

The left half of the book cover features a dense, abstract pattern of small, irregular, colorful shapes (yellow, blue, red, and black) scattered across a white background. These shapes resemble stylized, elongated letters or fragments of text.

John McH. Sinclair

*How to Use
Corpora in
Language
Teaching*

Studies in Corpus Linguistics

How to Use Corpora in Language Teaching

Studies in Corpus Linguistics

Studies in Corpus Linguistics aims to provide insights into the way a corpus can be used, the type of findings that can be obtained, the possible applications of these findings as well as the theoretical changes that corpus work can bring into linguistics and language engineering. The main concern of SCL is to present findings based on, or related to, the cumulative effect of naturally occurring language and on the interpretation of frequency and distributional data.

General Editor

Elena Tognini-Bonelli

Consulting Editor

Wolfgang Teubert

Advisory Board

Michael Barlow

Rice University, Houston

Robert de Beaugrande

Federal University of Minas Gerais

Douglas Biber

North Arizona University

Chris Butler

University of Wales, Swansea

Sylviane Granger

University of Louvain

M.A.K. Halliday

University of Sydney

Stig Johansson

Oslo University

Susan Hunston

University of Birmingham

Graeme Kennedy

Victoria University of Wellington

Geoffrey Leech

University of Lancaster

Anna Mauranen

University of Tampere

John Sinclair

University of Birmingham

Piet van Sterkenburg

Institute for Dutch Lexicology, Leiden

Michael Stubbs

University of Trier

Jan Svartvik

University of Lund

H-Z. Yang

Jiao Tong University, Shanghai

Volume 12

How to Use Corpora in Language Teaching

Edited by John McH. Sinclair

How to Use Corpora in Language Teaching

Edited by

John McH. Sinclair

John Benjamins Publishing Company
Amsterdam/Philadelphia



™ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Cover design: Françoise Berserik

Cover illustration from original painting *Random Order*

by Lorenzo Pezzatini, Florence, 1996.

Library of Congress Cataloging-in-Publication Data

How to use corpora in language teaching / edited by John McH. Sinclair.

p. cm. (Studies in Corpus Linguistics, ISSN 1388-0373 ; v. 12)

Includes bibliographical references and indexes.

1. Language and languages--Computer-assisted instruction. I.

Sinclair, John McHardy, 1933- II. Series.

P53.28 .H69 2004

418'.00285-dc22

2003067697

ISBN 90 272 2282 7 (Eur.) / 1 58811 490 2 (US) (Hb; alk. paper)

ISBN 90 272 2283 5 (Eur.) / 1 58811 491 0 (US) (Pb; alk. paper)

© 2004 – John Benjamins B.V.

No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

	List of contributors	vii
	Introduction <i>John Sinclair</i>	1
The corpus and the teacher		
<i>In the classroom</i>	Corpora in the classroom: An overview and some reflections on future developments <i>Silvia Bernardini</i>	15
<i>In preparation</i>	What teachers have always wanted to know – and how corpora can help <i>Amy B M Tsui</i>	39
Resources – Corpora		
<i>Corpus variety</i>	Corpus linguistics, language variation, and language teaching <i>Susan Conrad</i>	67
<i>Spoken – general</i>	Spoken corpus for an ordinary learner <i>Anna Mauranen</i>	89
<i>Spoken – an example</i>	The use of concordancing in the teaching of Portuguese <i>Luísa Alice Santos Pereira</i>	109
<i>Learner corpora</i>	Learner corpora and their potential for language teaching <i>Nadja Nesselhauf</i>	125

Research

<i>Composition</i>	The use of adverbial connectors in Hungarian university students' argumentative essays <i>Gyula Tankó</i>	157
<i>Textbooks</i>	A corpus-driven approach to modal auxiliaries and their didactics <i>Ute Römer</i>	185

Resources – Computing

<i>Basic processing</i>	Software for corpus access and analysis <i>Michael Barlow</i>	205
<i>Programming</i>	Simple Perl programming for corpus work <i>Pernilla Danielsson</i>	225
<i>Network</i>	Learner oral corpora and network-based language teaching: Scope and foundations <i>Pascual Pérez-Paredes</i>	249
Prospects	New evidence, new priorities, new attitudes <i>John Sinclair</i>	271
	Notes on contributors	301
	Index	305

List of contributors

Silvia Bernardini

SSLMIT

University of Bologna

Corso della Repubblica 136

47100 Forlì, Italy

Amy B M Tsui

Chair Professor

Faculty of Education

The University of Hong Kong

Pokfulam Road, Hong Kong SAR

Susan Conrad

Department of Applied Linguistics

PO Box 751

Portland State University

Portland OR 97202-0751, USA

Anna Mauranen

Professor of English

Head of School

School of Modern Languages

and Translation Studies

FIN-33014 University of Tampere

Finland

Lúisa Alice Santos Pereira

Centro de Linguística

da Universidade de Lisboa

Av. Prof. Gama Pinto, 2

1649-003 Lisboa, Portugal

Nadja Nesselhauf

English Department

University of Basel

Nadelberg 6

4051 Basel, Switzerland

Gyula Tankó

Assistant Lecturer

Department of English

Applied Linguistics

Eötvös Loránd University

Ajtósi Dürer sor 19-21

1146 Budapest, Hungary

Ute Römer

English Department

University of Hanover

Königsworther Platz 1

30167 Hannover, Germany

Michael Barlow

Department of Applied Language

Studies and Linguistics

The University of Auckland

Fischer Building

18 Waterloo Crescent

Auckland, New Zealand

Pernilla Danielsson

Centre for Corpus Research

School of Humanities

University of Birmingham

Edgbaston

Birmingham B15 2TT, UK

Pascual Pérez-Paredes

Departamento de Filología Inglesa

Campus de la Merced

Universidad de Murcia

30071 Murcia, Spain

John Sinclair

via Pandolfini 27

50122 Firenze, Italy

Introduction

John Sinclair

Substantial collections of language texts in electronic form have been available to scholars for almost forty years, and they offer a view of language structure that has not been available before. While much of it confirms and deepens our knowledge of the way language works, there is also a fascinating area of novelty and unexpectedness – ways of making meaning that have not previously been taken seriously. Further, in studying corpora we observe a stream of creative energy that is awesome in its wide applicability, its subtlety and its flexibility.

This cornucopia has not been welcomed with open arms, neither by the research community nor the language teaching profession. It has been kept waiting in the wings, and only in the last few years has any serious attention been paid to it by those who consider themselves to be applied linguists. For a quarter of a century, corpus evidence was ignored, spurned and talked out of relevance, until its importance became just too obvious for it to be kept out in the cold.

The reasons for this neglect of vital information need not detain us long. Just as the first electronic corpora were taking shape in the early nineteen-sixties,¹ the focus of linguistic theory was shifting from the study of empirical data to the study of the mental processes that together are often called the language faculty. This approach preoccupied most linguists until recently, and may still be the dominant paradigm world-wide. After a few awkward attempts at the application of mentalist theory to language teaching, its relevance was generally accepted as minimal, and so a gap opened up between the theory of language and the teaching of languages, to the great detriment of the teaching profession. Applied linguists, whose jobs were originally designed to mediate between theory and practice, took on the additional burden of providing quasi-theoretical underpinning for the linguistic side of language pedagogy, but their descriptions were not detailed enough to provide a firm foundation.

As a consequence, enquiry about the nature and structure of languages was discouraged, and everyone's attention turned to methodology.

The first signs of the language teaching profession taking an interest in corpus work came in a recognition that the teaching of lexical and phraseological structures needed a higher priority than they currently had, and – a little later – that reliable information about these structures could not be retrieved by introspection. Shortly after this the study of language variety found a new accuracy because comparisons could be made of substantial corpora, and terminology began to re-integrate with text, from which it had been separated by inadequate theories of meaning (Pearson 1998). Now corpora, large and small, are seen by many teachers as useful tools, and are being put to use more and more every day. Access has become fairly easy on standard small computers, user-friendly software is available for most normal tasks, websites are accumulating fast, and corpora are almost part of the pedagogical landscape.

To make good use of corpus resources a teacher needs a modest orientation to the routines involved in retrieving information from the corpus, and – most importantly – training and experience in how to evaluate that information. It is the second point that has caused much controversy, because a corpus is not a simple object, and it is just as easy to derive nonsensical conclusions from the evidence as insightful ones. Those who during the last decade tried to barricade the profession against the influence of corpora recycled the critical arguments of the theoreticians thirty years before, and we heard again that no corpus can be a totally accurate sample of a language, that occurrence in a corpus is no guarantee of correctness, that frequency is not a sound guide to importance, that there are inexplicable gaps in the coverage of any corpus, however large, etc.

That flurry of resistance is now largely behind us, and it is timely to consider the issue posed as the title of this book, how to use corpora in language teaching, since corpora are now part of the resources that more and more teachers expect to have access to.

Background to this book

The book was conceived as part of the activities of The Tuscan Word Centre, which is a non-profit company that exists to promote the scientific study of language. Its principal public activity is the regular organisation of short intensive courses, and in October 2001 it hosted a course with the same title as this book.² Experts in various aspects of the field were invited to lead topics, and a conscious effort was made to attract younger topic leaders rather than

the first generation of corpus linguists, who were hovering around retirement. The book was thus designed round seven papers from scholars with rising reputations, which were commissioned in advance of the course. I provided the overall design and a paper based on my contribution to the course.

The course was a popular and lively event, and the participants were invited to submit papers to join the commissioned ones. There was a good response, from which another four papers were chosen to give some representation to current research in Europe. Several of the papers were completed shortly after the course, and so make only passing reference to very recent publications – see particularly my comments below on Nadja Nesselhauf's survey of learner corpora. Short biographies on each of the participants can be found at the end of the book.

Design and content

The book begins with two papers that have the teaching process at the centre. **Silvia Bernardini** opens with an overview on the use of corpora in the classroom that highlights the pedagogic approaches rather than the data knocking at the door. She points out that after a quiet start the variety and energy of current work is impressive, and she goes on to set out her own approach, which points towards the future. It is a kind of discovery learning, harnessing powerful tools and resources as supports to the student.

While reviewing the whole field of corpus-oriented methods, Silvia's paper turns on more than one occasion to actual language data and the language user's response to it; this firmness of reference is characteristic of work in corpus linguistics, and will be found in several of the other papers.

Silvia is not only concerned with turning out students with an excellent command of English; many of them are destined to become professional translators, and so the development of problem-solving skills in an information-rich society has a special relevance to them, while being a fundamental resource for any language user.

The second paper concerns, as its title makes clear, "What teachers have always wanted to know – and how corpora can help". It is written by **Amy Tsui**, and it tells of a remarkable corpus-centred facility that has been made for the English teachers of Hong Kong. Most of the teachers there are native Cantonese speakers and have been trained locally; on the other hand the position of Hong Kong in the international trading community sets very high operational standards for English. The teachers' feelings of insecurity are shared in chat rooms, the language problems are assessed by an expert team under Amy's di-

rection, with reference to substantial corpus resources. Most of the queries are not unique to a single teacher, but recur frequently, and so they are posted in a growing database of immense value to the teaching community.

This pioneering work has been developing for almost a decade now, and is mature and well-established. As well as illustrating the kind of support that a community of language teachers needs and deserves, it also is a first reminder that well-distributed languages like English acquire a local flavour, setting tricky problems for teachers searching for appropriate models.

The second section of the book focusses on CORPORA themselves. As the primary source of data for this kind of language teaching, the way they are designed is of central importance.

There is nowadays a wide variety of corpora available, and also corpora which show variety within a single collection. This second kind of corpus allows researcher, teacher, student or any combination of these to explore the way in which language users make particular selections for particular occasions and particular tasks. Appropriacy of language to the purpose has always been an enduring problem for language learners, and Susan Conrad reviews the contribution that corpora can make in this important area.

She points out forcibly that attention to variation cannot be ignored in language learning, and it is not confined to specialised varieties, but pervades the central area of language use. This point is illustrated with an example that demonstrates that our received view of language use is not consistent with observation, and that the intuitions we have – even those of a native speaker – need to be complemented by corpus evidence.

Looking ahead to the section on computing which follows, Susan then describes a software tool that is capable of assessing several variables at the same time, thus giving substance to the notion of language variety.

Since the very beginning of corpus linguistics (Krishnamurthy 2004 (1970)), collections of spoken language – especially impromptu conversations – have exercised a particular fascination for researchers. They seem to catch the language off its guard, so to speak, and show its workings in a way that is often disguised in the blandness of writing. When computer typesetting became possible, there was an explosion of data from the printing industry that overwhelmed the relatively small collections of spoken language. Because there is as yet no chance of automatic transcription of ordinary conversations, there is a laborious and expensive process of transcription to be done, and that “reduces” the speech event into a written record of it, losing crucial information about the stressing, intonation, pausing and general delivery.

Despite this, and with promise of technical improvements on the way, recent years have seen a resurgence of interest in spoken corpora, and this is celebrated by **Anna Mauranen** in the next contribution to the corpus section. In a thoughtful state-of-the-art paper she considers the place and value of spoken corpora in the language teaching/learning process. This raises issues like authenticity, still a controversial topic in the classroom, and Anna takes a balanced attitude to it, joining other contributors to this volume in pointing out that corpus data is certainly superior to invented or adapted data. She stresses that some orientation is required for both student and teacher if they are to make the best use of corpora, and avoid the pitfalls of a procedure that is more complicated than it looks. Looking ahead, she points out that the proliferation of corpora will gradually displace the native speaker from central position as model and adjudicator of a language in use, and offer alternatives such as expert non-native speakers.

As an example of a large and recently-established spoken corpus, and what can be done with it, the next paper, by **Luísa Alice Santos Pereira** describes resource-building at the University of Lisbon, and some possibilities envisaged for applications such as language teaching. Portuguese is one of the most widespread languages of the world, with the fifth largest group of native speakers, and to make a reference corpus of it is a major task. Luisa's group, the Centro de Linguística da Universidade de Lisboa, has been accumulating resources for some years, and makes them available to the profession. One of their most impressive publications is a set of 4 CD-ROMs containing large samples of spoken Portuguese from the many countries where this language is in daily use. The samples are cleverly presented, with sound and transcript aligned.

Luisa gives several clear examples of the kind of information that is only obtainable from a corpus, and which is of great value to language learners and teachers, as well as to other professional users of language data. The differing frequencies of forms and lemmas is one important area for an inflected language, and the collocation profiles of near-synonyms are directly useful in the classroom. Her paper is full of information about the corpora and gives valuable addresses and links.

Finally in the corpus section **Nadja Nesselhauf** reviews the state of play in the making of corpora which are specially designed for research into language learning – the learner corpora. This initiative grew naturally from the large collections of learners' errors collected in several centres, and, led by the University of Louvain-la-neuve in Belgium has flowered into a many-faceted movement, collecting specimens of the language of learners with all sorts of language backgrounds. Nadja covers the whole world in her survey, showing

a remarkable amount and range of activity, and she sets out the advantages and limitations of using a learner corpus in support of language learning. She stresses that most applications of learner corpora require comparison with a standard corpus of native-speaker quality and reliability, and the potential of corpora to compare different varieties, introduced in Susan Conrad's paper, is taken further here. Nadja covers most of the important work in this important field and gives her own assessment of it.

Just as Nadja was finalising her paper, there was an important publication in the field of learner corpora (Granger et al. 2002). It was too late for her to include this work in her chapter, but she has in the meantime written a review of the book which is scheduled to be published in *IJCL* 8.2. With the review as a kind of appendix to her paper, Nadja's account of the field is fully up to date.

The next section gives a small selection of current RESEARCH interests, a glimpse of what is going on among the younger researchers. The paper from Gyula Tankó follows neatly from the discussion of the use of learner corpora, because it is a detailed research report on the differing uses of connectives between fluent Hungarian writers of English and similar writings from native speakers. Gyula first sets out the way connectives are presented in general grammars of English and in popular teaching materials, establishing the importance of corpus evidence in a complex area of central importance to effective written communication. Then he describes a small but well-focused corpus of Hungarian writers, and compares the number of connectives, the number of different types, and the choice of certain individual forms in his corpus as against a reference collection of native English writing. The results are extremely revealing, and Gyula goes on to discuss how the apparent divergent choices of the Hungarian writers might be guided into reliable and conventional patterns. Many of the points he makes echo Nadja's presentation of the use and value of learner corpora.

Next Ute Römer compares patterns of distribution of modal verbs in a corpus of spoken English with a group of texts culled from a best-selling German textbook for learners of English. Not only do the raw frequencies vary a lot, but since each modal has several meanings, Ute shows that the meanings chosen by the textbook writers have a different pattern of occurrence from that noted in the corpus of naturally occurring English. Ute closes with some recommendations for improving the representativeness of models of English presented to learners.

The pattern of Ute's findings echo one of Susan Conrad's examples, where again a piece of English, put forward as a model of a kind of English and probably written for the purpose, does not show the same features as are found in

appropriate selections from a corpus. Scholars have warned repeatedly that it is asking too much of the most able speaker of English to manufacture text without the constraints and support of a genuine communicative event.

We now turn to a section on **COMPUTING**, concerning the details of making corpora do what you want them to do. Frequently in publications in computational/corpus linguistics the work on the language texts and the work on the computer programs and other technical matters are kept separate – in different books, for example. The authors in this section argue that competent users of computational resources should have a detailed awareness of the jobs that are done and the facilities that are available from the technical experts. There is already a worrying lack of critical assessment of existing software and corpus resources from user groups, who are often so delighted to find something that “works” that they do not check what exactly it does or does not do.

First **Michael Barlow** shows how basic information can be retrieved from a corpus, and how it can be interpreted. Corpus evidence is essentially *indirect*, which means that it cannot be taken at face value but must go through a process of interpretation, and Michael makes it clear how careful it is necessary to be, and how apparently innocuous decisions at one point in the retrieval process can fundamentally affect the output. Anyone using a corpus should know the way in which the basic sorting and retrieving operations work, and how what seem to be simple and low-level decisions³ can have a profound effect on the evidence returned from a query. Michael regards the various operations like making word lists, concordances and collocational profiles as essentially rearrangements of the corpus, each allowing us a different viewpoint, each of them highlighting some patterns and obscuring others. This is a helpful concept when one is grappling with understanding what the computer is doing. Michael’s explanations are very clear and supported with copious examples throughout, and his presentation has the authority of one of the leading providers of corpus processing software, in MonoConc and ParaConc (see his website <http://www.ruf.rice.edu/~barlow/>). Perhaps the key point in Michael’s paper is that any display of corpus information is necessarily partial, and that important patterns may be concealed by the software settings and strategies. The evidence needs to be interpreted with some awareness of the design of the software query package.

The chapter by **Pernilla Danielsson** looks quite challenging at first, as she offers the reader the chance to write from scratch four fundamental programs for corpus handling – a tokeniser, a word splitter, a frequency counter and a KWIC concordancer. Many of Michael’s corpus rearrangements can be carried

out on a corpus of one's own choice using these tools, and Pernilla shows how easy it is to adapt these central programs for particular purposes.

In the daily business of using corpora there are frequently situations where a program needs a simple adjustment, or a file for input turns out to be in an inappropriate format, or it would speed things up if you could just stitch together two or three small programs without having to take the results from one and input them to the next – small jobs, without mystery, but much more convenient if the user can modify the files rather than call in an expert or – more likely – wait in the queue.

Pernilla shows that there are some arbitrary conventions to learn, and some procedures that reduce the likelihood of error, and then the programming gives great satisfaction and useful results for only a small input of labour and attention. She concentrates on the Perl language which is particularly favourable to text handling.

In my opinion these two chapters set out the minimum competence in, and awareness of, actual corpus computing that anyone using corpora extensively should have; many, of course, go far beyond this beginners' kit.

Finally in this more technical section there is a paper that combines the use of learner oral corpora and network-based language teaching, written by Pascual Pérez-Paredes and based on his own experience in Murcia. While this chapter could have been placed in the section on corpora, because it has strong links with both oral corpora (Anna Mauranen) and learner corpora (Nadja Nesselhauf), it is also valuable for its practical orientation in the use of technical facilities, and the integration of resources, software and hardware in support of the language learning. It is also the only paper to deal directly with computer-assisted language learning (CALL), an important movement that is developing in parallel with corpus-oriented language learning. Data-driven learning (DDL), which is often referred to in this book, is the cord that joins the two approaches.

