriented works, as shown classically in the family of Oxford dictionaries deriving their pedigree from the *OED*. A similar 'trickle down' occurs with corpora. Large corpora provide the basis for large, scholarly dictionaries and grammars. But they also have a clear spin-off in the direction of teaching materials, as has been demonstrated by the *Cobuild* family of publications produced by Collins (now HarperCollins). From the *Cobuild* point of view, a philosophy which links corpora in reference publishing to corpora in ELT materials development is spelt out in Dave Willis's book entitled *The Lexical Syllabus: a new approach to language teaching* (Willis 1990). In this philosophy, stress is laid on the importance of frequency of occurrence, a form of information which is for the first time starting to become widely and informatively available to the language teacher through corpora. Yet we still await a comprehensive frequency dictionary of English, and West's ancient pre-electronic *A General Service List of English Words* (West 1953), first published in 1936, has still not been superseded.

The contributions of corpora to language-teaching materials can be thought of under three headings: (a) first, the provision in abundance of frequency information (see Mindt, this volume, for the relevance of frequency to ELT materials); (b) second, the availability of copious examples of authentic language in use; (c) third, the provision of computer-delivered learning packages, such as we have already discussed above. In the present context, I will concentrate only on the first of these, frequency information.

Whether we focus on lexical frequencies, as Willis does, or extend our interest to grammatical frequencies (e.g. frequency of grammatical structures), the revival of serious interest in the relevance of frequency to language teaching is well overdue. True, there are well-known objections, in applied linguistics, to the use of frequency criteria in deciding what to teach, when and to whom. Among them is the argument that frequency is only one of a number of criteria for deciding teaching priorities, and perhaps not the most important (van Els *et al.* 1984: 210–12). Another argument is that corpora, at least as they exist at present, are not the most reliable sources of frequency data: for example, a corpus of adult written English (which is the easiest kind of computer corpus to obtain) is not a good guide as to what is most frequent in the spoken language of children. Available lexical frequency lists, unfortunately, are still predominantly based on written language. Against these arguments, Kennedy (1992) argues cogently that language teachers, syllabus designers, and materials writers are wilfully ignoring compelling frequency evidence already available. Whatever the imperfections of the simple equation 'most frequent' = 'most important to learn', it is difficult to deny that the frequency information becoming available from corpora has an important empirical input to language learning materials.

Language testing

The arguments about corpora in direct language teaching (see 2.1 above) are of equal relevance to language testing. Testing, like teaching, benefits from the conjunction of computers and corpora in offering an *automatized, learner-centred, open-ended* and *tailored* confrontation with the wealth and variety of real-language data (see Alderson 1996 for some aspects of this). In fact, it seems that corpus-based CALL, of the 'convergent' type discussed in 2.1 above, can be characterized as self-testing. In convergent language/

linguistics teaching and in language testing, the same advantages of the corpus appear: in both cases, there is a strict control on the nature of the task, so that automatic scoring in terms of 'correct' and 'incorrect' responses is feasible. At the same time, the corpus provides the advantage of using genuine 'real-life' samples in sufficient quantity so that, if required, selection of test samples can be randomized. In principle, authentic samples of the language can be automatically graded by a range of criteria. And the general paradigm whereby the learner is confronted with a text sample, the grammatical characteristics of which are stored in the computer, but not visible on the screen, can be applied both to computer-based teaching and computer-based language testing.

2.3 Further teaching-oriented corpus development

One of the continuing difficulties for applications of corpora to language/linguistics teaching is the lack of suitable corpora. It is a sad fact that the types of corpora which are most easily available for the computer today consist largely of written texts, whereas the types of corpora which would most faithfully reflect the priorities of language learning would contain at least as much spoken material as written material.

To put it more generally: for human beings, experience of language is primarily spoken and secondarily written; for computers, conversely, language is primarily written and secondarily (via transcription) spoken. This difference in humans' and computers' experience of language is reflected in the history of corpus linguistics. The first computer corpora (e.g. the Brown Corpus, 1961–64) consisted of written English: only later, in the computerizing of the London–Lund Corpus in the mid-1970s (see Svartvik 1990) was it possible to produce corpora of spoken English – naturally, through the mediation of transcriptions (renderings of speech in written form). This prioritizing of the written language has persisted through the history of corpus development. For example, the BNC consists of *c.* 90 million words of written English, and only *c.* 10 million words of spoken language. Why this discrepancy? The reason, simply, is that spoken language is much more expensive to collect, at the present stage of technological progress, than written language. It costs roughly as much to collect 10 million words of spoken language (requiring, as this does, manual transcription)

as to collect 90 million words of written English – and even then, the result is little more than a basic orthographic transcription.

On the other hand, the learning of spoken language is generally felt to be a *sine qua non* of language learning today: the idea of learning the written language, without being able to make use of the spoken language, makes good sense to a computer, but not to a human being.

This time-lag between the collection of written corpora and spoken corpora is no longer as important as it was. After all, there now exist corpora containing 10 million words or more of the spoken language. However, the general problem that the history of spoken and written language corpora reveals is that the corpora which are easiest to compile are not necessarily those which are most useful for language learning purposes. This leads to the question: what kinds of corpora do we need to develop, to make up the deficit between what corpora exist, and what corpora are needed for the best applications to language teaching? These resources are now being developed, albeit somewhat haphazardly.

LSP corpora

LSP, or 'Language for Specific Purposes', has an important place in the goals of language teaching. For example, many millions of people, throughout the world, need to know English particularly for a specialist subject which they are studying or professionally practising: science, technology, law, medicine – to mention a few. Although, again, the history of LSP thinking in applied linguistics has not favoured such an approach, it makes good sense to find out as much as we can about the linguistic characteristics of language varieties – including, for example, lexical frequencies, collocations, and characteristic grammatical structures. For this purpose, there is a clear need for LSP corpora. Such corpora are coming into existence gradually, by three different means. First, keen LSP linguists and teachers have developed their own corpora: early examples were the JDEST [science and technology] Corpus (Yang 1985) and the GPEC [petroleum industry] Corpus (Zhu 1989), both produced in China. Second, now that we have enormous general-purpose corpora, such as the BNC, it will be possible to select from such corpora a subcorpus dealing with a certain domain. (The informative written part of the BNC, for example, is divided into domains such as 'pure science', 'applied