Originally – that is some twenty years ago – the main difference between the two was that CALL dealt in small-scale programs and packages, often trimmed to what was the current capacity of computers that were affordable by teachers; in contrast corpus research was always conscious of the need to make larger and larger corpora to track down the recurrent patterns in the everyday language. Now, with substantial corpora available to all, there is not so much difference between them, and Pascual sees a valuable link in their common interest in learner oral corpora.

Pascual makes it clear that the technical breakthroughs of recent years, in corpus construction and networking, offer the prospect of new methodologies,

unimagined in the early work; in particular the creative moves available to the student working in a well-designed local area network are much improved.

The final section is entitled PROSPECTS, and contains only one paper, my own. I have been interested for many years in the revelations about language that arise in corpus investigation, because they have been so unexpected. At the start of the Cobuild project in 1980 I assumed that the use of a corpus would improve accuracy and comprehensiveness, and would speed up the process of lexicography because of the clarity of the descriptions and the organising power of the computer. Some of this proved to be correct, but I grossly underestimated the effect of the new information that the corpus supplied, and in particular the total lack of fit between the evidence coming from the corpus and the accepted categories of English lexicography. The Cobuild team had to reconceptualise the dictionary in the light of the early evidence.

It was clear not only that matters of detail needed to be revised, but descriptive categories and, later, theoretical positions. Changes in priorities gradually gave a different shape to the model of language, e.g. from concentration on the word as the carrier of lexical meaning I moved to the notion of the *lexical item*, which can be several words in length, and now give it pride of place as the prime carrier of lexical meaning. This in turn opens up a more complex descriptive apparatus for lexis, with at least two levels in a hierarchy.

As I contemplated changes of this kind, I realised that they were likely to have a profound effect on the teaching and learning of languages, because the new descriptions would represent language in a different way. This effect would take place regardless of whatever pedagogical precepts were fashionable, regardless of the stance, welcoming or – more commonly – discouraging, of applied linguists. If resistance to the new ideas remained strong, the problem would appear insuperable, and the profession of language teacher could become extremely depressed and heavy with warring factions, because, viewed through a traditional model, the new categories and statements are atomised into a mass of apparently unconnected detail and seem confusing and impossible to assimilate. Since language teaching is well known for its conservatism, the prospect was grim.

So I decided in my contribution to this book to approach the issues through a discussion of some well-known features of language and its teaching that are often held to be problem areas, and see if a revised perspective, informed by corpus evidence, gave promise of improving the situation.

Acknowledgement

Ute Römer, in addition to her own contribution to this volume, took on the job of reading proofs with me, for which I am most grateful, and which speeded up the production process.

Notes

1. See Francis and Kučera 1979 and Krishnamurthy (Ed.) 2004 (1970).
2. Several participants on this course, including some of my co-authors, were aided by grants from the European Commission, under contract no. HPFCT-CT-1999-00224. The Commission's support is gratefully acknowledged.
3. A recent example that was reported to me concerned the Bank of English, where it appeared that on one day there were lots of instances returned of the word "Taliban", and a few days later none at all. It is most unlikely that the corpus was tampered with, and indeed the word reported missing is definitely still there in numbers. The most likely cause of this is the setting, somewhere in the software, of the "case sensitivity". If the query is case sensitive, then a search for "taliban" will be unsuccessful, but if the case is insensitive then all the instances of "Taliban" will be returned by that search.

References

- Francis N. & H. Kučera (1979). *Manual of Information to Accompany a Standard Sample of Present – day American English*. Providence: Brown University Press.
- Granger S. J. Hung & S. Petch-Tyson (Eds.). (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.
- Krishnamurthy R. (Ed.). (2004 [1970]). *English Collocation Studies: The OSTI Report* (by John Sinclair, Susan Jones and Robert Daley). Birmingham: Birmingham University Press.
- Pearson J. (1998). *Terms in Context* [Studies in Corpus Linguistics 1]. Amsterdam: John Benjamins.

The corpus and the teacher

In the classroom

Corpora in the classroom

An overview and some reflections on future developments

Silvia Bernardini

University of Bologna

By stages we have been able to move much closer to a situation where we can give the hoped-for response: ‘go to any of the labs, hit the icon which says “Corpus” and follow the instructions on the screen’. (Fligelstone 1993: 101)

Within corpus-aided language pedagogy, a distinction can be made between uses of corpora as sources of descriptive insights relevant to language teaching/learning, and uses of corpora that directly affect the learning and teaching process(es). This chapter, which is concerned with the second of these two aspects, retraces the development of data-driven/discovery learning approaches, presents their rationale and describes some relevant corpus typologies and applications, with special reference to the fields of LSP and translation teaching. It suggests that the challenge for corpus-aided discovery learning, now that corpus construction and access have become easier, is to make sure that these powerful tools and methodologies find a role in the language classroom – for communicative reasoning-gap activities, strategic and serendipitous learning as well as reference purposes – as central as that they have already secured in other areas of applied linguistics.

1. Introduction

Corpora seem to have entered the classroom from the backdoor. Whilst corpus data have long established themselves as *the real language data* (paraphrasing Cobuild’s famous catchphrase), sweeping away resistance as to their descriptive and, more controversially, pedagogic value, the actual use of corpora in language learning settings has for a long time remained somewhat behind such momentous breakthroughs. This now seems less true, however, judging from the number of conference papers, software applications and corpora address-

ing the issue of “how best corpora and corpus linguistics can aid language learning and teaching”, as opposed to “what language facts of relevance to language learning and teaching can be derived from corpora”. The latter is an equally interesting, but arguably different issue, which is discussed in a number of contributions to this volume and will be only slightly touched upon here. Instead, this paper focuses on the first issue, the theoretical and practical implications of the body of work dealing with corpora in the classroom, looking back on early insights and ahead to future developments. Particularly, we shall focus on those ideas that have helped us rethink *language pedagogy* from a corpus perspective, in the same way as we are witnessing an increasing interest in rethinking *language description* and *linguistic theory* from a corpus perspective.¹

2. Bringing corpora to the classroom

2.1 Data-driven Learning (DDL) or “The learner as researcher”

Johns’ (e.g. 1991) work on *data-driven learning* has proved extremely influential and ground-breaking in showing the relevance of corpus analysis techniques to the wide and varied audience of language teachers and students around the world. Much if not all subsequent work in this area owes something to Tim Johns’ pioneering efforts, which constitute a truly “applied linguistics” approach, in Widdowson’s well-known terms (1984).

Johns suggests that learners should be guided to *discover* the foreign language, much in the same way as corpus linguists discover facts of their own language that had previously gone unnoticed. A similar viewpoint is expressed by Leech (1997: 10) who claims that

The critical and argumentative type of essay assignment [...] should be balanced with the type of assignment [...] 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.

This shift of emphasis from deductive to inductive learning routines has wide-ranging effects on: (a) the teacher, who becomes a coordinator of research, or facilitator; (b) the learner, who learns how to learn through exercises that involve the observation and interpretation of patterns of use; (c) the role of

pedagogic grammars, whose level of abstraction often works against their effectiveness. A classic case might be article usage, a well-known problem area for many foreign learners of English, even at advanced levels. Its intricacies make this aspect of English lexico-grammar little amenable to neat classifications where corpus work, on the contrary, can provide enough evidence and stimuli for the learner to arrive at developmentally-appropriate generalisations (i.e. accounts that are not necessarily correct and exhaustive, but agree with the learner's current language system). Although descriptive grammars like the *Longman Grammar of Spoken and Written English* (Biber et al. 1999) have recently addressed the issue of the inadequacy of traditional grammars in coping with corpus evidence, offering corpus-derived insights and information about frequencies, Johns' claim seems to go beyond, and suggest that even corpus-based grammars do not offer the same potential as corpora in the development of abilities to "identify – classify – generalise" on the basis of language experience, one of the abilities on which learning in general, and autonomous learning in non-institutional settings in particular would seem to rely. Learner empowerment is a common thread within the body of work discussed in this paper, and one of the most interesting aspects of pedagogy in a corpus perspective.

2.2 Language learning as (schema-based) restructuring

Whilst Johns' approach focuses on the role of corpus use in the development of learning capacities and in the establishment of a non-authoritarian learning environment, a number of scholars have suggested that concordancing in particular may prove unique in the acquisition and restructuring of competence.

Language learning may be viewed as an inductive process in which meaning and form come to be associated. This view agrees well with the cognitive psychology work on memory known as *schema theory* (a schema is a trace left by an event we experience, individualised and selected for remembering according to our "appetite, instinct, interests and ideas" (Bartlett 1932:206)). Language learning in a schema perspective is a process that involves the development or adjustment of real world knowledge structures or *schemata* appropriate to the target language culture, and the matching of these with relevant pragmatic and linguistic schemata. By providing access to authentic interaction (both written and spoken, both monological and dialogical), corpora offer an ideal instrument to observe and acquire socially-established form/meaning pairings. In other words, they allow learners to observe *what* is typically said in

given circumstances, and *how* it is typically said, and to relate the two (cf. the Sinclairian *idiom-principle* e.g. J. M. Sinclair 1991a).²

Barlow (1996:30) claims that schema-meaning pairings are constructed on the basis of repeated experiences of instances of language use. A concordancer may short-cut this lengthy process since, “by concentrating and manipulating instances of a language phenomenon, [it] makes the patterns stand out clearly”. Similarly Aston (1995), points out that concordancing can highlight patterns of repetition and variation in text, thus favouring the analysis of larger and more specific schemata into smaller and more general ones, or else the opposite process, the synthesis of smaller and more general schemata resulting in larger but more specific ones.

To take a well-known example, Swales’ (1990) CARS (Create a Research Space) Model may be viewed as providing a large schema (i.e. accounting for a substantial chunk of discourse) which is however restricted in its application, or “specific”: virtually only contemporary research article introductions in English are likely to set off by “establishing a territory” (e.g. *The study of x is an important aspect of...; it has been claimed that y*), then “establishing a niche” (e.g. *These studies, however, suffer from x*) and finally occupying it (e.g. *This paper argues that y*). Yet if we deconstruct (“analyse”) this large, specific schema, and take its steps one by one, we may find that we are dealing with smaller schemata appropriate to other instances of discourse, say research article *conclusions*, which are therefore more general in scope. These suggestions apply equally well to phraseological regularities and idioms. As claimed by Danielsson (2001:97) “as the units [...] get longer on the syntagmatic scale, the paradigmatic choices tend to get fewer”. On a similar vein, Cignoni et al. (2002:129) discuss a common type of idiom variation, which consists in making them more specific by the addition of words that link them to the context (thus producing, for instance, *toe the education authority line* from the more general *toe the line*).

In general terms, the suggestions relating to analysis and synthesis of linguistic and situational schemata discussed above would appear to be in agreement with Sinclair’s work on the lexical item (e.g. 1996) as a unit of analysis, showing patterns of variable context-specificity/generalizability and clear form-meaning correlations. They are also consistent with current applied linguistics approaches which see processes of analysis and synthesis as lying at the basis of knowledge restructuring. This can be defined as “willingness and capacity [...] to reorganize [one’s] underlying and developing language system, to frame and try out new hypotheses and then act upon the feedback which is received from such experimentation” (Skehan 1996b:22). Below (3.) I shall

suggest that not only restructuring but also fluency and accuracy, the goals of language education in Skehan's approach, may gain from experience of corpus work. Knowledge restructuring in particular may be encouraged through the combined use of reference corpora and specific corpus typologies, such as "translation" and "learner" corpora. Let us turn to consider what these are and what role they might play in a foreign language classroom and/or in a translation classroom.

2.3 Learner and translation corpora for language learners and translation students

Learner and translation corpora have been used in language and translation classrooms with encouraging results. Learner corpora, which contain samples of learner writing alongside comparable samples (by text type and age) of native speaker writing, for instance, have been used to develop writing CALL (Computer Assisted Language Learning) software (Milton 1998) and to develop materials and activities for use in the ELT classroom (Granger & Tribble 1998). The assumption behind these attempts is that the learning process may be aided by form-focused instruction and access to focused negative evidence. In other words, if learners are presented with concordances showing the typical errors they (statistically) appear to make, and with similar textual environments where the same structure is used appropriately, they may find it easier to become aware of more or less fossilized characteristics of their interlanguage, thus potentially initiating a process of knowledge restructuring. Though doubts have been raised in the past as to the role played by negative evidence in Second Language Acquisition (see e.g. the pessimistic conclusions reached by Schwartz & Gubala-Ryzak 1992:35), this contrastive, form-focused approach provides an interesting alternative or addition to *standard* DDL, and develops the idea that *authenticity* may be a condition of the learner's engagement with a text, or the perception that a text is somehow relevant to her concerns (see Widdowson 1991, 1992, and 2000 on authenticity and for a critique of pedagogic corpus use). We shall go back to a general discussion of this point below. From the viewpoint of learner corpora in the classroom, an even more radical attempt at bringing together the concerns of the learners and more traditional corpus-aided language learning is described by Seidlhofer (2000a). Her starting point is a view of language learning as an intertextual activity, since "we access any text we come across via our knowledge of other, previously encountered texts, in a continual process of reconstruction of our individual and social realities" (ibid.:211). In this view, the emphasis is not so much on mistakes, and the

implicit recognition of superiority of a native variety is absent. In groups, learners compare each other's ways of carrying out a language-based task, discuss procedures and results, and come up with questions they would like to raise. A solution to these is then searched for in a reference corpus. This approach has a number of advantages: corpora are used to answer learner-generated questions, thus ensuring motivation; a view of language use as inherently intertextual is brought home to the learners, relieving them of the processing efforts of composing utterances from scratch; errors and conformance to a *target* norm (whose status is more and more under scrutiny) are de-emphasised (more on this issue below).

Translation corpora have also become important instruments in the education of translators. Parallel corpora (Source Text – Target Text corpora) can act as expert systems, drawing the learner's attention to (un)typical solutions for typical problems found by *mature*, expert translators. If one views translation education as a process of “acculturation, [...] of becoming increasingly proficient at thinking, acting and communicating in ways that are shared by the particular knowledge communities of which we are striving to become members” (Kiraly 2000: 4), the relevance of parallel corpora becomes evident.

On the other hand, bi- or pluri-lingual comparable corpora (collections of texts in more than one language, usually assembled on the basis of their text-type and content) have proved invaluable sources of information about typical turns of phrases, collocations, terms and their lexico-syntactic environment etc., resulting in translated texts which *read well*, conforming to the norms of the target language discourse communities (Gavioli & Zanettin 2000). Educating learners to use comparable corpora as reference tools in their everyday activity may result in better-documented, more accurate as well as more fluent translations. Since corpora may need to be (re-)assembled for each specific translation project, a number of researchers have emphasised the importance of DIY corpus construction skills (Maia 2000; Varantola 2003; Zanettin 2002). Apart from the direct effect of teaching learners to develop their own reference tools, the activity of corpus construction has also been found to have consciousness-raising effects of wider import (Maia 2000), as learners appreciate the problems involved in such operations as text selection, sampling, OCR, encoding etc., thus becoming better corpus users as well as more careful text analysers.

The development of bi-directional corpora (like the English-Norwegian Parallel Corpus, in which English originals and Norwegian translations are matched with comparable Norwegian originals and English translations) is an attempt to increase the consciousness-raising function of translation cor-

pora further (Bernardini 2002). By allowing learners to carry out comparisons of (1) original and translated language; (2) source texts and target texts; (3) comparable sets of bilingual (sub)corpora, bi-directional corpora may provide rich and varied stimuli for research, appealing to students interested in corpus linguistics, literary and translation studies and so forth. More importantly perhaps, a modular, flexible resource may highlight the operation of norms at different levels of specificity, thus favouring the observation of schemata and the evaluation of their applicability to different settings (Aston 1995, see 2.2 above).

This point applies equally well to translation and LSP teaching, which will be taken up in the following sub-section.

2.4 Learning LSP with corpora

ESP teachers were among the first to appreciate the pedagogic potential of corpus work. In line with the objectives set out in the introduction, here we shall not discuss work focusing on descriptively-oriented pedagogic issues such as syllabus development, specialised corpus construction and analysis and so forth (see e.g. Flowerdew 1996; Tribble 1997).

As we have seen, classroom concordancing with ESP students has been argued to be particularly promising because it highlights context-bound regularities, favouring the formation of large specialised schemata. Furthermore, it may provide learners with the cognitive and technical capacities required for text and corpus analysis, arguably among the most valuable objectives of an ESP course. In her discussion of the use of LSP corpora in (specialised) translation and interpreter education, Gavioli suggests that these corpora can be used to teach students to interpret instances of language production as *samples* rather than *examples*, “identifying recurrences and inferring patterns which appear in some way typical of certain contexts” (2000: 129). This involves developing a *researcher* attitude towards data, rather than trusting unquestioningly the authority of the teacher. Since students are expected to act as *participants* in discourse as well as discourse *observers* (Gavioli & Aston 2001), the observation of typical ways of organising language within particular genres can easily be “authenticated” through use, to adopt Widdowson’s terms (1984: 218). In other words, students browse corpora in search of information they require to complete a communicative task, analyse the results, choose a solution that appears to satisfy their needs, and adapt it to these.

A similar point is made by Mparutsa et al. (1991: 130) who found that

the experience of using the concordancer [...] challenges the role of a set text in the learning process. The text shifts from being an inviolable authority to something which students can question, explore and hopefully come to understand.

This, it is suggested, seems especially important in those educational settings where taking responsibility over one's own learning is traditionally not encouraged by teachers. Whilst this effect of DDL seems valuable in most fields of study, Mparutsa et al. found different focuses of attention suggest themselves in the course of activities in different areas. In "English for Economists", the focus was on terminology and lexicon, in "English for Geologists" on how to process information, and in "English for Philosophers" on patterns of categorisation and cohesion in texts. In other words, DDL has been found to operate not only at the formal level, on the surface of texts, but also at a deeper, conceptual level. Critical discourse analysis along the lines of Stubbs (1996, 2001) may prove educationally appropriate for language learners with a specific interest in a given knowledge domain, providing opportunities for (a) interaction, (b) observing the conventions operating in the field, and (c) developing capacities for text- and corpus-analysis:

[...] the possibility of student-tutor discussion of citations, where the student can contribute his/her developing subject knowledge and the tutor can contribute knowledge of language functions, can give a sense of joint discovery leading to illumination of the text. (Mparutsa et al. 1991: 131)

Analysis of text using a concordancer is not merely automatic. Users must make decisions at all stages of the process [...]. Concordancing is therefore in no way a substitute for critical thinking, but rather a tool which can be used investigatively, to enhance the interpretative power of the scholar (Kowitz & Carroll 1991: 135).

3. Discovery Learning (DL) or "The learner as traveller"

Building on the insights described above, I have proposed an approach to learning from corpora in which learners are guided to browse large and varied text collections in open-ended, exploratory ways. The view of 'learning as discovery' is easily and profitably adaptable to a corpus environment, thanks to the richness of the data and the endless possibilities offered by software programs which are more and more often designed with learners in mind. Whereas the *learning as research* approach favoured by Johns (1991) and Gavioli (2000)

implicitly assumes that learners share the same interests, competences and capacities as (adult) teachers or linguists, the *learning as discovery* view makes no such claim. Instead, it encourages learners to follow their own interests whilst providing them with opportunities to develop their capacities and competences so that their searches become better focused, their interpretation of results more precise, their understanding of corpus use and their language awareness sharper. This may be confusing at first, as learners are asked to abandon deeply rooted norms of classroom behaviour, but soon becomes liberating for both teachers (who can stop pretending to be sources of absolute and limitless knowledge) and learners (who start to see themselves as active participants in the teaching-learning process).

For a number of years I have tried out this approach with advanced learners of English in their last year of studies as undergraduates at the School for interpreters and translators of the University of Bologna at Forlì (Bernardini 2000, 2002). The response has always been very encouraging, despite a certain resilient technophobia among students. Those who accept the challenge are shown how to use the British National Corpus (BNC) with its interrogation software, SARA (Dodd 2001), and a number of other resources available on the Faculty local network, including parallel and comparable corpora in various languages, using WordSmith Tools (Scott 1996). As they build up experience in choosing resources, designing queries, interpreting results, and so forth, they are progressively given more and more freedom. At the end of the course, they are asked to carry out a self-initiated project involving corpus browsing, whose results *and* strategies will be discussed in class. The most unusual aspect of this assignment is a recommendation to follow up new strands of research that might suggest themselves in the course of their work, or to make note of them for future use. Encouraging a student to let irrelevant issues distract her from her work is not a common attitude in the Italian school system, and learners need some time before they convince themselves this is actually acceptable. Hopefully, in time they appreciate that discoveries are often made when least expected, and that *serendipitous* findings may be rewarding and encouraging in (language) learning.

Let us look in more details at the kind of work learners may be faced with. A typical first day activity may require participants to use the BNC to interpret the meaning of a rather obscure and elliptical newspaper article headline such as the following:

Blair hailed for staunch support of America (*Washington Times*, electronic edition, 06/11/01).

Problems raised by this short headline include:

1. Deciding whether *hailed* is a past participle or simple past form (ambiguity due to copula ellipsis causing problems to many).
2. Realising that a search for *hailed* will only retrieve instances of this word form, not other word forms belonging to the lemma *HAIL*.
3. Disentangling different patterns/meanings by sorting solutions, grouping relevant ones and discarding irrelevant ones (e.g. instances of the colligation *HAIL* + *as* + NP were grouped together as relevant (cf. Figure 2), whereas instances of the semantic preference *HAIL* + *means of public transport*, usually *cab* or *taxi* were noted as interesting but then discarded as irrelevant to the present analysis).
4. Observing a similarity of meaning between patterns such as *hailed as supporter* and *hailed for support*.
5. Grouping collocates of relevant solutions according to common traits (e.g. noticing that the nouns *epitome*, *hero*, *success*, *sensation*, *lords* and the adjectives *new*, *historic*, *great*, *dominant*, *major* etc. appearing in the co-text of “hailed as” are all emphatic, appreciative words, cf. Figure 2).
6. Deciding how common is *staunch* as a modifier of *support(er)* and its synonyms and/or antonyms (see Table 1 for a list of the ten most frequent collocates of the adjective *staunch* in a span of ± 4 words)
7. Wondering what other nouns and adjectives are typically modified by *staunch* (the reference to political and religious matters is obvious in collocates such as *Marxist*, *Methodist*, *Monarchist*, *Royalist*, *Thatcherite* and so forth, see Table 1 and Figure 1).
8. Identifying occurrences that show clear similarities with the pattern under study, and which should therefore be particularly focused upon (cf. the following solution identified as ‘relevant’ by one learner; “When Iraq’s tanks rolled into Kuwait last August, the Moroccan king again proved himself a *staunch friend* to America and Saudi Arabia.” *The Economist*, BNC-ID: ABE). This pattern-matching ability is fundamental to corpus analysis, but also, arguably, to language learning and communication in general (Beaugrande & Dressler 1981).
9. Reflecting on the text typological restrictions associated with such lexico-syntactic observations, and generally on the usefulness of linguistic patterns for inferring the typology a text may belong to.
10. Defining a more relevant sub-corpus (according to text type, e.g. “newspaper texts”, or according to domain, e.g. “texts about world affairs”) in which to conduct further queries.

ase. He was such a loyal, staunch and tender-hearted friend of my family, and he sink there I built a good staunch bench [...] about like that square, put a vision's daughter Mary, was a staunch Catholic. Consequently, during the last month Tait and Stewart were staunch churchmen; their book set out to show that Mont and New Hampshire staunch conservative New England states that are traditions my family have been staunch Conservatives, but I'm afraid my immediate ssable... 'Eliot, then, had a staunch defender. |
 ng a burger.' | Mitterand, a staunch defender of French culture, may be reluctant in the late 1920s and as a staunch Nazi supporter he had enjoyed rapid promotion. President Bush was a staunch opponent of abortion under all but the most a Mr Hanmer who was a staunch Royalist. When she was a girl of eight, she ework, her writings reveal staunch Royalist views and a distinctly Anglican Reative years) he has been a staunch servant through thick and thin. A career ave ra's and was to become a staunch source of support to her over the years. One ave Theosophy and I am a staunch supporter. I join the Liberal Catholic Church for Spelthorne, last year a staunch supporter on the Commons committee ex hearted support. | Another staunch supporter is Wulstan Atkins, Elgar's godson about Party candidate, is a staunch Thatcherite, sometimes justly described as s is advantages. If you're a staunch union member there is advantages. With these and clubs and have four staunch volunteers and three or four helpers who la

Figure 1. 20 randomly selected occurrences of *staunch* as an adjective in the BNC (sorted to the right)

university college has been hailed as a boost for the area by Education Secretary (1986). This has been hailed as a key to managing increasingly complicated class constructor and been hailed as a Brazilian Ferrari or Chapman. In fact, incess. The decision's being hailed as a legal milestone. Nick Clark reports. don prison, near Bicester is hailed as a prison of the future. Members of the same fashion then promptly hailed as a conquering hero when his team carried court date. || A DRIVER was hailed as a hero last night after helping rescue earlier. The legislation was hailed as a cautious first move by the Saudi government. ish flag of convenience was hailed as a lucrative alternative, beneficial to the nation in autumn 1989, it was hailed as a bold and novel decision through which of a foreigner already being hailed as an England batting hero long before his her's leadership was rightly hailed as an inspiration worldwide. How Mr Lawser Edward Elgar was once hailed as England's answer to Beethoven. But a me. | Although the move was hailed as sensational at first sight, the vague negotiations. Incredibly, he was also hailed as the saviour of the Conservative Party in gy. The agreement has been hailed as the first of a series intended to tackle 15th January). She has been hailed as the pioneer of a literary genre for the 'usk of the coconut, is being hailed as the new alternative to peat, it has manifested the results of a large survey, hailed as the 'Italian Kinsey report'. In which he hich could scuttle what was hailed as the most significant arms deal reached

Figure 2. 20 randomly selected occurrences of *hailed as* in the BNC (sorted to the right)

Table 1. The ten higher scoring collocates of the adjective *staunch* in the BNC (z-score order, span ± 4 , only words occurring more than 3 times in the collocation range)

Collocate	TAG	Frequency	Z-Score
royalist	SUBST	8	176.7
supporter	SUBST	36	113.5
methodist	SUBST	4	55.4
advocate	SUBST	7	46.4
defender	SUBST	8	38.8
flow	SUBST	13	37.1
ally	SUBST	8	34.5
friend	SUBST	20	23.0
opponent	SUBST	6	22.5
conservative	SUBST	5	18.7

From this guided, “convergent” task, learners soon go on to more autonomous, “divergent” ones (in Leech’s (1997) terms) typical of discovery learning experiences.

To give just one example, while carrying out a translation into English of a European report on youth policies a learner wondered what the difference(s), if any, might exist between the two plural forms of the noun *competence* (*competences*|*ies*), and in what cases the noun *skill* might be more appropriate as a translation equivalent for Italian *competenza*. She began by categorising the collocates of both nouns in an attempt to identify common traits that might lead to hypotheses as to which term is used in what cotext/context, and subsequently tried to restrict the search to “social science” and “world affairs” texts, where less technical uses would be less likely to clog up the concordance.

At this stage, while no final answer for her question had been found, she had had some experience of genuine occurrences of each term in context, and occasions to reflect on them, having paged through the solutions many times, formulating, testing, and revising hypotheses. More interestingly, a number of curious words and new structures, often unknown, were present in the concordance outputs, and these provided subjects for further searches and discussions with the rest of the class.

One of these, the word *foibles*, led to a search for *foible*|*foibles*, and from there to further unknown expressions, such as *true-blue*. A search for this word suggested two related meanings, a more general one, meaning “marked by unswerving loyalty”, and a more specific one referring to Conservative supporters or politicians. It was then hypothesised that the colour *blue* refers to Conservative party members and consequently, that the colour *red* may refer

to Labour party members. Further queries supported this hypothesis, in some cases offering glosses, as in “the Hon. Samuel, of Slumkey Hall, successful Blue (Tory) candidate in the Eatanswill election”, provided information about the world as well as the language (more or less figurative references to the *red rose*, to *Mr Kinnock*, and to accusations that the latter had stolen the emblem of the Duchy of Lancaster. . .) and led to identifying a number of expressions that seemed typical of electoral propaganda. Some of these were parts of short letters to a newspaper or journal editor, starting “Sir – . . .”. A sub-corpus of these texts was defined, so as to be further analysable in class, with the aim of determining the structural and lexico-syntactic patterns associated with this “text type” (see Figure 3).

Defining a sub-corpus through the specification of required lexico-syntactic structures, though rather unusual as a classroom activity, and somewhat controversial as a heuristics for corpus construction (see e.g. Sinclair submitted), is in line with current work on text typologies (see e.g. Biber et al. 1998), and may be conducive to the development of the capacities needed for the construction of web-based DIY corpora (see Varantola 2003; Zanettin 2002), an important aspect of translator’s and interpreter’s professional expertise.

Work of this type would appear to be coherent with the views on language learning outlined above, and conducive to similar results relating to knowledge restructuring, critical autonomy, researcher skills, language awareness, opportunities for communicative interaction and so on (see Section 2 above). Furthermore, there seem to be a number of further advantages:

- Learners are encouraged to become more autonomous in their studies, taking responsibility for their own learning. Discovery learning activities are designed to favour learner-centred, open-ended, tailored learning. These qualities, according to Leech (1997: 11–12), “are fully realized only where the program is fully adaptable to the learner’s individual needs and preferences [,] 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.” The importance of autonomy and self-direction in language learning is nowadays widely recognised as an important objective and guiding principle in language pedagogy (cf. e.g. issue 23,2 of the journal *System*, dedicated to this theme (1995)). It is all the more important when one of the aims of instruction is to prepare students to go on learning the language autonomously according to their professional (or other) needs. This seems to be the spirit, for instance, of a recommendation of the Council of Europe suggesting that “language pedagogy [...] should [...]

develop explicit objectives and practices to teach methods of discovery and analysis” (Kettemann 1996).

- In turn, autonomy and responsibility are arguably conducive to increased motivation to learn and consequently to increased learning effectiveness (see Dickinson 1995 for a review of the literature supporting this claim).
- A supportive, non-authoritarian environment is created: the teacher is not artificially setting up tasks requiring learners to provide information that she already has; rather, everyone in the classroom is actively trying to find the solution to a problem, discussing a solution proposed by one of the participants, guessing at the meaning of an expression, and so forth. In this framework, the teacher acts as a learning expert rather than a language expert.
- Discovery learning is not only empowering for learners, but for teachers as well, especially if non-native speakers of the language they teach. Being life-long language learners as well as teachers, they possess an invaluable repertoire of learning strategies and experience of difficulties and successes that students can draw from, whilst their limited intuitions concerning acceptability and appropriateness are less crucial a problem than they used to be. For this reason, among others, I believe that corpus-based discovery learning can facilitate a process of democratisation of the learning setting, contrary to the fears of a number of applied linguists (e.g. Widdowson op.cit.; Cook 1998, more on this issue below, Section 4.).
- Discovery activities require learners to focus on form as well as meaning, and provide a learning environment where noticing the correlations between the two (i.e. that different patterns are associated with different meanings, Sinclair 1996) is facilitated (cf. the example of *hail*, in which different senses could be disentangled by way of reference to different collocational and colligational patterns). They also encourage learners to link observation and participation in discourse, allowing them to discuss findings in pairs or small groups before undertaking more structured written or spoken reports. The value of post-task activities involving public performance is highlighted by Skehan (e.g. 1996a), who claims that such activities can infiltrate a concern with syntax and analysis into the task, “reminding learners that fluency is not the only goal during task completion, and that restructuring and accuracy also have importance” (ibid.:55/56). Thus, the three goals of language learning in Skehan’s framework (*accuracy*, *fluency*, and *restructuring*) would appear to be coherent with activities of corpus-aided discovery learning.

| SIR – I am distinctly uneasy about Peterborough’s account of Emma Te
 | Sir, – It saddened me to see Mr Reg Cleaver describe the Jews as ‘an
 | Sir, – The new Mayor of Woodbridge, Mr Tony Hubbard, is encouraging
 | Sir, – I write to you in my capacity as president of the Institut des Revise
 | Sir, – Ian Luder’s letter in your November issue (p 6) does not consider
 | Sir – I wish to comment on the letter from (Nature 360, 704: 1992) on
 | Sir, – ‘It is proposed to retain the Ipswich airport as a two runaway, grass
 | Sir, – As the last senior partner of the Dearden Farrow, I was interested to
 | Sir, – I have read with great interest the recent reports and subsequent
 | Sir, – The Chancellor’s proposal to add VAT to domestic fuel bills has
 | Sir, – I was interested in your article about Channel 7 (see ACCOUNTAN
 | Sir, – I write to draw attention to an inaccuracy in the headline and the fi
 | Sir, – I am indebted to Mrs Swindin for her lucid article of May 10, in whi
 | SIR – The late, great Mr. X, of whom there has never been a more acute
 | Sir, – In a letter to the EADT (March 18), Andrew Blake defends animal e
 | Sir! – It’s only £599! One for the road – we road test a GriD laptop in th
 | Sir: – I note the new unpriced first-class postage stamps are black. Are the
 | Sir: – Now that the dust kicked up by the mass raid on the Broadwater Far
 | Sir, – [Tutorial Programs in Phonetics & Linguistics] I was formerly a Lect
 | Sir: Your contributor William Rees-Mogg identified his list of 50 ‘masters

Figure 3. Extract from a Concordance for *Sir* in a sub-corpus of letters

4. The past, and the future

The views on corpus use in the classroom discussed in the previous sections not only show how, in the last ten years, teachers and applied linguists have become more and more interested in the corpus linguistics approach. They also suggest that descriptive insights and research methodologies have not simply been borrowed from the descriptive paradigm, but have been adapted, reformulated and often extended in various ways to fit pedagogic concerns and priorities, whilst preserving the most interesting and innovative aspects (e.g. the operation of the idiom principle, the role of collocational phenomena in structuring and interpreting discourse, the links between lexico-syntactic and situational/discoursal/text-typological observations and so forth).

The uses of corpora described here are far from prescriptive and “integralist” (Cook 1998), although many of them rely on native speaker language performance. They in no sense prescribe that learners imitate “the most usual, the most frequent or, in short, the most clichéd expressions” (ibid.). I think, however, that most of the researchers and teachers whose work has been referred to here would agree that it is important for their students to (learn to) *understand* these expressions and reflect on their implications in the context of a given discourse setting. The language and learning awareness that discov-

ery learning may favour, also through the observation and evaluation of native norms are a prerequisite for autonomy and assertion, and constitute an antidote against uncritical submission to those same norms. As claimed by Sinclair, in corpus-inspired pedagogy “rules are not restrictive, they are not “do not” rules”; they are “try this one” rules where you can hardly go wrong. There is an open-ended range of possibilities and you can try your skill [...] trying to say what you want to say” (1991b: 493).

The current debate on (the teaching of) English as an international language is rightly, I think, questioning the status of the idealised “native-speaker” as a target model and stressing the need for non-colonising attitudes, that stir clear of acculturation practices. Corpus access in the language classroom may be a powerful tool in this sense, since it allows observation of instances in which a norm has been respected, and others in which it has not, resulting in ironic, creative, dissonant effects, or in a misunderstanding. The ease of access to instances of language performance makes it possible for learners to rely less on one or two individuals with their idiosyncracies and their limited intuitions. If they can also work with corpora in their native language, this may convince them of the unreliability of their own intuitions about their mother tongue, resulting in a heightened attention to (un)typical ways of saying in any languages they know. Lastly, corpora of English as an international language are also seeing the light in Austria, Finland and Spain, with the aim of collecting and making available “unscripted [...] communication among fairly fluent speakers from a wide range of first language backgrounds whose primary and secondary education (and socialization) did not take place in English” (Seidlhofer 2000b).³ Accusations of linguistic imperialism are therefore, I would suggest, very wide of the mark, if one considers the theoretical insights and practices described in this paper.

A second aspect which seems increasingly to have come to characterise corpus use in the classroom in the last few years, is the interaction between corpora and web-based learning and CALL environments. This is a relatively recent but fast growing phenomenon, coherent with the autonomising and non-authoritarian approach to classroom concordancing described above, reflected in (web addresses in appendix):

- The multiplication of web-based concordancers (e.g. *WebCorp*, *KWICFinder*, and *WebKWIC*).
- The integration of concordancing in more complex environments, such as the lexical database *Wordnet* or, more interestingly for our concerns, the Hong Kong-based Virtual Language Centre *Web Concordancer* for Chinese,

English, French, and Japanese, and the University of Montreal-based *Complete Lexical Tutor*, a set of tools which include, among others, frequency analyses and vocabulary profiles of any texts, vocabulary tests and reading and listening facilities for English and French. The latter are integrated with concordancing software and easy access to *WordNet*, so as to facilitate reading/listening comprehension and acquisition (Cobb et al. 2000).

- The development of corpus-based grammars and tutorials accessible online (e.g. *SEU* and *Chemnitz Internet grammars*) and the possibility of accessing a range of corpus resources on line (the *W-3 corpora project* at Essex, the impressive collection of corpora at the *Institut für Deutsche Sprache*, Mannheim, the *Translational English Corpus* (TEC) at UMIST, the *English Norwegian Parallel Corpus* (ENPC) at Oslo University to name but a few; a more extensive list of links is provided by Barlow (online)).

Thirdly, the appearance of academic articles and conference papers investigating the effectiveness of concordancing with language learners and the interaction between activities, strategies and learning outcomes seems finally to be filling a significant gap. Cobb (1997), for instance, finds that the efforts of using concordances to work out the meanings of new words appear to result in a gain in the ability to transfer word knowledge to novel situations. This finding is consistent with views of vocabulary learning as being influenced by the processing demands of particular activities and by the processing strategies adopted (Robinson 1995). Bernardini (2000) and Kennedy and Miceli (2001), on the other hand, discuss common errors made and strategies adopted by learners and suggest ways in which these might be limited/optimised. Interestingly, though the groups of students described in these articles come from different language backgrounds and are at different language proficiency levels, similar conclusions are reached regarding the need for careful guidance and attention to the development of corpus-investigation skills, especially at early stages. Language and learning awareness are thus confirmed to be among the key concerns of data-driven learning.

5. Conclusion

How can corpora and corpus linguistics aid language learning and teaching, then? In this paper I have suggested that their potential may reside not only in the descriptive insights corpora give access to. More importantly (albeit perhaps less obviously), corpora and corpus analysis tools would seem to provide

one of the most powerful tools made available to date for classroom discovery learning activities. In line with current views of language learning and teaching, such activities are designed to engage the learners' interest, to be motivating – thus conducive to autonomy – to focus attention on both form and meaning, to provide opportunities for (in)formal interaction, to encourage the setup of a relaxed atmosphere and so forth (see Section 3 for details). Although discovery learning is in no sense inherently corpus-based, and undoubtedly pre-dates the entrance of corpora into the classroom (and corpus linguistics itself, at least in its modern sense), corpora and corpus-inspired views of language and linguistics have been facilitative and instrumental in setting up and developing this approach to language learning.

In this paper I have attempted to retrace the development of soft and hard data-driven/discovery learning approaches, the rationale behind them, some relevant corpus typologies and applications, particularly in the fields of LSP and translation teaching, with their early appreciation of the combined potential of corpus tools – as sources of linguistic insights and as stimuli for discovery learning.

Recently, technological improvements (huge hard disks, fast processors, the WWW, to name but the first that come to mind) have made corpora (and corpus-based learning tools) easier and cheaper to construct, download, buy and/or access online. More importantly perhaps, *corpus* has become less of a buzzword and more of a necessary, acknowledged reference source for students, linguists, language professionals (teachers, translators, technical writers, lexicographers etc.). As a consequence, discovery learning is now a workable option for many teachers, that can easily be adapted and made to appeal to most students, not necessarily very advanced ones or language specialists.

One last note of caution. Today, many of us can proudly tell our students, eager to consult corpora, to 'go to any of the labs, hit the icon which says "Corpus" and follow the instructions on the screen' (Fligelstone 1993:101). This is no doubt a very welcome improvement to our teaching settings, provided that self-access for reference purposes is seen as part of a learning experience, rather than the sole mode of access to corpora. This is because learners require guidance and heightened awareness to learn from corpora and much of their potential (for strategic learning, serendipity, reasoning-gap, as well as for stimulating communicative activities) would be lost if learners did not have a chance to carry out relatively complex analyses, requiring them to observe phraseological regularities and restrictions, and the functions associated with them, and to share their opinions and findings. In other words, corpus-learner, and indeed corpus-teacher interaction are not replacements for learner-learner

and teacher-learner interaction, but rather should be seen as an added value offered by corpus-aided discovery learning.

Notes

1. Here reference is made to the title of the TALC 1998 proceedings edited by Burnard and McEnery, *Rethinking Language Pedagogy from a Corpus Perspective*.
2. Cf. also the point made by Aitchison (1994) in her (non-corpus-based) discussion of two senses of the adjective *old* (namely *former* and *mark of affection*). These, she claims, may be distinguished mainly by way of reference to their typical lexico-syntactic patterns, i.e. a possessive pronoun would seem to be associated more with the *mark of affection* sense, as in *my old friend Tom*, than with the *former* sense).
3. There are signs that the status of International (or Lingua Franca) English and the problems involved in its representability and describability through corpora are moving from the periphery to the core of corpus linguistics concerns. See on this subject the discussion in the *corpora* list archive during the months of November and December 2001 (<http://www.hit.uib.no/corpora/2001-4/>).

References

- Aitchison, J. (1994). *Words in the Mind*. Oxford: Blackwell.
- Aston, G. (1995). Corpora in language pedagogy: Matching theory and practice. In G. Cook & B. Seidlhofer (Eds.), *Principles and Practice in Applied Linguistics: Studies in Honour of H. G. Widdowson* (pp. 257–270). Oxford: Oxford University Press.
- Barlow, M. (1996). Corpora for theory and practice. *International Journal of Corpus linguistics*, 1(1), 1–37.
- Bartlett, F. C. (1932). *Remembering: A Study in Experimental Social Psychology*. Cambridge: Cambridge University Press.
- Beaugrande, R.-A. de & W. U. Dressler (1981). *Introduction to Text Linguistics*. London: Longman.
- Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 225–234). Frankfurt am Main: Peter Lang.
- Bernardini, S. (2002). Serendipity expanded: Exploring new directions for discovery learning. In B. Kettemann & G. Marko (Eds.), *Teaching and Learning by Doing Corpus Analysis. Papers from the Fourth International Conference on Teaching and Language Corpora, Graz 19–24 July 2000* (pp. 165–182). Amsterdam: Rodopi.
- Biber, D., S. Conrad, & R. Reppen (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, D., S. Johansson, G. Leech, S. Conrad, & E. Finegan (1999). *The Longman Grammar of Spoken and Written English*. London: Longman.

- Burnard, L. & T. McEnery (Eds.) (2000). *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt am Main: Peter Lang.
- Cignoni, L., S. Coffey, & R. Moon (2002). Idiom variation in Italian and English. *Languages in Contrast*, 2(2) (1999), 119–140.
- Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System*, 25(3), 301–315.
- Cobb, T., C. Greaves, & M. Horst (2000). Can the rate of lexical acquisition from reading be increased? An experiment in reading French with a suite of on-line resources. In P. Raymond & C. Cornaire (Eds.), *Regards sur la Didactique des Langues Secondes*. Montréal: Éditions Logique. Online: <http://132.208.224.131/ResearchWeb/> consulted: 25.11.2003.
- Cook, G. (1998). The uses of reality: A reply to Ronald Carter. *ELT Journal*, 52(1), 57–64.
- Danielsson, P. (2001). The Automatic Identification of Meaningful Units in Language. PhD dissertation. Gothenburg University.
- Dickinson, L. (1995). Autonomy and motivation – A literature review. *System*, 23(2), 165–174.
- Dodd, A. (2001). *Sara 0.98*. Oxford: Oxford University Computing Services.
- Fligelstone, S. (1993). Some reflections on the question of teaching from a corpus linguistics perspective. *ICAME Journal*, 17, 97–109.
- Flowerdew, J. (1996). Concordancing in language learning. In M. Pennington (Ed.), *The Power of CALL* (pp. 97–113). Houston, TX: Athelstan.
- Gavioli, L. (2000). The learner as researcher: Introducing corpus concordancing in the classroom. In G. Aston (Ed.), *Learning with Corpora* (pp. 108–137). Houston, TX: Athelstan / Bologna: CLUEB.
- Gavioli, L. & G. Aston (2001). Enriching reality: Language corpora in language pedagogy. *ELT Journal*, 55(3), 238–246.
- Gavioli, L. & F. Zanettin (2000). I corpora bilingui nell'apprendimento della traduzione. In S. Bernardini & F. Zanettin (Eds.), *I Corpora nella Didattica della Traduzione. Corpus Use and Learning to Translate* (pp. 61–80). Bologna: CLUEB.
- Granger, S. & C. Tribble (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on Computer* (pp. 199–209). Harlow: Longman.
- Johns, T. (1991). Should you be persuaded – Two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom Concordancing* [ELR Journal, 4] (pp. 1–16).
- Kennedy, C. & T. Miceli (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning and Technology*, 5(3), 77–90.
- Kettemann, B. (1996). "Concordancing in English Language Teaching". Online: <http://www-gewi.kfunigraz.ac.at/ed/project/concord1.html>, consulted: 25.11.2003.
- Kiraly, D. (2000). *A Social Constructivist Approach to Translator Education*. Manchester: St. Jerome.
- Kowitz, J. & D. Carroll (1991). Using computer concordances for literary analysis. In T. Johns & P. King (Eds.), *Classroom Concordancing* [ELR Journal, 4] (pp. 135–149).
- Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, A. M. McEnery, & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 1–23). London: Longman.

- Maia, B. (2000). Making corpora – A learning process. In S. Bernardini & F. Zanettin (Eds.), *I Corpora nella Didattica della Traduzione – Corpus Use and Learning to Translate* (pp. 47–60). Bologna: CLUEB.
- Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (Ed.), *Learner English on Computer* (pp. 186–198). Harlow: Longman.
- Mparutsa, A., A. Love, & A. Morrison (1991). Bringing concord to the ESP classroom. In T. Johns & P. King (Eds.), *Classroom Concordancing* [ELR Journal, 4] (pp. 115–134).
- Robinson, P. (1995). Attention, memory, and the ‘noticing’ hypothesis. *Language Learning*, 45(2), 283–331.
- Schwartz, B. & M. Gubala-Ryzak (1992). Learnability and grammar reorganization in L2A: Against negative evidence causing the unlearning of verb movement. *Second Language Research*, 8(1), 1–38.
- Scott, M. (1996). *WordSmith Tools*. Ver. 3.0. Oxford: Oxford University Press.
- Seidlhofer, B. (2000a). Operationalizing intertextuality: Using learner corpora for learning. In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 207–223). Frankfurt am Main: Peter Lang.
- Seidlhofer, B. (2000b). Towards the teaching of lingua franca English: The Vienna ELF Corpus. *TALC 2000 Pre-Conference Volume*. Graz: University of Graz.
- Sinclair, J. M. (1991a). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. M. (1991b). Shared knowledge. In J. E. Alatis (Ed.), *Georgetown University Round Table on Language and Linguistics 1991* (pp. 489–500). Washington, DC: Georgetown University Press.
- Sinclair, J. M. (1996). The search for units of meaning. *Textus*, 9(1), 75–106.
- Sinclair, J. M. (submitted). Corpus and text – Basic principles. Paper submitted to M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice*.
- Skehan, P. (1996a). A framework for the implementation of task-based instruction. *Applied Linguistics*, 17(1), 38–62.
- Skehan, P. (1996b). Second language acquisition and task-based instruction. In J. R. Willis & J. D. Willis (Eds.), *Challenge and Change in Language Teaching* (pp. 17–30). Oxford: Heinemann.
- Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.
- Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.
- Swales, J.M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.
- System (1995). *Autonomy, Self-direction and Self-access in Language Teaching and Learning* (Special issue 23:2).
- Tribble, C. (1997). Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching. In B. Lewandowska-Tomaszczyk & J. P. Melia (Eds.), *PALC '97. Practical Applications in Language Corpora* (pp. 106–118). Łódź: Łódź University Press.
- Varantola, K. (2003). Translators and disposable corpora. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora and Translator Education* (pp. 55–70). Manchester: St. Jerome.

- Widdowson, H. G. (1984). *Explorations in Applied Linguistics* 2. Oxford: Oxford University Press.
- Widdowson, H. G. (1991). The description and prescription of language. In J. E. Alatis (Ed.), *Georgetown University Round Table on Language and Linguistics 1991* (pp. 11–24). Washington, DC: Georgetown University Press.
- Widdowson, H. G. (1992). Communication, community and the problem of appropriate use. In J. E. Alatis (Ed.), *Georgetown University Round Table on Language and Linguistics 1992* (pp. 305–315). Washington, DC: Georgetown University Press.
- Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21(1), 3–25.
- Willis, J. D. (1996). A flexible framework for task-based learning. In J. R. Willis & J. D. Willis (Eds.), *Challenge and Change in Language Teaching* (pp. 52–62). Oxford: Heinemann.
- Zanettin, F. (2002). DIY corpora: The WWW and the translator. In B. Maia, J. & M. Ulrych (Eds.), *Training the Language Services Provider for the New Millennium* (pp. 239–248). Porto: Universidade do Porto.

Appendix

Web addresses (consulted: 11.11.2003):

<i>WebCorp</i>	http://www.webcorp.org.uk/
<i>KWiCFinder</i>	http://miniappolis.com/KWiCFinder/KWiCFinderHome.html
<i>WebKWIC</i>	http://miniappolis.com/WebKWiC/WebKWiCHome.html
<i>WordNet</i>	http://www.cogsci.princeton.edu/~wn/
<i>Web Concordancer</i>	http://www.edict.com.hk/concordance/
<i>Complete Lexical Tutor</i>	http://132.208.224.131/
<i>SEU Grammar</i>	http://www.ucl.ac.uk/internet-grammar/
<i>Chemnitz Internet Grammar</i>	http://www.tu-chemnitz.de/phil/InternetGrammar/publications/
<i>W-3 corpora project</i>	http://clwww.essex.ac.uk/w3c/
<i>Mannheim corpora</i>	http://www.ids-mannheim.de/kt/corpora.shtml
<i>TEC</i>	http://www.ccl.umist.ac.uk/staff/mona/tec.html
<i>ENPC</i>	http://www.hf.uio.no/iba/prosjekt/
<i>Barlow's CL page</i>	http://www.ruf.rice.edu/~barlow/corpus.html

In preparation

What teachers have always wanted to know – and how corpora can help

Amy B M Tsui

The University of Hong Kong

This chapter discusses how corpus evidence is used to address teachers' questions about English grammar and points out that corpus linguistics has a unique contribution to make in raising teachers' sensitivity to linguistic features and patterns. For nine years now, English language teachers in Hong Kong have been soliciting advice from language specialists on a website, *TeleNex*, with regard to difficulties that they have with English grammar. More than one thousand questions that they sent to the website over a period of seven years were analyzed and six major types of questions were identified. This paper focuses on three types. The first type is lexical items which, although synonymous, have different usage, and teachers have difficulties explaining the difference to students; also items which appear to be synonymous, and teachers have problems differentiating them. The second type relates to linguistic evidence that contradicts the prescriptive grammar rules that teachers had been taught when they were students. The third type is lexical collocations that teachers try to rationalize. When answering teachers' questions, corpus evidence was used by the language specialists supporting the website to explore answers with teachers. The use of corpus evidence in addressing these questions is not only powerful and convincing but also leads to discoveries of patterns and meanings which have not been attended to in the standard libraries or in common practice.

Corpora and linguistic description

Before the existence of corpora, linguistic descriptions relied very much on native-speaker intuition and introspection. They describe what people *know about* language, or what they *perceive* language to be, rather than how language *is used*. The easy accessibility of huge bodies of naturally occurring texts on the computer has made it possible for us to test the robustness of linguistic de-

scriptions which were based on introspection and elicitation, and to gain new insights into language structure and use. It has helped us to gain a better understanding of how language is actually used rather than how language is perceived to be used. Examining specific instances of language use gives us insights into how language works which would never have been obtained by simply introspecting about the language system. Such insights result in our construing the linguistic system in a different way.

So far, corpus-based studies have focused on four main types of description and analysis: lexical collocation by examining the frequency and context of occurrence of linguistic items (see for example, Sinclair 1991; see also Kjellmer's (1994) dictionary of collocations based on the Brown Corpus), syntactic patterning based on co-occurrence of grammatical word-class tags, genre analysis based on the co-occurrence of groups of linguistic items and processes (see for example, Biber 1988), and discourse structure and cohesion in spoken and written English (see for example, Carter and McCarthy's *Spoken English Corpus* at the University of Nottingham – e.g. Carter & McCarthy 1997). (See Kennedy 1998 for a summary of corpus-based studies.) The findings of the above studies, particularly word-based studies, have important implications for second or foreign language teaching, as we shall see in the next section.

Corpus-based studies and ESL / EFL teaching

In EFL and ESL situations, learners do not have the same amount of exposure to the target language as they do in L1 situations. Therefore, they are unlikely to acquire the language efficiently without systematic guidance on linguistic forms. By focusing on words which have a high frequency of occurrence and by concentrating on the usual rather than the exceptional, teachers can help learners acquire the language more efficiently, especially at elementary and intermediate levels. The findings of corpus analysis can be used as a basis for selecting and sequencing linguistic content, as well as for determining relative emphases.

A number of studies have observed discrepancies between corpus findings and the selection of and emphasis given to linguistic content in ESL and EFL textbooks and curriculum. As early as the sixties, George (1963, cited in Kennedy 1998:283) studied a corpus of English made in Hyderabad that was based on written texts and found that the highest frequency of occurrence of the simple present is not to indicate habitual or iterative actions, such as "I go to school by bus every day." (5.5%), but rather the actual present, such as "I

agree with you” (57.7%) or neutral time, such as “My name is Mary.” (33.5%). His findings converge with a more recent grammar of English compiled by Mindt (2000) based on corpora totaling 240 million words of spoken and written English. Mindt found that the three prototypes which make up the majority of all cases of the present forms of verbs are the extended present, the actual present and the timeless present. This is contrary to the emphasis given to the habitual present in most ESL and EFL textbooks as the major function of the simple present.

Holmes (1988) compared a corpus analysis and a textbook analysis of epistemic modality and found that, like most textbooks, important epistemic uses of modal verbs are under-taught and that lexical verbs expressing modality, such as *appear*, *believe*, *doubt*, and *suppose*, nouns such as *possibility*, *tendency*, *likelihood*, and adverbials, such as *perhaps*, *of course*, *probably*, tend to be given little pedagogical attention. Ljung (1991) compared the EFL textbooks at upper secondary level in Sweden with the Cobuild corpus and found that 20% of the most frequent one thousand words in the learners’ texts did not occur in the most frequent one thousand words in Cobuild. Biber, Conrad and Reppen (1994) examined the structural options for postnominal modification and the attention given to these options in popular grammar ESL and EFL textbooks. They found that typically more pedagogical attention was paid to finite and non-finite relative clauses than prepositional phrases as noun modifiers, in contrast with their analysis of the *LOB* corpus, which shows prepositional phrases as noun modifiers occurring far more frequently than relative clauses (see also Quirk et al. 1985: 1274). Kennedy (1998) observes that similar incompatibility can be found in the pedagogical focus on grammatical quantifiers such as “all” and “every” in many textbooks to indicate the concept of totality when in both written and spoken corpora totality is much more commonly lexically marked, such as *entirely*, *completely*, *whole*, *throughout*.

The above very brief summary of some comparative studies of corpora and ESL and EFL textbooks show the relevance of corpus studies to ESL and EFL teaching and learning. One of their major contributions is to provide objective quantitative evidence of the distribution of linguistic items on which the goals and content of the curriculum can be based.

Corpus analysis and teachers’ language awareness

One area that is under-explored is the relevance of corpus linguistics to teacher education, particularly in the area of teachers’ language awareness (see also

Allan 1999; Berry 1994; Hunston 1995). In the last decade or so, more attention has been paid to the importance of raising teachers' language awareness (see for example the collected papers in Bygate et al. 1994; Hawkins 1999; James & Garrett 1991).

This chapter hopes to emphasise that teachers' language awareness is one area in which corpus linguistics has a unique contribution to make. It examines over one thousand grammar questions that English teachers in Hong Kong sent over a period of seven years to a website, *TeleNex*, to seek advice and demonstrates how empirical linguistic data which show the context and frequency of occurrence of the linguistic items in question can be a powerful tool to raise teachers' linguistic sensitivity, to help teachers question long-standing assumptions, and to gain new insights into language structure and use.

To contextualize the discussion on teachers' grammar questions, some brief background information about *TeleNex* is provided below.

TeleNex

TeleNex is a website for English language teachers in Hong Kong. It is developed and managed by TELEC (Teachers of English Language Education Centre), housed in the Faculty of Education at The University of Hong Kong.¹ On the *TeleNex* website there is a conference area in which a number of discussion corners have been set up to which teachers can send questions or comments. One of these corners is the Language Corner (formerly the Grammar Corner) that focuses on issues relating to the English language. Teachers send questions to this corner to ask for advice when they come across discrepancies between grammatical statements made in textbooks and by various sources of authority, such as the Examinations Authority (which is responsible for administering public examinations) and the Education Department (the equivalent of the Ministry of Education), when there are differences in opinion regarding English grammar amongst themselves and between native- and non native-speakers of English, and when they cannot find the answers in reference grammars and dictionaries. (For a detailed discussion of the design of *TeleNex* and studies of the discussions in the conference corners, see Tsui 1996; Tsui & Ki 2002.)

Teachers' questions are responded to by their peers, who are registered users, and language specialists in TELEC. In the rest of this chapter, I shall examine the questions sent by teachers over a period of seven years, and discuss the ways in which corpus data, including corpora which were compiled

by TELEC (called *TeleCorpora*) as well as other corpora, such as Cobuild, were used to help teachers address questions which they came across in their every day teaching. Other corpora such as the *British National Corpus* (BNC) were also examined in the discussion.

TeleCorpora consist of a corpus of modern English and a learner corpus. The latter is still being constructed and at present is around 2.2 million words, consisting of a corpus of primary students' written English and a corpus of secondary students' written and spoken English. The former started as a 5 million-word corpus of modern English (hereafter referred to as Modern English Corpus, MEC). MEC consists of 2 million words of feature articles from an English newspaper in Hong Kong, the *South China Morning Post*,² 1 million words of spoken English from radio programs such as radio phone-ins and panel discussions, casual conversations and lectures, and 2 million words of literary, academic and journalistic texts. The SCMP feature articles were written or edited by native speakers of English in Hong Kong. Though the native awareness of the writers or editors may be tempered slightly by local words and wordings, they were definitely "expert users" of English (Carter & McCarthy 2001). The *South China Morning Post* corpus has now grown to more than 20 million words. *TeleCorpora* is now available for on-line access by teachers registered as users (<http://www.telenex.hku.hk>).

Teachers' questions and corpus evidence

An analysis of the questions sent to the Language Corner over the period of seven years since the website came into full operation in 1994, show that they largely fall into one of the following six types. The first type has to do with synonymous lexical items. Some lexical items are largely synonymous but have different usage. Teachers are aware of the difference in usage but have problems explaining the difference to students, for example, "tall" and "high". Some lexical items appear to be synonymous but teachers are not sure if they are "absolute synonyms" (Lyons 1981; Partington 1998), for example, "day by day" and "day after day". The second type relates to linguistic evidence that contradicts the prescriptive grammar rules that teachers had been taught when they were learners, the most frequently asked being subject-verb agreement and the use of the definite article. The third type concerns lexical collocations which teachers try to rationalize but cannot. The fourth type consists of lexical items which teachers take to be absolute synonymous but have been asked by students to explain whether there is any difference in meaning, for example,

“big” and “large”, “lastly” and “finally”. The fifth type are prescriptive stylistic rules which seem to have been passed on from generation to generation but are queried by teachers, for example, the rule that one should not begin a sentence with “Because”, “And” and “But”. Finally, the sixth type concerns lexical items which students find confusing because the translation of these items into the students’ first language is either identical or very similar, for example “find” and “look for” which will have very similar translations.

Because of the limit of space, I shall focus on the first three types of questions. (For a discussion of the other three types, see Tsui 2001.) Teachers’ questions and responses will be cited, followed by the use of corpus evidence in addressing their questions.³ The messages will be reproduced verbatim. I shall also discuss how in the course of addressing teachers’ questions, analyses of corpus data led to insights about linguistic patterns and meanings which have not been given much attention in standard libraries or in common practice.

Synonymous lexical items

One of the most frequently asked questions is whether there is any difference between words that are commonly taken as synonymous. There are cases in which teachers were not aware of any difference in meaning and usage, such as “big” and “large”, “lastly” and “finally”. But there are some in which they were aware of a difference in usage but could not quite articulate what the difference was, for example, “tall” and “high”.

Tall versus high

The following is a message sent by a teacher, Teacher 1, who said that she knew that “tall” and “high” have different usage, but she could not quite explain the difference to her students.

The words ‘tall’ and ‘high’ have similar meaning but different usage. I have no problems in using the words myself. However, I find it difficult to explain the difference between these two words to my students. Is there any suggestions of teaching these two words?

Teacher 2 responded by saying that she had similar problems but all she could do was to give them examples and explain the examples in Chinese.

My ss (students) have the same problems as yours. I’ve come across sentences like these:

He is 1m 65 cm high.

How high are you?

All I can do is to write the sentence on the board, explain(s) them in Chinese, (and they laughed), and ss most likely will get the difference. But I don't really know what else I can do.

The difference in usage between “tall” and “high” is particularly difficult for Chinese learners to grasp because there is no such distinction in Chinese. Both words will be translated as the same word in written Chinese. Therefore explaining the words in Chinese does not really help.

Teacher 3 responds to Teacher 1's question by providing an explanation from the *Longman Dictionary of Contemporary English*, which states that “high” is used for measurement of most things but not people, especially when we are thinking only of distance above the ground, such as “a high shelf”, “a high building” and “a high mountain”, whereas “tall” is used for people and ships, for example “a tall man” and “a tall ship”. It further added that “tall” is used for things that are high and narrow. It gave the examples of “a tall/high building” and a “tall/high tree”.

In response to the teachers' questions, TELEC staff searched *MEC* in *TeleCorpora* and examined the nouns that were modified by “high” and “tall”. What emerged from the search was that there was a tendency for “high” to be used in a metaphorical sense with more abstract nouns whereas “tall” tended to be used more frequently with concrete nouns such as people, trees and buildings. The following concordance lines were provided to the teachers.

[Concordance A]

g at least as serious in areas of high **industrial growth** as in the p
 amentals in the region will offer high **growth potential**.” The ass
 huric acid – the latter at a very high **concentration**. Following Le
 l schools suffered low morale and high **wastage rates** because of inad
 just 50 cents, and the relatively high **increases** of \$2 and \$5 would
 he next 10 years, and be accorded high **priority**,” it says. {para} M
 4,000-strong HKMA claimed that a high **inflation rate**, which pushed
 enveloped countries tend to charge high **accounting rates** because they
 ssion. I must say we don't have a high **calibre** of people coming into
 rket slump had been compounded by high **inflation**, overheating and ti
 senger movement also incorporates high **standards** of monitoring with
 ing and smashing up Hongkong is a high **price to pay**’ {article} Howel
 s death, concluded that she was a high **suicide risk**. {para} Ms Lai a
 e findings by a legal expert that high **costs** have kept civil cases o
 Hong Kong to remain a bastion of high **ethical standards** in business
 se they work and because of their high **visibility**. But they should n
 f failure. My career started on a high **note**, then plummeted to icy d

transfer beam at the bottom of a tall **building**. And there isn't at
enough urinals so "small boys and tall **men** won't be embarrassed"; m
sustained by daily rainfall, of tall **trees**, bananas, relatives of
ecause their remote locations and tall **chimneys** ensured the smoke wa
e Gaulle or Wellington? For every tall **leader** there is a small one.
} is that the man with glasses, a tall **man**? {B} mmm {A} mm-hmm {B} S

This was a useful start to get the teachers to think about a word not in isolation but in terms of their "semantic preferences" (Sinclair 1991).

A further analysis of *MEC* showed that there are 1,779 instances of "high" whereas there are only 96 instances of "tall". Except for 9 instances in which "tall" is used idiomatically, such as "a tall order", "walk tall", the rest are used in the context of talking about the height of people, buildings, and things (see the above concordance lines for "tall"), with the highest frequency of "tall" co-occurring with people (about 50%) followed by buildings and structures (about 35%). In other words, the semantic preference of "tall" is quite restricted. By contrast, the contexts in which "high" is found is much more wide-ranging, including amount, intensity, quality and relative quantity. Taking 10 as the cut-off point for frequency yielded the following nouns that co-occur with "high" in the corpus (see Table 1).

A further search was conducted on the *BNC* which is much bigger and contained a larger variety of texts than the *MEC* in *TeleCorpora*. The findings converged with those yielded by the search on the *MEC*: there were 38,188 in-

Table 1. Nouns and their frequencies of co-occurrence with "high"

High	
Level(s)	77
interest	43
Cost(s)	36
Price(s)	26
degree	25
risks	24
quality	22
Standard(s)	20
inflation	18
profile	15
proportion	14
speed	14
frequencies	11
rate	11
ground	10

stances of “high” but only 4,329 instances of “tall”. One hundred instances of “high” were downloaded from the *BNC* and an analysis was conducted. Out of the 100 instances of “high”, only 31 instances are used in the context of measurement of things, for example, “a high mountain”, “high railings”. Except for 11 instances where “high” is used as part of a proper noun, such as the High Court, High Commission, High School, High Street, and High as referring to God, the rest, that is, 58 instances, are used with abstract nouns. In other words, “high” is in fact less frequently used to indicate measurement in height. A further analysis of these 58 instances shows that there are 33 instances where high is used to indicate quantity, frequency and intensity, such as high ratio, high rates, high speed, and high incidence, and 25 instances are used in attitudinal contexts, to indicate quality, for example, high class, high prestige, high standards, and so on. For the latter, most of the instances carry a positive semantic prosody (Sinclair 1996; cited in Tognini-Bonelli 2001: 111). For example,

[Concordance B]

..... He also had a **high opinion** of British travellers and officials
 We grew up in a world of chain store **high fashion**, middle-of-the-road revolution,
 n form, Josh Gifford, introduces a potentially **high class** recruit to the chasing ranks in French
 oduce FM radios, then television sets – and **high fidelity** reel-to-reel tape recorders which
 me women will use it because for them it’s a **high prestige** variety.
 ucation gave local education authorities very **high profiles** in the lives of ordinary people;
 He holds the rabbit in **high regard** as a sporting quarry.
 network to promote, stimulate and encourage **high standards** of health and safety at work.

In order to find out if the question that the first teacher asked reflects a common problem amongst students, the student corpus in *TeleCorpora* was checked. It was found that the problems that students have do not seem to be related to the distinction between “high” and “tall”, but rather related to inappropriate collocation which was very likely to be influenced by the students’ L1.

[Concordance C]

sionals and workers enriched with **high skills** are increasing. The f
 e studying but also help to pay a **high attention** and most incentive
 though Hong Kong is a place with **high freedom**. But as citizens in
 has not been completed. They **have high curiosity** to their environmen
 ver, you employ more teacher with **high qualifications** in English or

The word “high”, as used by the students, is likely to be a direct translation from the Chinese word “high” which could mean “a high degree of” or “so-phisticated”. In different contexts, the Chinese translation of “a high degree of”

have fairly different meanings. For example, “high skills” means sophisticated skills, “high attention” means to pay a great deal of attention, and “high curiosity” means extremely curious. In different contexts, the word “high” takes on different meanings.

Day by day versus day after day

There are also words or phrases which look almost synonymous but teachers feel that there may be some difference in meaning, that is, they may not be “absolute synonyms”. However, they do not quite know what the difference is. An example is “day by day” and “day after day”.

In response to a teacher’s question (Teacher 4) regarding the difference between these two phrases, the search conducted on the *MEC* showed only two instances of “day after day” and seven instances of “day by day”. The number of instances was too small for any patterns to be noticed, though *TELEC* staff had a hunch that the semantic prosody of “day by day” was positive or neutral whereas that of “day after day” was negative. A subsequent search was conducted on the 20-million-word corpus of Cobuild, and the following response was sent to the teacher:

You may remember in my previous message that I said I wasn’t really sure what was going on, and suggested that there were two possible explanations: either the *TELEC* corpus was not large enough to provide an accurate picture, or our intuitions were not reliable.

I have recently checked these two phrases in Cobuild’s 20 million word corpus, and find that the data available from the larger corpus supports our original feeling that *day after day* is normally used to convey some kind of negative feeling, possibly of frustration with the monotony, whereas *day by day* is typically used either for neutral situations, or to convey a sense of something worthwhile, developing in a steady, positive manner. Here are some citations which illustrate this tendency:

“Day after day” used to convey some kind of **negative feeling**

[Concordance D]

eel his father’s death, spending day after day in silence, self-
angry and wretched. I have stood day after day watching the wago
tation of soup. Thus we starved day after day and night after night
to windward in 60 knots of wind day after day after day.”
ur people, in the poor women who day after day discover suffering

that I couldn't face meeting him day after day, and partly that I'd
 want them to do? Sit at home day after day contemplating suicide,
 than I can stand the horror, day after day at the court and in
 cally repeats the words 'Muslim' day after day. My friends, b
 fter year .. the same old thing day after day. And the impulse came

As can be seen from the above concordance lines, the phrase "day after day" co-occurs with lexical items which denote negative experiences, events and feelings, such as "death", "suffering", "suicide", "horror", "starved", and "angry and wretched". In other words, the semantic prosody is negative.

By contrast, "day by day" occurs in contexts that are either neutral or positive. In the following concordance lines, the contexts in which "day by day" are neutral, like tours, plan, people living side by side, horoscope.

[Concordance E]

on all the tours, with maps, day-by-day itineraries and other
 transfer, the plan is to take it day by day in the Alps, then
 people who live side by side and day by day with Israelis These
 No I just take life as it comes day by day. Mm. No I'm not
 mean if I took my horoscope day by day which I n+ I don't

"Day by day" in the following concordance lines conveys a sense of development and progress, as pointed out by a TELECOM staff member in his reply to the teacher's question.

[Concordance F]

your competence is being increased day by day. Amleto felt
 Having mastered the art of living day by day, we can always man
 smoking, your health is improving day by day, so why not take it
 as a result her memory is improving day by day. In particular, she
 They discovered, bit by bit, day by day, what algebra was

The contexts in the above lines are positive: they are about increase in competence, mastering the art of living, improvement in health, improvement in memory and discovery of knowledge.

Sometimes, the negative semantic prosody of "day after day" is being exploited in a positive context to emphasize an element of repetitiveness. The following line appears on the packaging of Vichy hair conditioner.

Smooth and manageable hair, day after day.

In the above advertisement, the advertiser exploits the repetitiveness denoted by “day after day” to emphasize the consistent effect that the conditioner will have on the hair.

Grammar rules and conflicting evidence

Teachers are often troubled by the fact the grammar rules that they have been taught as students do not accord with the authentic linguistic examples that they encounter (see also Tognini-Bonelli 2001). Indeed, in the Language Corner, one type of most frequently asked questions has to do with teachers who try to apply some usage rules but are confronted with conflicting evidence. The most frequently asked questions pertain to subject-verb agreement and the use of definite articles.

Subject-verb agreement

The following are some sample messages sent by teachers:

Teacher 4

Hello! Which one is correct?

There is a man and a woman outside.

Or

There are a man and a woman outside.

Please give some comments, any one.

Teacher 5

Hi,

What should we use in the following sentences? Is or are?

1. There _____ an apple and some oranges on the table.
2. There _____ some oranges and an apple on the table.

Thanks

It seems to me that ‘are’ is okay in both. Is there any rule here?

Teacher 6

To me, I will use “There are a man and a woman ...”. It is because we are talking about two persons.

I want to make sure that if this is grammatically correct.

We can see from the response given by Teacher 6 that she was trying to apply the rule of subject-verb agreement to the example provided by Teacher 4. Teacher

5 intuitively felt that “are” can be used in both sentences but she was looking for some rules.

TELEC staff responded to the teachers’ questions by pointing out that usually the singular form of ‘be’ is used when the first noun that follows is singular and the plural form of ‘be’ is used when the noun group after it is plural (see also *Collins Cobuild English Grammar*, p. 416). However, a search through the corpus does show an instance of the following:

According to PACE, suspects can only be detained at designated police stations where **there are a custody and a reviewing officer**.

In other words, while what is stated in the Cobuild English Grammar is correct, teachers may benefit from knowing that occasionally the plural form of ‘be’ is actually used even when the noun following is singular. Therefore, the question is not about possibility but about probability of usage.

What is interesting is that the corpus data show that “there’s” is often used in informal and spoken English before a plural noun. Although Quirk et al. (1985) have also made this point, it is much more convincing to provide teachers with corpus evidence. A search of the *MEC* showed that there are 1,693 instances of “there’s”. The following concordance lines were provided to teachers.

[Concordance G]

ollaboration is concerned I think **there’s 2 reasons**. The first is that this I
when we first married that, er, if **there’s any small decisions**, well, the wif
...talked about that a bit think if **there’s any other reasons**, and most importa
know. – Seriously? – No, no, no. **There’s buses and things**, they ve body, not just in the
.....er – people isn’t there **there’s more mothers**, yeah yeah in and
.....looking around. I mean, **there’s more books** there than t the dickens is going on,
.....from home as well C: And **there’s some jobs** you can earn some money?

By contrast, there are only three instances of “there’re” before a plural noun.

[Concordance H]

uickly have a look at these i mean **there’re a lot of them**. we obviously haven
if you start at the ends . because **there’re no unknowns** at the ends you see s7
ike that . moving along . and that **there’re two sorts** of forces that operate

The above discussion led to further questions of a similar nature from other teachers regarding whether the singular or the plural verb should be used in the context of “one of the + plural noun”.

Teacher 7

Should you say:

1. Peter is one of the richest boys that have/has ever studied in our school.??
2. He is one of the writers who were/was honoured yesterday.??
C/F
3. One of the boys was punished yesterday. CORRECT
4. One of the writers was honoured yesterday. CORRECT

Teacher 8 responded as follows:

I've asked my panel and she said,

1. Peter is one of the riches boys who **has** ever studied in our school.
2. He is one of the writers who **was** honoured yesterday.

A search of *MEC* in *TeleCorpora* showed that there are 348 instances of “one of the ... (be) ...” and both the plural and singular “be” forms are used. There is a higher frequency of occurrence of the plural form but the singular form is used often enough to be regarded as an acceptable alternative.

Some further examples of questions of similar nature are whether a plural or singular verb should be used after the structure “none of the ...” and “more than one ...”. For example,

Teacher 8

Should we use a singular or plural verb after the structure ‘none of the’?

Teacher 9

Hi, would somebody be kind enough to tell me why we should use a singular verb after ‘more than one player’?

It is clear from the teachers’ questions that they were puzzled by the lack of agreement between the subject and the verb. As a *TELEC* staff member pointed out to the teachers, technically, “none” means literally “not one”, and it seemed to be more logical to use a singular verb. However, because “none of” functions as a quantifier, it is often followed by a plural noun, and therefore a plural verb is used. A search of the *MEC* found both singular and plural verbs being used. For example,

[Concordance I]

None of the studies **has** survived. Lautrec and Bernard may have both
 None of the five **was** masked when they hauled the driver out of his
 None of the oeuvre catalogues **includes** it. Yet it was a pa
 None of the conference recommendations **is** legally binding the estab
 none of the members who head Omelco panels **were** allowed to act as sp
 none of the 36 Vietnamese going home **have** volunteered, unlike some o
 none of the blanks **are** filled in and they go ahead and dredge anyway
 None of the members **are** elected to the board, while four ex-officio

The above discussions, sparked off a series of questions from teachers posing related questions asking whether one should say “There **are** no students in the room.” or “There **is** no student in the room.”; “I have no **friends**.” or “I have no **friend**.” The use of corpus evidence is particularly helpful because the question is not about possibility but probability. Moreover, teachers came up with so many variations of the subject-verb agreement structure that it would not have been possible to provide some kind of comprehensive guiding principle. The best solution is to get them to look for corpus evidence themselves.

Definite articles

The presence or omission of the definite article is another problematic area for teachers. They have difficulties finding some kind of consistency in the rules for using the definite article. For example, they have been told that the definite article should be used if there is only one of a kind being referred to, such as the sun, the moon, and the earth, the name of a country, and before a position, such as the Chairman and the Secretary. However they have also come across cases where the definite article was missing. For example, “He was elected Chairman of the Association”, “She was appointed secretary of the committee.”

Sometimes, the questions asked by the teachers can be very specific and it would not be possible to answer them without consulting a corpus. Take for example the following message sent by Teacher 7:

Hello! Although I have been an English teacher for about 4 years, I still sometimes have difficulty in using articles. I would be very grateful if someone can help me in the following problem:

They watched television.

They listened to music.

They listened to the radio.

So, should I say:

“When the teacher was teaching, they listened to the walkman.”
Or “When the teacher was teaching, they listened to walkman.”

One TELEC staff member responded by pointing out that the definite article “the” is used when referring to systems of communication or mass media, such as “the radio”, “the telephone” and “the mail”. He observed that the use of “the” is a bit variable in television because a search through the corpus showed instances of television, with and without “the”. For example,

[Concordance J]

I watched **the television** last night and was gripped by the horror and ...
Jazz was watching **the television**...
...an enjoyable play on **the television** ...
to get some exciting bands on **the television**...

week and is also permitted to **watch television** at certain hours. His
er with a blanket and went to **watch television**. But when she checked on

Yes. 1. Well, did you (um) **watch television** the other night when

The question about “walkman” is a bit more problematic. There is not a single instance of “walkman” in the *MEC*. The TELEC staff introspected that because “walkman” is in fact a brand name which has been generalized to refer to any small portable cassette player rather than a system of communication, the tendency would be to use it as an ordinary countable noun, and therefore the indefinite article “a” or a possessive pronoun would be used. He conducted a further search on the Cobuild corpus and found 28 instances of “walkman”. The corpus data confirmed his introspection.

Rationalization of collocations

The third type of frequently asked question has to do with rationalization of collocations. Teachers often try to look for rules governing which words can go with certain words and why. For example, the following is a message from Teacher 8 asking whether one can say “well-experienced”.

If we say someone is experienced, we mean this person has certain knowledge or expertise. Do we have ‘well-experienced’ as well? If so, does it mean there is an even higher level of expertise?

One of the TELEC staff who is a native speaker of English replied as follows:

I couldn’t make up my mind about well-experienced. We say well-educated, well-brought-up etc. So why shouldn’t we say well-experienced?

However, one minute it sounded OK but the next it didn't so I looked it up in the corpus.

There weren't any examples of well-experienced at all. To express an even higher level of expertise, the examples from the corpus showed that people use adverbs such as *very* and *vastly* – but these don't seem to be very common. ...

Mind you, ..., even if there is such a word as well-experienced, I'm now not so sure that it means more than experienced. Similarly, is a well-educated person more educated than an educated one?

A search conducted on the *MEC* in *TeleCorpora* on the word “experienced” showed that there are 105 instances of “experienced” used as an adjective. There are only 25 instances where “experienced” was modified by intensifiers like “very”, “supremely”, “vastly” and by comparatives and superlatives like “more” and “most”.

A further search was conducted on the *BNC* on “experienced” and yielded 8 instances of “well experienced”.

[Concordance K]

the intention being (and Dudek was very **well experienced** in this sort of work) for him to take
 As a rule, the examining doctor will be **well experienced** in dealing with sexual symptoms
 Firms which are **well experienced** in overseas employee transfers often
 to be numerous and very well armed – though how **well experienced**?
 A life science graduate already **well experienced** as a CRA, you should ideally have
 encouragingly, our Export Department is **well experienced** and appears to be well placed to
 ever worked in pubs; CVs of head chefs **well experienced** in take-aways; you read the CV of
 I mean as you as you say, Nick's a long serving and **well experienced** reliable

To investigate whether there is any difference in the behavior between “experienced” and other adjectives that take “well” as the modifier, a search was conducted on “well” in the *BNC* and it yielded the following compound adjectives: “well qualified”, “well educated”, “well organized”, “well equipped”, and “well-known”. To see whether the rare occurrence of “experienced” being pre-modified by the adverb “well” has to do with the semantics of “experienced”, a search was conducted on the modifying adverbs, “highly”, “very”, “poorly” and “badly”. The following are the results of the search (see Table 2).

The figures in Table 2 show that there are several ways in which “experienced” behaves differently from the other five adjectives. First, despite the fact that “well experienced” is found in the *BNC*, its occurrence is far fewer

Table 2.

	well	highly	very	poorly	badly	very well
experienced	8	33	78	0	0	1
qualified	89	86	1	1	0	8
educated	76	39	3	8	2	7
organized	57	39	3	7	6	10
equipped	151	1	0	20	3	11
known	1750	0	0	21	0	52

than the rest. Second, while there is a large number of instances of “experienced” taking the intensifier “very”, there are very few or no instances of the other five adjectives co-occurring with “very”. Third, these five adjectives, however, take the intensifier “very” when they combine with “well” to form compound adjectives. Fourth, while “educated”, “organized”, “equipped” and “known” can be modified by “poorly” and / or “badly”, “experienced” does not. These four characteristics suggest that it is likely that “experienced” denotes a positive quality which renders the modification by “well” superfluous and the contradictory modification by “poorly” and “badly” unacceptable. By contrast, except for “qualified”, the other adjectives can be modified by adverbs denoting negative qualities, suggesting that they can be used neutrally, though they commonly denote positive qualities.

Implications for language teacher education

In the above discussion, we have seen that teachers often look for generalizations about grammar rules so that they can provide some guidelines to their students. This is perfectly legitimate especially in second language learning situations where learners do not have the same amount of exposure to the language as in first language learning situations. The problem is whether the rules and generalizations indeed capture how language is actually used rather than how language is perceived to be used, and whether they reflect the dominant patterns of use. The easy accessibility of corpora allows teachers to check prescribed rules and generalizations against linguistic data, encourages them to be sensitive to patterns that emerge from the data and to make their own interpretations and generalizations of these patterns (see also Hunston 1995).

Indeed, the constant use of corpus evidence in addressing teachers’ questions by TELEC staff helped teachers to reflect on their knowledge of the language as well as critically examine grammar rules and patterns that they

had always taken for granted. They began to look at corpus evidence for answers, instead of just relying on dictionaries and reference grammars. More and more messages like the following have emerged in the Language Corner in recent years.

When I was at school, one of my English teachers taught me patterns like:

I help him **with** his English / (to) do something.

I assist him **in** doing something.

I wonder if it's alright to say, for example, "I help him **in** doing sth." What does the Bank of English say about the collocations that go with "help".

The process of using corpus evidence to address teachers' questions is beneficial not only to the teachers but also to TELEC staff themselves because in this process they often notice linguistic patterns and pragmatic loads carried by linguistic items that they were not aware of previously, some of which were not attended to in the standard libraries. The following is just one example, among many, of how a question from a teacher can lead to interesting discoveries of linguistic facts.

Imply and infer

These two words cause some confusion because some people use the word "infer" to mean "imply", as observed by the *Collins Cobuild English Dictionary* (p. 862, entry "infer"). For example,

The police **inferred**, though they didn't exactly say it, that they found her behaviour rather suspicious.

A teacher sent a question regarding these two words:

Would you please provide some examples for me to explain the use of the two words (imply & infer)? Thank you very much!

Another responded by saying that "to imply" means "to suggest something indirectly", whereas "to infer" means "to guess something is the case or to conclude".

The explanation provided by the above teacher was very much in agreement with those provided by the *Collins Cobuild English Dictionary* for "infer" and "imply", which is given below.

If you **infer** something is the case, you decide that it is true on the basis of information that you already have. *I infer from what she said that you have not been well.*

1. If you **imply** that something is the case, you say something which indicates that it is the case in an indirect way. *'Are you implying that I have something to do with those attacks?' She asked coldly.*
2. If an event or situation **implies** that something is the case, it makes you think it likely that it is the case. *Exports in June rose 1.5%, implying that the economy was stronger than many investors had realized.*

Instead of just restricting the discussion to the meanings of these two words, TELEC staff searched the *MEC* and noted that there seems to be a number of instances where “imply” is modalized, and where the writer makes it clear that what has been said does not imply what the reader has inferred. In fact, out of a total of 59 instances of “imply” found in *MEC*, there are 17 instances of “imply” being modified by either modal verbs, “may”, “might”, “would”, “could” or lexicalized modality, such as “seem to”, “appear to”, “tend to”. For example,

[Concordance L]

from the Reiss and Wagner study **may** imply only that, in their procedure, generation to this on the ground that it **may** imply that women generally have weaker characteristics. That was because they **tended** to imply that beliefs, laws, and principles were. None the less many writers **seem** to imply that it is, and as a result their position found in the work of many **seems** to imply. Rather, modern anthropologists tend to imply in the English 19th Century novel and **will** imply as he does so that these so-called pressure to purchase flats on the market **would** imply additional demand and upward pressure on money from the Land Fund **would** imply that Chinese officials would be appointed to an office in Taiwan since it **would** imply recognition. {para} “Now, one of the vic buildings of Venice. This **would** imply that the development is itself something

There are 18 instances of negative forms of “imply”, such as “does not imply”, “should not be taken to imply”, and “would be unfair to imply”, and so on. For example,

[Concordance M]

about fuel being consumed we **don't** imply that energy is lost. There's still the circumstances **need not necessarily** imply a change in the specific ability of the environment inhibition **does not necessarily** imply a rejection of all other theories of evolution for the latter **does not necessarily** imply nationalism. Despite the official position not an optimistic sign. It **does not** imply the future is a bright one.” In the next definition, Category II **did not** imply buildings were dangerous. The use of α and β ; **should not be taken to** imply a belief that α and β ; is the sole deontic “truthistic” biologists **do not wish to** imply that animals are actuated by a concept of Irish catholicism is **not meant to** imply a negation of its positive aspects: 1935. It **would be misleading to** imply that mucus and lysozyme were Florey's (p. 29). It **would be unfair to** imply that we were working in a kind of vacuum

A search of the *MEC* for instances of “infer” showed that there is a similar tendency for “infer” to be modalized. There are 14 instances of “infer”, out of which nine co-occur with the modals “can”, “could”, and lexicalized modality, such as “plausibly” and “willing”, and two are in negative forms.

[Concordance N]

your body and I ‘have” mine we **can** infer the general features of each other’s
 it to make things difficult we **can** infer what gets done in the normal way of
 evidence from which the jury **could** infer _ and did infer _ that he intended t
 hich the jury **could** infer _ and did infer _ that he intended to cause grievous
 recording changes. We **can** **plausibly** infer that crime has been increasing in th
 developmental biology **not to try to** infer developmental mechanisms from final
 out development **to resist trying to** infer how a structure develops by only loo
 doctor was none the less **willing to** infer from the circumstances that there ha
 worker, Clark Hull, **endeavoured to** infer the processes within the black box t

The common characteristics shared by “imply” and “infer” is that they both pertain to what is not explicitly stated. Therefore, people tend to hedge statements about implications and inferences with modals or lexicalized modality. The much higher proportion of the negative form of “imply” as compared to that of “infer” suggests that there is a difference in the semantics of the two lexical items. The negative form of “imply” serves to pre-empt possible misinterpretations of what is not directly said. This kind of evidence helps teachers to understand better the difference in meaning between these two items spelled out in the *Collins Cobuild English Dictionary*. These characteristics would not have been easily detected without the help of the corpus and the concordancer showing the environments in which these two words occur.

Studies of applications of corpus linguistics to second language teaching and learning have emphasized the importance of adopting a data-driven approach to language learning so that learners go through a process of self-discovery (see for example Johns 1991). The discussion in this paper shows that it is equally important, if not more important, for teachers to go through this process of self-discovery and to experience formulating generalizations about linguistic patterns that they have observed so that they own the grammar as much as linguistic researchers.

Notes

1. TELEC has received funding from a number of sources to develop this website since its inception in 1993. The author wishes to thank the Hong Kong Telecom Foundation (now PCCW), the Hong Kong Jockey Club, the Hong Kong Government Language Fund and Quality Education Fund for their generous support.
2. Feature articles are downloaded to the Centre two or three times each week by the *South China Morning Post*. The author wishes to thank the *Post* for their generosity and assistance in providing the texts.
3. The author wishes to thank the teachers whose messages she has cited for allowing her to do so in this chapter. She would also like to thank her colleagues at TELEC for allowing her to cite their responses to teachers, particularly Quentin Allan whose responses constitute a major part of the data cited in this chapter.

References

- Allan, Q. (1999). Enhancing the language awareness of Hong Kong teachers through corpus data: The *TeleNex* experience. *Journal of Technology and Teacher Education*, 7(1), 57–74.
- Berry, R. (1994). Using concordance printouts for language awareness training. In C. S. Li, D. Mahoney, & J. Richards (Eds.), *Exploring Second Language Teacher Development* (pp. 195–208). Hong Kong: City University Press.
- Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D., S. Conrad, & R. Reppen (1994). Corpus-based approaches to issues in applied linguistics. *Applied Linguistics*, 15(2), 169–189.
- Bygate, M., A. Tonkyn, & E. Williams (Eds.). (1994). *Grammar and the Language Teacher*. New York: Prentice Hall.
- Carter, R. & M. McCarthy (1997). *Exploring Spoken English*. Cambridge: Cambridge University Press.
- Carter, R. & M. McCarthy (2001). Size isn't everything: Spoken English corpus, and the classroom. *TESOL Quarterly*, 35(2), 337–340.
- Collins Cobuild English Grammar* (1990). London: HarperCollins.
- George, H. V. (1963). *Report on a Verb-Form Frequency Count*. Monograph 1. Hyderabad: Central Institute of English.
- Hawkins, E. W. (1999). Foreign language study and language awareness. *Language Awareness*, 8(3/4), 124–142.
- Holmes, J. (1988). Doubt and certainty in ESL textbooks. *Applied Linguistics*, 9, 21–44.
- Hunston, S. (1995). Grammar in teacher education: The role of a corpus. *Language Awareness*, 4(1), 15–31.
- James, C. & P. Garrett (Eds.). (1991). *Language Awareness in the Classroom*. London: Longman.

- Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (Eds.), *Classroom Concordancing, ELR Journal 4* (pp. 1–16). Birmingham: CELS University of Birmingham.
- Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Longman.
- Kjellmer, G. (1994). *A Dictionary of English Collocations Based on the Brown Corpus*, 3 Vols. Oxford: Clarendon Press.
- Ljung, M. (1991). Swedish TEFL meets reality. In S. Johansson & A.-B. Stenström (Eds.), *English Computer Corpora* (pp. 245–256). Berlin: Mouton de Gruyter.
- Lyons, J. (1981). *Language and Linguistics*. Cambridge: Cambridge University Press.
- Mindt, D. (2000). *An Empirical Grammar of the English Verb System*. Berlin: Cornelsen.
- Partington, A. (1998). *Patterns and Meanings. Using Corpora for English Language Research and Teaching* [Studies in Corpus Linguistics 2]. Amsterdam: John Benjamins.
- Quirk, R., S. Greenbaum, G. Leech & J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins.
- Tsui, A. B. M. (1996). The participant structures of *TeleNex* – a computer network for ESL teachers. *International Journal of Educational Telecommunications*, 2(2/3), 171–197.
- Tsui, A. B. M. (2001). What teachers have always wanted to know – and how corpora can help. Paper presented at the conference on How to Use Corpora in Language Teaching, The Tuscan Word Centre, Italy.
- Tsui, A. B. M. & W. W. Ki (2002). Socio-psychological dimensions of teacher participation in computer conferencing. *Journal of Information Technology for Teacher Education*, 11(1), 23–44.

Resources – Corpora

Corpus variety

Corpus linguistics, language variation, and language teaching

Susan Conrad

Portland State University

Many teachers view variation as an annoying aspect of language use that needs to be ignored, believing that the conditions associated with the choices among different grammatical variants are too complex, subtle, or unimportant to consider when designing teaching materials. With examples from corpus-based work, this paper argues that variation is a crucial aspect of naturally-occurring language, that grammatical variants often have very clear associations with certain factors, and that ignoring this variation has undermined the effectiveness of teaching materials. First, a corpus-based analysis of the linking adverbial *though* is compared to its treatment in four ESL textbooks – illustrating that ignoring variation between conversation and academic prose results in misleading descriptions in the textbooks. Second, the patterns of numerous co-occurring linguistic features are compared between a practice lecture from an ESL textbook and a corpus-based analysis of lectures recorded at American universities. The analysis shows that the textbook lecture does not give students practice with some of the most important groups of features that make the naturally-occurring lectures different from other forms of discourse. In addition, the example illustrates how a software program developed for a corpus-based research project can also serve as a useful pedagogical tool.

In his review of the *Longman Grammar of Spoken and Written English* (Biber et al. 1999), a corpus-based reference grammar, a reviewer bemoans his experience looking up information about when to use a past perfect construction. He warns readers:

After a great deal of searching, you may find the answer is no more definitive than ‘it depends.’ Of course, this is quite probably the right answer but it doesn’t help students finish their homework or pass exams. (Ward 2000: 167)

Reading the review, I was shocked. I agree that the book is not suitable for an ESL student who needs basic structural information. But why would a language teacher seem to be surprised at the answer “it depends”? He even seems disappointed that a large reference grammar would cover so many conditions about use of past perfect – its frequency across different registers, verbs that are common and uncommon with it, its co-occurrence with adverbials and dependent clauses, and conditions that affect the choice of simple past versus past perfect.

On reflection and after discussion with several language teachers, I think that these comments are a reflection not just of frustration at wading through a large reference book, but of a widely held attitude in language teaching: variation is just an annoying aspect of language use that needs to be ignored. Conditions affecting the choice between one structure or another are considered too complex and subtle, and not particularly important to know. Teachers and students seek only definitive answers – such as being able to identify what is grammatical and ungrammatical. Therefore, it is only to be expected that homework and exams do not reflect any information about making appropriate choices based on contextual conditions.

Interestingly, many teachers are exposed to ideas of language variation during their training programs. Whenever language is viewed as a means of communication, variation is a central concept. In systemic functional linguistics, for example, tying linguistic choices to their context of use is a central concern. As Thompson explains in *Introducing Functional Grammar*:

Functional Grammar sets out to investigate what the range of relevant choices are, both in the kinds of meanings that we might want to express (or functions that we might want to perform) and in the kinds of wordings that we can use to express these meanings; and to match these two sets of choices.

(Thompson 1996:8)

Similarly, many teachers-in-training take sociolinguistics classes and become familiar with ideas of variation. They study Hymes’ well-known SPEAKING acronym, used as a mnemonic for situational factors that are important to consider when examining a communicative event (see, e.g., Wardhaugh 1998:242–244). Changes of the circumstances for each factor will affect language choices. Thus, the concept of language variation is central to understanding the ethnographic framework. Yet when it comes down to a very concrete feature of language – a question about grammar, for example – teachers don’t expect to have to think about variation.

In this paper, I argue that we need to change this view of language variation. Currently, by minimizing the importance of variation, we are misrepresenting

language in materials that we use with students. In teacher-training programs we are creating false expectations for teachers. With exams that do not reflect an understanding of how language use varies, we are limiting how well we can assess a student's language proficiency for a given domain.

In discussing the importance of variation, I will use evidence from corpus-based analyses. More than any other approach in linguistics, corpus research has allowed us to understand patterns of variation more comprehensively. One approach that can be taken is to investigate the features of a particular variety that set it apart from other varieties. For example, work on the CANCODE project, a corpus of spoken British English, has highlighted a number of features of spoken English that are typically not represented in grammars based on written language, including pre- and post-posed items (topics and tails) and certain kinds of ellipsis (Carter & McCarthy 1995; Hughes & McCarthy 1998; McCarthy 1998). Academic or discipline-specific language has also been studied. For example, Coxhead (2001) presents an academic word list based on a corpus analysis – specifying the words that are common in academic texts across a variety of disciplines. Specialized vocabulary or lexico-grammatical patterns have been identified in disciplines such as plant biology (Williams 1998) and cancer research articles (Gledhill 2000). Patterns among a variety of language features have been identified in a number of science and other disciplines (e.g., Atkinson 1996; Conrad 2001; Biber & Finegan 2001).

All of this work highlights language variation by showing features that are characteristic of a particular variety. However, it is still possible for many teachers to consider such variation the concern of language teaching for “special purposes” – whether a very technical purpose such as writing scientific research articles or a less technical one, such as informal conversation. Such studies rarely discuss the need for variation to be considered in all teaching contexts, including grammar courses (though see Biber, Conrad, & Reppen 1998; Conrad 2000 for discussions of the myth of “general” English).

Another perspective taken by corpus studies is to describe variation in the use of a specific feature of language, rather than to characterize a variety. Variation in a word's lexical associations are often examined. For example, Biber (1993) shows how different senses of words correspond to different groups of collocations; thus, variation in meaning corresponds to variation in lexical associations. Much work associated with the Cobuild project has highlighted the importance of lexico-grammatical associations in language (e.g. Sinclair 1991; Hunston & Francis 1998, 2000). For example, verbs often have a number of possible complementation patterns, but some are much more frequent than others. The verb *provide* is typical with the pattern “V n with n” (“provide

someone with something”) but also occurs with “V n to n” pattern – “provide something to someone” (Hunston & Francis 2000: 97). The factors associated with the use of these different patterns contribute to our understanding of language variation. However, for some teachers it is difficult to understand how to apply such information in the classroom.

In this paper, I focus on two examples of corpus-based work to show just how important variation is for language teaching. I demonstrate that a lack of attention to variation can undermine teaching materials, and also describe how a tool designed for research on variation can also be a useful pedagogical tool. For the examples, I focus largely on grammatical structures, on English, and on variation across registers – that is, varieties defined by their situation of use, such as conversation and academic prose. The principles, however, are applicable to all features of language, to all languages, and to all kinds of variation.

The linking adverbial *though*

Linking adverbials – also referred to as “linking expressions,” “conjunctive adverbs,” “conjuncts” or “connecting words” – are a kind of adverbial that explicitly state the relationship between two units of discourse. The linking adverbials *therefore* and *thus*, for example, signal a result relationship. Because linking adverbials make discourse relationships explicit, they are recognized as important devices for creating cohesion in a text, and are therefore covered in many ESL textbooks.

Here I focus on linking adverbials that are used for contrast and concession, such as *however* and *though* in the following examples:

- (1) Some small farms are as efficient as some larger ones and at all sizes there are instances of high and low efficiency. *However*, there does appear to be a threshold somewhere between the two- and three-man unit... (Academic Prose)¹
- (2) This is where I take the uh, it's the exam. I forget the name of it *though*. (Conv)

Note that these linking adverbials differ from subordinators such as *although* and *whereas* because the linking adverbials are separate elements of independent clauses. Subordinators always occur at the beginning of a clause and mark a subordinate clause, as in this example (subordinate clause in []):

- (3) I could overhear all this, [*although* Ken didn't think I could hear what's going on]. (Conv)

The item *though* can be either a subordinator (as a variant of *although*, sometimes occurring as *even though*) or a linking adverbial. As a subordinator, it occurs at the beginning of the subordinate clause:

- (4) [*Though* he is not religiously observant], he says that he considers himself a Jew by identity. (Newspaper)

As a linking adverbial, it is often in final position, as in (2) above. However, it can occur in other positions in a clause, as in this example:

- (5) This time, *though*, she nodded. (Fiction)

In speech, the linking adverbial *though* tends to sound parenthetical in its intonation. This is often reflected with the use of commas in written prose, as in (5).

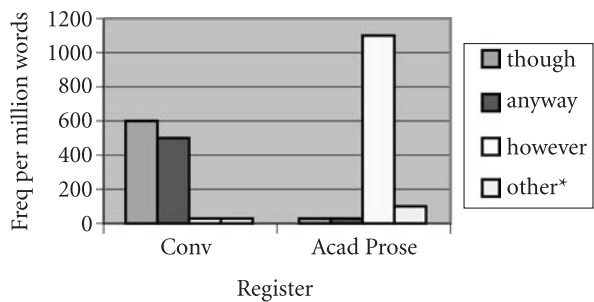
It turns out that the adverbial *though* provides a useful illustration of the way in which corpus-based research can make us more aware of strong patterns of variation and provide data about language use to compare with treatments in textbooks. First, consider the frequency of linking adverbials of contrast/concession in two registers: conversation and academic prose (Figure 1). The results are condensed from a larger study within the *Longman Grammar of Spoken and Written English* (Biber et al. 1999), and are based on an analysis of about 5 million words of academic prose and 6.5 million words of conversation, covering both British and American English. (Full details about the corpus can be found in Biber et al. 1999, Chapter 1. For this analysis, British and American English conversation were analyzed separately, but the results for the two were the same so they are not displayed independently here.)

Clearly, there is a considerable difference between the two registers for the most common linking adverbials. In conversation, *though* is most common. (*Anyway* is also common but its function is less clearcut since many occurrences could also be classified as discourse markers.) In academic prose, *however* is by far most common. There is also more diversity in academic prose. Four other adverbials occur at least 100 times per million words: *nevertheless*, *rather*, *yet*, *on the other hand*.

In quantitative terms, *though* is clearly an important adverbial of contrast/concession in conversation. An examination of its contexts of use provides further insight into its usefulness. Consider these examples:

- (6) Jeez. Oh, maybe I won't go. I should *though*, I feel that I should.
- (7) A: I love your outfit.
B: Well thank you. I feel kinda sleazy *though*.
- (8) [Remarking on a penalty call during a football game]
A: Oh, that's outrageous.
B: Well, he did put his foot out *though*.
- (9) [Trying on a new piece of clothing]
A: Actually it looks alright.
B: Yeah it doesn't look bad.
A: I would shorten the waist a little *though*.

Many occurrences of *though* operate within an utterance of a single speaker, as in (2) and (6) above. However, an important function of *though* is apparent when it is used during exchanges between speakers. *Though* provides a means of disagreeing in a less direct way than with *but* or *however*. The linking adverbial seems concessive in meaning, as though the second speaker is not contradicting the first, but just adding information that needs also to be taken into account. The disagreement is softened even when there is a clear contradiction in the speakers' ideas – for example, in (8), where Speaker B believes the referee's penalty call was correct and Speaker A does not. In sum, then, *though* can play an important interactional function in face-to-face conversation, softening statements of disagreement.



*“Other” represents the count for any one of four other adverbials: *nevertheless*, *rather*, *yet*, *on the other hand*. Some occurrences of *anyway* could alternatively be classified as a discourse marker.

Figure 1. Most common linking adverbials of contrast/concession (Condensed from the *Longman Grammar of Spoken and Written English*, p. 887)

Given this useful function of *though*, it seems likely that textbooks seeking to improve student's spoken proficiency would include *though* when covering linking adverbials. To investigate this, I examined four ESL textbooks. The books have been published within the last five years, were advertised by publishers as representing the latest developments in grammar teaching, and all claim to help students with both spoken and written language (Azar 1999; Frodesen & Eyring 1997; Preiss 1998; Steer & Carlisi 1998).

An analysis of the four books' coverage of linking adverbials of contrast and concession reveals a mismatch between the corpus evidence and what is covered in the textbooks (Table 1). Only one of the four books covers the use of *though* as a linking adverbial at all, and that book lists it only as showing contrast, not concession. None of the books have an example of *though* used to soften disagreement between speakers. Furthermore, the one book that suggests specific contrast linking adverbials to use in speech suggests *however* and *on the other hand* – two linking adverbials that are far more common in academic prose than conversation.

These books of course have many other positive features; their coverage of linking adverbials of contrast and concession is only a small part of the book. However, this coverage provides a very concrete example of how ignoring variation can detract from effective teaching materials. Linking adverbials used in writing are covered, but the most frequent linking adverbial of contrast/concession in conversation is not covered in most of the books. Its important concessive use in conversation is not covered even when it is included. The information for students is thus misleading – some implicitly, by not covering a common, useful item, and some explicitly, by suggesting that *however* and *on the other hand* are common in conversation.

Table 1. Coverage of *though* as a linking adverbial in four ESL textbooks

Book	Coverage of <i>though</i> as a linking adverbial	Information specific to linking adverbials in speech
Azar 1999	– none (gives examples of <i>though</i> as a subordinator)	– none
Frodesen & Eyring 1997	– example with a single speaker – listed as contrast connector, not concessive	– none
Preiss 1998	– none	– Spoken connectors listed: <i>however</i> & <i>on the other hand</i>
Steer & Carlisi 1998	– none	– none

Comparing many language features simultaneously

One of the advantages of corpus-based analysis techniques is that they make it possible to study a large number of language features simultaneously. This perspective is valuable as we seek to understand variation in texts because texts are, of course, composed of many different features. If we choose to focus on only one feature – such as linking adverbials in the last example – we see one view of how registers differ. Considering other features may give a very different picture of similarities and differences between the registers.

In this section I will illustrate how a software tool created to analyze the patterns of co-occurring features in texts – designed for research and test-development purposes for Educational Testing Service – can also serve as a useful pedagogical aid. The software tool will be used to examine the similarities and differences between a practice lecture from an ESL textbook and actual class sessions recorded at five universities. The lectures come from the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) Corpus, developed for Educational Testing Service (see Biber et al. 2002). As Table 2 shows, there are 176 class sessions in the corpus, and they come from a variety of disciplinary areas. They were collected in five different regions in the United States, and total over a million words. They are called “class sessions” rather than “lectures” because virtually every session had some interaction in it, even if it was in a large lecture-style classroom. No labs or study groups are included as class sessions.

The ESL textbook lecture will be compared to the class sessions using multi-dimensional analysis. This technique was developed by Biber (1988) to investigate the patterns of linguistic variation across different registers. The “dimensions” are groups of linguistic features that co-occur with markedly high frequency in texts. The dimensions are identified quantitatively, using a fac-

Table 2. Class sessions in the T2K-SWAL corpus

Disciplinary field	# of classes	Appx number of words
Business	36	240,000
Education	16	137,000
Engineering	30	171,000
Humanities	31	249,000
Natural Science	25	160,000
Social Science	38	294,000
Total	176	1,251,000

tor analysis. They are then interpreted functionally, based on the situational and cognitive functions that the features serve and on the relationships among registers along the dimensions.

Table 3 provides a summary of the dimensions of variation, while all of the linguistic features for each dimension are listed in the appendix. As an example, consider Dimension 1. It is named “Involved vs Information Production” to reflect the type of variation it covers: both a difference in purpose (interactional vs. informational) and a difference in production circumstances (real-time production vs. time for planning and revising). A text with a very “positive” score on this dimension would have relatively common use of the “positive” features: first and second person pronouns, *be* as a main verb, contractions, verbs that convey private practices like *think* and *believe*, markers of vagueness and general hedges such as *sort of* or *something like*, and present tense. It would be produced under real-time constraints, without time for planning or editing. A text with a more negative score would have more nouns, attributive adjectives, and prepositions as well as more diversity in word choice and longer words. This type of informationally dense prose is generally possible only with time for planning and revising.

To understand differences in texts along Dimension 1, consider these two passages:

(10) Class Session

Teacher: I guess uh... I'm trying to think of other levels here but the question that you have to ask is what kind of resources do you have internally? And what do you have to get externally? And what are you good at and what are you not good at? To be able to really do good innovation to get products out and I contend also to have good e-commerce sites and good e-business sites is that you have some combination of some of these things. The more you have a whole set of resources it's more likely that you are going to have a competitive advantage, and then the question is which one of these do you have...

(11) Textbook

The formation of a separate socialist bloc would insulate the East from the coming economic chaos in the West and enhance socialist economic development. The primary motivation, however, was political. A separate Eastern economic bloc, in the Soviet Union's view, would provide a buffer zone of friendly, that is, Communist states on its borders and would prevent Germany or other “hostile” Western powers from posing a threat of military invasion.

Table 3. Summary of dimensions in multi-dimensional analysis

Dimension	Description and functions	Most Important Language Features*
1. Involved vs. informational production	Interactive discourse with interpersonal concerns and high involvement vs. densely informational discourse written with time for planning and editing	More positive scores: first and second person pronouns, <i>be</i> as a main verb, contractions, “private” verbs like <i>think</i> and <i>believe</i> , vague/indefinite words, present tense More negative scores: nouns, attributive adjectives, prepositions (prepositional phrases), long words, diverse vocabulary
2. Narrative vs. non-narrative concerns	Stereotypical past-time narration	More positive scores: past tense, perfect aspect, third person pronouns More negative scores: lack of the positive features
3. Elaborated vs. situation-dependent reference	Reference depending on time and place vs. reference depending on less context-dependent elaboration	More positive scores: relative clauses, nominalizations More negative scores: adverbs, adverbials of time and place
4. Overt expression of persuasion	Overly persuasive or argumentative discourse	More positive scores: many modals, conditionals, verbs that present plans (such as <i>propose</i>) More negative scores: lack of the positive features
5. Impersonal style vs. non-impersonal style	Discourse with more focus on the participants vs. with a focus on events and circumstances	More positive scores: passive constructions and linking adverbials More negative scores: lack of the positive features

*The “positive” and “negative” designations have no evaluative meaning; they are purely artifacts of the factor analysis, designating sets of features that are in complementary distribution. When the positive group of features occur in high frequency, the negative features tend not to.

The class session has many more of the positive features of Dimension 1. First person pronouns are used as the instructor expresses his ideas, and second person pronouns are used in questions that present general conditions for all businesses to assess (not actual questions for the students). Main verb *be* is used (e.g. *what are you good at?*), as are contractions (*I’m, it’s*), the private verb *think* and vague expressions such as *these things*. The passage is in present tense. Of course, negative features of Dimension 1 are also present – noun + attributive adjective combinations such as *good e-commerce* and noun phrases that

incorporate noun, adjectives and prepositions, such as *a whole set of resources*. Information is conveyed, but the overall character of the discourse remains more typical of involved, interactional speech than informational prose.

The textbook, on the other hand, has information that is densely packed into noun phrases and prepositional phrases – *formation of a separate socialist bloc; socialist economic development; a buffer zone of friendly, that is, Communist states on its borders*. The effect of time for planning and revising is apparent in the more careful word choice and more dense packing of information.

Texts can be given a “score” along a dimension by adding together the frequency of the features in the particular text. The class session text above would have a positive score on Dimension 1 because of its relatively high use of the positive features and low use of the negative features, while the textbook would have a very negative score because of its greater use of negative features and relatively low use of positive features. In addition to single texts, many texts from a single register – such as class sessions – can be analyzed and a “mean dimension score” for the register can be calculated. Individual texts or registers can then be compared quantitatively to each other along all five dimensions, giving multiple perspectives on their similarities and differences, as the following example illustrates.

To facilitate multi-dimensional comparisons for Educational Testing Service, we designed a software program to compare the T2K-SWAL corpus to new texts in order to see how typical the new text was of a register. (This program is part of the larger corpus compilation and analysis project funded by Educational Testing Service. See Biber et al., in press, for complete details.) For this example, I consider a sample from an ESL textbook designed to help students prepare for university classes (Dunkel, Pialorsi, & Kozyrev 1996). I chose this book because it was highly recommended by teachers in the intensive English language program in our department. The book’s objective is to help students develop their listening skills for regular university-level courses.

When the program is run, the results show a comparison of the “target text” (in this case, the text from the ESL listening comprehension book) to selected texts in the T2K-SWAL corpus (in this case, the class sessions) along each of the five dimensions. Table 4 displays the output file. For each dimension, the information given includes: the mean for the class sessions (the “Comparison Texts Mean”), the standard deviation for the class sessions (“Comparison Texts SD”), the dimension score for the ESL textbook passage (“Target Text Score”), and whether the passage is within a 95% confidence interval (roughly within two standard deviations of the mean for the class sessions). This confidence interval measure provides a means of seeing how typical or atypical the new text

Table 4. Output of multi-dimensional analysis software program

COMPARISON OF A TARGET TEXT TO THE T2KSWAL CORPUS				
Target text name = C:\ESLCORP\autotagd\LECTURE5.ASC				
NON-LINGUISTIC FEATURE CATEGORIES SELECTED:				
Text Category: CLASS SESSION				
NUMBER OF COMPARISON TEXTS ANALYZED: 176				
Dimension	Comparison Mean	Texts SD	Target Text Score	Within 95% CI?
1	27.66	10.46	-13.42	no
2	-2.28	1.20	-1.44	yes
3	-2.96	2.59	6.02	no
4	2.07	2.44	-1.10	yes
5	-1.16	0.93	4.48	no

is: those outside of the 95% confidence interval are quite unusual, relative to the corpus.

The table shows that for 2 of the 5 dimensions, the lecture passage is within the 95% confidence interval. Thus, the practice ESL passage is similar to the class sessions in that it has relatively few features of narration (Dimension 2 – mostly past tense, perfect aspect, third person pronouns) or overt persuasion (Dimension 4 – primary conditionals, modals and verbs overtly showing persuasion).

On the other hand, the practice lecture passage is very different from the class sessions with respect to Dimensions 1, 3 and 5. The differences are as follows:

- Compared to the class sessions, the practice lecture has a much more negative score on Dimension 1. The practice lecture has more dense packing of information with nouns, attributive adjectives, and prepositions, and a greater diversity in word choice, while the class sessions have more use of the features of interaction and involvement.
- Compared to the class sessions the practice lecture has more use of elaborated reference (a more positive score on Dimension 3). The practice lecture uses more relative clauses and nominalizations while the class sessions have more adverbs and place and time adverbials.
- Compared to the class sessions, the practice lecture has a greater use of features of impersonal style (a more positive score on Dimension 5). There are more passive constructions, with fewer agents and actors than is typical in the class sessions.

Overall, these differences suggest that the practice lecture is more clearly oriented towards conveying information, while the naturally-occurring class sessions have more features of involvement, show more real-time processing constraints, and have more context-dependent reference.

These differences are apparent in a sample of the practice lecture, especially when compared with the class session in (10).

(12) Passage from ESL textbook practice lecture

Culture influences and establishes how people interact with one another (or do not interact with one another). In particular, culture influences the rituals that take place in the classroom setting, and influences the ways students participate in the classroom discourse. It also influences the esteem in which teachers are held... North American students of European origin are usually more talkative in class and more willing to share their opinions than students of Native American heritage or from Asian backgrounds. This difference is directly related to cultural values about learning and education, and classroom behavior. EuroAmerican students' culture teaches them that learning is shaped and helped by their talk and active participation in exploring or discussion issues. Asian students, however, are generally taught that they will learn best by listening to and absorbing the knowledge being given to them by the teacher.

The practice lecture has a number of long noun phrases, incorporating nouns, attributive adjectives and prepositional phrases (e.g. *North American students of European origin; cultural values about learning and education, and classroom behavior*). There are few occurrences of features of "involved discourse," such as the first or second person pronouns, contractions, and private verbs of the class session passage. While the class session uses adverbs such as *here*, the practice lecture is more decontextualized in reference. There are many nominalizations (e.g. *education, behavior, participation*) and more relative clauses modifying nouns (*esteem in which teachers are held*). The practice lecture also has more passive constructions (e.g. *are held, is related, is shaped and helped*). The overall difference in the practice lecture and class session is striking. In many ways, the practice lecture seems more like the textbook sample in (11), rather than the spoken discourse of the class session.

The analysis of this practice lecture is not meant to imply that the textbook, or even the particular passage, will not give students any useful practice with listening comprehension. Rather, it shows that this passage shares some characteristics with class sessions, but not others. The analysis of the T2K-SWAL corpus found that class sessions were a rather unusual register: they have an

informational purpose, but the real-time, face-to-face nature of the discourse gives it a much more interactive, involved character than we typically associate with an informational, academic purpose. An analysis such as this can help teachers and materials writers see that – with this passage – students will not get practice with the more involved, personal, situation-dependent features that are characteristic of class sessions. In designing a book, including lectures with a more involved style would be useful, or a teacher may want to supplement the course book with other class sessions (for instance, on videotape, which may make the reference and other features easier to understand).

Conclusion

An awareness of variation and information from corpus-based research is already being incorporated into classrooms by some teachers. Collections such as Aston (2001), Burnard and McEnery (2000), and Wichmann et al. (1997) provide numerous teaching applications of corpus work, and Tomlinson's (1998) collection for materials writers includes specific articles on incorporating corpus data and activities (see Chapters 1–3).

However, teaching practices that are informed by knowledge of variation in language use are more the exception than the rule. With the foregoing examples I have hoped to show that an awareness of variation is important in all teaching. Even an apparently straightforward task such as presenting common, useful linking adverbials can be misleading without knowing how their use varies across registers. Similarly, not analyzing the language features in texts from a variety of perspectives may result in materials that do not give students practice with some language patterns that are common in the real situations that they are preparing for.

The examples included here show that acknowledging variation in teaching and materials development is not an insurmountable problem. Results of analyses here demonstrate that there are very strong patterns in language use. With such analyses, we find out not only that the answer to most questions about language use is “it depends,” but we also can answer the question “What does it depend on?” This is the question that teachers, materials writers, and students should be asking when they wonder about the use of a language structure.

Acknowledgements

I would like to thank participants at the First Inter-Varietal Applied Corpus Studies (IVACS) International Conference in Limerick, Ireland for their stimulating discussion on the issues raised in this paper.

Note

1. Text samples are taken from the Longman Spoken and Written English Corpus (samples 1–9; see Biber et al. 1999, Chapter 2 for details of the corpus) and the TOEFL 2000 Spoken and Written Academic Language Corpus (samples 10 and 11; see Biber et al. 2002 for details of that corpus).

References

- Aston, G. (2001). *Learning with Corpora*. Houston, TX: Athelstan.
- Atkinson, D. (1996). The philosophical transactions of the Royal Society of London: A sociohistorical discourse analysis. *Language in Society*, 25, 333–371.
- Azar, B. (1999). *Understanding and Using English Grammar* (3rd ed.). New York: Longman.
- Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D. (1993). Co-occurrence patterns among collocations: A tool for corpus-based lexical knowledge acquisition. *Computational Linguistics*, 19, 531–538.
- Biber, D. & Finegan, E. (2001). Intra-textual variation within medical research articles. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-Dimensional Studies* (pp. 108–123). Harlow: Longman.
- Biber, D., S. Conrad & R. Reppen (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, D., S. Conrad, R. Reppen, P. Byrd, & M. Helt (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, 36, 9–48.
- Biber, D., S. Conrad, R. Reppen, P. Byrd, M. Helt, V. Clark, V. Cortes, E. Csomay, & A. Urzua (in press). *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus* [TOEFL Monograph Series]. Princeton, NJ: Educational Testing Service.
- Biber, D., S. Johansson, G. Leech, S. Conrad, & E. Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Burnard, L. & T. McEnery (Eds.). (2000). *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang.
- Carter, R. & M. McCarthy (1995). Grammar and the spoken language. *Applied Linguistics*, 16, 141–158.
- Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly*, 34, 548–560.

Conrad, S. (2001). Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-Dimensional Studies* (pp. 94–107). Harlow: Longman.

Coxhead, A. (2001). A new academic word list. *TESOL Quarterly*, 34, 213–238.

Dunkel, P., F. Pialorsi, & J. Kozyrev (1996). *Advanced Listening Comprehension* (2nd ed.). Boston, MA: Heinle & Heinle.

Frodesen, J. & J. Eyring (1997). *Grammar Dimensions 4: Form, Meaning, and Use* (2nd ed.). Pacific Grove: Heinle & Heinle.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, 19, 115–135.

Hughes, R. & M. McCarthy (1998). From sentence to discourse: Discourse grammar and English language teaching. *TESOL Quarterly*, 32, 263–287.

Hunston, S. & G. Francis (1998). Verbs observed: A corpus-driven pedagogic grammar of English. *Applied Linguistics*, 19, 45–72.

Hunston, S. & G. Francis (2000). *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English* [Studies in Corpus Linguistics 4]. Amsterdam: John Benjamins.

McCarthy, M. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

Preiss, S. (1998). *Northstar* (advanced). New York: Longman.

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

Steer, J. M. & K. A. Carlisi (1998). *The Advanced Grammar Book*. Boston, MA: Heinle & Heinle.

Thompson, G. (1996). *Introducing Functional Grammar*. London: Arnold.

Tomlinson, B. (1998). *Materials Development in Language Teaching*. Cambridge: Cambridge University Press.

Ward, C. (2000). Review of *Longman Grammar of Spoken and Written English*. *RELC Journal*, 13, 165–167.

Wardhaugh, R. (1998). *An Introduction to Sociolinguistics* (3rd ed.). Oxford: Blackwell.

Wichmann, A., S. Fligelstone, T. McNery, & G. Knowles (1997). *Teaching and Language Corpora*. London: Longman.

Williams, G. C. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, 3, 151–171.

Appendix

Summary of the factor analysis from Biber (1988)
(Numbers indicate the factor loadings for each feature on the dimension.)

Dimension 1: Involved versus Informational Production

Involved Production	
private verbs	.96
THAT deletion	.91
contractions	.90

present tense verbs	.86
2nd person pronouns	.86
DO as pro-verb	.82
analytic negation	.78
demonstrative	
pronouns	.76
general emphatics	.74
first person pronouns	.74
pronoun IT	.71
BE as main verb	.71
causative	
subordination	.66
discourse particles	.66
indefinite pronouns	.62
general hedges	.58
amplifiers	.56
sentence relatives	.55
WH questions	.52
possibility modals	.50
non-phrasal	
coordination	.48
WH clauses	.47
final prepositions	.43

Informational Production

nouns	-.80
word length	-.58
prepositions	-.54
type / token ratio	-.54
attributive adjs.	-.47

Dimension 2: Narrative versus Non-Narrative Discourse

Narrative Discourse

past tense verbs	.90
third person pronouns	.73
perfect aspect verbs	.48
public verbs	.43

synthetic negation	.40
present participial clauses	.39

Non-Narrative Discourse
[No negative features]

Dimension 3: Elaborated versus Situation-Dependent Reference

Elaborated Reference	
WH relative clauses on object positions	.63
pied piping constructions	.61
WH relative clauses on subject positions	.45
phrasal coordination	.36
nominalizations	.36

Situation-Dependent Reference	
time adverbials	-.60
place adverbials	-.49
adverbs	-.46

Dimension 4: Overt Expression of Persuasion or Argumentation

Overt Expression of Persuasion or Argumentation	
infinitives	.76
prediction modals	.54
suasive verbs	.49
conditional subordination	.47
necessity modals	.46
split auxiliaries	.44

[No negative features]

Dimension 5: Impersonal Style versus Non-Impersonal Style

Impersonal Style

conjuncts	.48
agentless passives	.43
past participial	
adverbial clauses	.42
BY-passives	.41
past participial	
postnominal clauses	.40
other adverbial	
subordinators	.39

Non-impersonal Style

[No negative features]

Spoken – general

Spoken corpus for an ordinary learner

Anna Mauranen

University of Tampere

Speech corpora have taken longer to make their way to the classroom than their written counterparts, but interest in their use is growing fast. The world of academic English teaching is taking them on board, in response to the need for professionals to communicate in English. Spoken corpora are laborious to compile, and differ in pedagogic terms from written corpora. This paper discusses three distinguishing features: the status with respect to authenticity, corpus utility for immediate communication vs. learning, and the role of a corpus in discovering formulaic expressions. It is argued that spoken corpora score high in authenticity and classroom usability, and that they offer direct access to characteristics of speech, so often inadequately described in textbooks. User-friendly programs and thoughtful teacher education are also called for to ensure the advantages of corpora to all learners. It is also pointed out that what the academic learner needs is likely to be international English rather than a native variety or two, and that corpus compilation should begin to reflect this.

1. Introduction

Enthusiasm for the pedagogic possibilities of corpora has been steadily growing in the last few years. However, as is familiar from many other linguistic and language teaching enterprises, the first steps have been taken almost exclusively in the written domain. So far the pedagogical applications of speech corpora have received scant attention, with few exceptions (e.g. McCarthy 1998, 2001; Swales 2001; Zorzi 2001). In part, this is probably a consequence of there being less descriptive research in spoken than written corpora, but in part there are genuine differences between the most obvious pedagogical applications for speech and writing.

One of the general strengths of corpora is that they can show that which is typical, or common in the language. So for instance if the most frequent use of

the verb ‘to think’ is not with the meaning referring to some ponderous mental process, but with the meaning ‘have an opinion’ (“*I think I had an extremely fair hearing,*” he said afterwards) this can be seen in a corpus and shown to students by a teacher, or be discovered by students on their own.

On the basis of examples like this, we can replace recommendations of language use which are solely based on tradition or teacher intuition. It has become a common finding that what is taught as functional language use is not necessarily in agreement with what is frequent in the language, or even appears at all. Such findings seem to be particularly typical of speech, so that it is not unreasonable to expect corpus data to be helpful in simply providing more relevant information to base pedagogical practices on.

Spoken corpora differ from written in at least three pedagogically relevant ways:

1. As written transcriptions of sound recordings, they can be seen as further removed from their origins than written data is, which poses a challenge to considerations of authenticity.
2. A speech corpus has fewer obvious uses for immediate productive tasks than a written one. Therefore, pedagogical emphases need to shift from those suited for writing.
3. Prefabricated elements have become a standard element in communicative spoken language teaching, while in written language they are still a fairly new discovery. Yet the actual expressions taught as typical and useful rarely have attested origins, and their variability and roles in communication have been very imprecisely described.

I shall discuss each of these issues, with a particular group of learners and teachers in mind, and a particular corpus as well. The target group consists of intermediate or advanced EFL learners with basically an instrumental motivation for using English. The teachers involved are not corpus enthusiasts, but experienced teachers specialising in teaching spoken academic English. The aim is to chart uses and possibilities for a corpus as part of an ordinary spoken English course at an ordinary university language centre, and to discuss issues relating to relevant corpus types and tasks for immediate learning and teaching needs.

As ESP (English for Specific Purposes) continues to be a growth area within EFL all over the world, a university language centre is an increasingly typical learning environment for wide populations of learners of English. After decades of careful exploration of texts for reading and writing, the EAP (English for Academic Purposes) world is beginning to show more interest in

speech. An EAP learner is an interesting learner because he or she is an ordinary learner compared to a university student majoring in English, whose perspective has tended to dominate the discussion of pedagogical corpus applications, despite the pioneering work in pedagogic corpus use by Tim Johns (e.g. 1991, see also Mparutsa et al. 1991).

2. Authenticity

'Authentic' has become a loaded word in applied linguistics. Its positive connotations have been much exploited, as has its semantic bifurcation. Both of its frequent uses are essentially positive. In the first, it is used in a purely factual sense meaning 'genuine', so for instance in arts, a painting or an object may or may not be authentic. This is a dichotomous concept. The second sense is also positively evaluative, more in everyday use, and means that something is so similar to the real thing that it is almost like the thing itself. This is a relative concept. While perhaps the most widely discussed issues have centred around the authenticity of the language brought into the classroom, Widdowson's much repeated idea in relation to L2 learning is that authenticity refers to a learner's response. Let us first look at linguistic authenticity.

Authenticity cannot be brought into the classroom in the strict sense of 'genuine', apart from the language used for classroom management. Whatever records of 'used language' (Brazil 1995) we bring to learners from outside the classroom, they are necessarily severed from their original contexts and therefore inauthentic in the first sense. Yet they may be more or less authentic in the second sense of 'similarity', and whether we find them satisfying depends on our model of language as an object of learning. As any model picks up some aspects of the object to be modelled and ignores others, the diversity of models leads to disputes ranging from questions of whether to use whole discourses vs. extracts to issues concerning the editing out of hesitation, ungrammaticality and misunderstandings. For example, in pursuing marketable authenticity, many textbook producers seem to suffer from a 'facsimile syndrome' which prompts them to reproduce newspaper articles and railway tickets in a faithfully copied outward form. This would appear to attempt an approximation of the 'genuine' sense, but inevitably falls short and counts as a 'copy'.

A more fundamental difference between written and spoken data in the classroom derives from their different degree of completeness as discourse: written text remains interactively incomplete without a reader, while spoken interaction is co-constructed by participants as it unfolds, and is interactively

complete in recordings. This allows more authentic responses in the ‘genuine’ sense to copied writing than to recorded speaking even if the reader is not part of the intended audience of the text; recordings of spoken dialogue retain a status of ‘non-participant observation’. Yet non-participant observation is far from irrelevant to learners – it is a major strategy in learning from audiovisual media, or coping among groups of native speakers. Thus, even if the participant response is missing, recordings of spoken interaction which was not originally produced for purposes of language didactics (and was therefore authentic, or genuine, for the original speakers) have a strong claim to authenticity in the ‘similarity’ sense for classroom learners, whether searched with corpus methods or other means.

Because reproducing linguistic material in writing or sound has become technically easy, a satisfying model for authentic material would currently demand criteria which at least reproduce the language uttered, even if other aspects would have to be sacrificed. Compared to written language, speech corpora appear further removed from the original language event, in being transcribed, which imposes a change of mode, and a reduction of features, like missing out on sound and non-verbal interactive clues. These limitations seriously reduce the information available from the data. Given that overcoming the technical obstacle of sound concordancing in synchrony with the transcript is being conquered, speech corpora will soon provide richer representations, but technical perfection will not alter the impossibility of achieving the genuine.

However, limitations of a similar kind are inevitable whenever we focus in on some aspect of language: by highlighting some aspects we turn our attention away from others. This is unavoidable, given the possibilities of human attention generally and classroom practices specifically (for example talking in the L2 to L1 peers achieves authenticity of dialogic interaction, but loses on the need to use the target language for communication). A speech corpus helps freeze real speech for observation, and, like corpora generally, provides a large number of instances of an item in use. It can therefore achieve a high level of authenticity in the similarity sense by representing crucial aspects of used language, such as the actual wordings and sequences used, with repetitions, overlaps, hesitations and misunderstandings coded. If it can do this, it is worth bringing into the classroom even if it is technically less than perfect, and for example uses a broad transcription system or lacks audio or video recordings.

In this way, a speech corpus can provide data which helps observe aspects of speech whose salience for participants in real ongoing situations may be relatively low (for example because communicative content normally receives

the greatest proportion of conscious attention), but which are nevertheless frequent and play important roles in constructing discourse. Thus the alienating effect that frozen and transcribed speech produces may be pedagogically as useful as examples traditionally have been – for demonstration, temporary isolation and focus. But unlike traditional examples presented for special attention, these are attested instances of language use, and allow us to change the amount of context and the foci of the searches, by learners or by teachers.

So far, we have considered authenticity as a property of language brought into the classroom. However, the main argument against corpus authenticity as Widdowson (e.g. 2000) has repeatedly presented it, rests on the notion that in a pedagogic context authenticity is not an objective characteristic of texts, but one which users bring to texts, by their ability and willingness to re-contextualise them. His objection to corpus data (as well as other ‘real’ language) is their lack of authenticity for students, on account of the impossibility of replicating the original contextual conditions of the language in the classroom. Obviously, it must be accepted that contextual conditions cannot be replicated in the classroom – or any other place apart from the original context of utterance. This is also conceded by McCarthy (2001), who nevertheless criticises Widdowson for talking from the viewpoint of the savant applied linguist, not the learner, and states that “What we really need to test authenticity and relevance is learners’ own reports.” (2001: 138). This is an important point, independently of how reliably obtainable or interpretable learner’s reports may be, but, sadly, McCarthy goes on to do exactly what he accuses Widdowson of – speculates on behalf of the learners whose response he is appealing to. The dispute is disturbingly similar to the issue of ‘equivalence’ in translational theory when it is defined not as a linguistic similarity, but as ‘an equivalent response in the target audience’. To my knowledge, nobody has carried out comparative research on such responses.

Measuring learner response is indeed welcome, but it is extremely likely that the experiencing of classroom material as ‘authentic’ or ‘sufficiently authentic’ by learners is a matter of pedagogical practice as much as it is of the actual language material in the classroom. Classroom learning, in any subject, is based on a certain suspension of disbelief. Learners are socialised into certain models of learning and classroom credibility; learner response is a social response, contextualised and situated, and it is futile to try to unearth a ‘pure’ response from ‘the universal learner’. Authenticity as a learner response is, therefore, like other classroom practices, socially negotiable and historically changeable. Educators can promote learners’ acceptance of corpora in

the classroom as authentic data – or reduce it, but it is a matter of conscious pedagogic choice, not a law of nature.

A corpus can also model authentic learning in other ways. Faced with corpus data, the learner's experience resembles in certain ways that which he or she has in an L2 situation, surrounded by masses of L2 language, of which there is an urgent need to make sense. In both situations, learners need good strategies to make sense of the relevant bits of the language, as well as to sort out what might be learned. In the first situation, communication strategies are likely to predominate, in the case of the corpus perhaps the latter. This is another naturalistic aspect of corpus data, strengthening the authenticity of learner response in the 'similarity' sense. The main limitation of corpus experience then derives from the fact that the language mass is not encountered in unfolding situations where the learner is a participant. The similarity, and the strength, lies in exposure to masses of naturally-occurring L2 data.

In any case, approximating naturalistic learning is not a central goal in itself for classroom learning. It is not at all clear that naturalism has any wholesale advantages over classroom learning in second language acquisition – there is evidence to support both sides. In a classroom, corpus data is light years ahead of invented examples in authenticity, and to make the most use of that data is a matter of pedagogic intervention in the learning process – which largely entails focusing the learner's attention on what might be relevant.

3. Communicative utility

A written corpus can be used for a number of practical purposes (reading, writing, translation) to solve immediate problems of language use. Thus, it can prove its usefulness to L2 (and why not L1) users with no reference necessarily to actual *learning* results.

An example of communicative utility without language learning motives can be seen in a study on professional translators working in industry who were introduced to using corpora as part of their work (Jääskeläinen & Mauranen 2000, 2001). Although they had been using computers for virtually all their work, they had never yet heard of corpora. Having been taught basic corpus skills and being provided with corpora, they kept a journal of their use of the corpus for a couple of weeks, from which emerged the most typical uses summarised here (Jääskeläinen & Mauranen 2001):

1. Most often for text production:
 - Seeking confirmation to their own intuition
 - Checking variants suggested by reference works (dictionaries, wordlists)
 - Checking which prepositions and /or adjectives go together with a term
 - Checking different uses of certain terms
 - Looking for idiomatic expressions
 - Learning how to use a new expression
2. For text comprehension: trying to figure out the meaning of an expression from the context

Any of these uses would equally well suit students faced with a writing task, such as translation, composition, essay, or report. In other words, the L2 user need not internalise the outcome of a particular search or store it in a readily accessible form in his or her memory, and still the corpus provides a rich source for solving communicative problems.

In the case of a speech corpus, this option is more limited. Clearly, when preparing for a spoken presentation, a learner may use a speech corpus as reference material in a similar way. But equally clearly nobody can carry out a conversation by using a corpus for help – the communication must be delayed in one way or another for corpora to be of use, just as is the case with most reference materials. A speech corpus can be extremely useful for checking the meaning of a puzzling expression heard in speech but not adequately described in reference materials, or in preparing a spoken presentation, but in both cases the immediate communication is removed in time from using the corpus. Given the increasing need in many walks of life to prepare spoken presentations, I can see openings for specific purpose speech corpora in both the academic and the business world, for example, but it is hardly fruitful to conceptualise professionals as ‘learners’ or to see this as part of the general pedagogical usefulness of corpora. It seems, then, that the emphasis with speech corpora must be on accumulating factual knowledge about typical ways of expressing something in the target language, along with developing fairly high-level cognitive skills, such as ‘language awareness’.

4. Formulaic expressions

The third difference between spoken and written data mentioned above is the position of prefabricated elements (or gambits, formulae, lexical phrases –

these semi-fixed, semi-formulaic expressions go under many names and definitions but are roughly comparable) which have long been a standard element in communicative language teaching, and in teaching spoken skills in general. Very often, though, teaching prefabs is based on questionable data, such as materials writers' intuition or tradition. Many researchers have shown that a number of the routine-like formulae which have for a long time constituted the staple diet of spoken language teaching simply do not occur in real speech (e.g. Aston & Burnard 1998; Burdine 2001; Lawson 2001), and even if they do, they tend to be infrequent or show internal variation. More transparent prefabs, involving combinations of individual words with their typical lexical and grammatical environments (e.g. *let me/us just say/mention/add/digress...*), have been excluded from many traditional categories of idioms or prefabs. Such conceptualisation of routine-like expressions in speech is in marked contrast to prevailing approaches to the written language, where the myth of elegant variation is much more persistent, and the discovery of semi-fixed expressions is in its fairly early stages. Apparently also repeated chunks are more characteristic of speech than writing. The teaching of speech thus has its special problems with respect to these items.

Prefabs or formulae need to be empirically validated for adequate description, since intuition does not seem to be a particularly reliable guide to their frequency or variability. Pedagogic descriptions tend to be far from adequate in L2 teaching materials, and despite ideological lip-service to the priority of speech, spoken language is not often very realistically depicted in textbooks or reference books. One example from the Michigan Corpus of Spoken Academic English (MICASE, see Simpson et al. 1999) to illustrate the point: in all the reference books I have found, the meaning of BE + *the case* has been something to the effect that by saying that something is the case, the speaker means that he or she believes this to be true. Yet the examples in the corpus are virtually all associated with non-factuality, i.e. something either not being the case, or being so conditionally or speculatively:

(1)

i just wanna say that *is not always the case*. and i can give, uh one a contradiction. cuz contradictions just *can't be the case*. so like if, let's that's, no one knows if that might be, *might be the case*. but that is a this was a revolution age. maybe. and if *if that's the case* then we're living well *i hope that's that's the case*. i expect that would be the case

The exception, as one might surmise, is where the positive meaning receives particular emphasis:

- (2) she would again have to support me and *indeed that was the case* because my father died when i was in high school

Spoken language is therefore a domain where learners need to work out many linguistic features on their own, because they cannot expect enough help from textbooks, teachers, or reference materials.

Zorzi (2001) mentions discourse markers in speech as an example, and she may well be right in assuming that on the whole these are not adequately described for learners. Yet in EAP teaching they receive a reasonable amount of attention even in speech, no doubt influenced by the fact that written discourse markers have been highlighted in academic reading and writing courses for quite a while. But in the description of frequent prefabs, academic speech is largely virgin territory. Descriptions emerging from the MICASE and the T2K-SWAL (see, Biber et al. 2001) corpora have not found their way to teaching materials on a large scale yet. Thus, as we can expect prefabs to cover around 70% of ongoing speech (Erman & Warren 2000), the learner is faced with a formidable task.

Prefabs are normally thought to be usefully memorised as unanalysed wholes, which are then assumed by some scholars (e.g. Nattinger & DeCarico 1992; Aston 1995) to transform into analytical structures as the learner's repertoire develops. It seems that the chunking of linguistic material according to repeated sequential units is spontaneous and part of normal language processing. Chunking builds up long-term memory representations of repeated sequential units (such as prefabs), and frees processing capacity for analysing new meanings encountered in ongoing discourse (Ellis 1996). Moreover, socio-interactive benefits also appear to follow from using prefabs, in helping speakers manage discourse situations (Wray & Perkins 2000; Wray 2000). In these respects it seems that the same spontaneous processes take place in both first and second language acquisition. Whether explicit learning (or teaching) can aid such processes of implicit learning is less clear, and the assumption that chunks stored in memory turn into analytical structures as they seem to do for L1 learners remains controversial (e.g. Wray 1999, 2002).

The assumption that units memorised as wholes for routine usage can be subsequently reanalysed into productive rules in L2 learning is reminiscent of McLaughlin's (1987) information processing model in second-language learning. McLaughlin postulates that sequences are automatised as a result of repeated activation, stored in the long-term memory, and eventually accompanied by a restructuring of the learner's linguistic system. It is unclear whether a learner's automatised chunks are reorganised internally, which presumably

would be detrimental to their continued availability as single units, or whether the automatised units remain intact in themselves but they or their individual elements enter into new configurations in the system. The latter would seem to be more plausible, allowing for multiple representations of items, which is a prominent feature in Wray's (2002) model of formulaic language.

Insofar as encountering sequential units of the target language develops a sense of frequently repeated sequences in the language, a corpus search would seem to offer an ideal source for learning. However, the kind of repetition that leads to a transfer of units into the long-term memory is not likely to be the kind of repetition achieved by a corpus search, which gives all the instances at once. Therefore it is better suited for bringing a pattern to a learner's attention in the sense that Schmidt (1994) refers to as 'noticing', i.e. bringing some stimulus into focal attention, without which learning cannot take place.

To what extent could or should pedagogics step in and facilitate the creation of appropriate chunking, or help construct multiple representations, if these are spontaneous processes? That is, if learners are sensitive to input regularities and extract probabilistic patterns on that basis, to what extent can pedagogic intervention help this process?

For some learners, this may be more useful than for others, since it is likely that different learner types weight gestalt learning and analytic learning differently in L2, as they do in L1 (see, Wray 1999, 2002). On the whole, it looks like research evidence is basically in support of form-focussed classroom instruction (Lightbown 2000). Even if learners construct their own rules, learning with corpora is explicit and largely form-focussed learning, with conscious effort to discover ways of saying things in the target language. But as is well known, learners may know a rule in the target language without being able to use it in real-life communication. For delayed communication this is not a problem because a rule or regularity needn't even be learned, it just needs to be observed to be applied, but for speech, seeing the relevant phraseology on screen cannot be expected to turn into fluency without intermediate steps of applying the discovery to tasks requiring its use. If the psycholinguistic reality of prefabs is that they are stored and retrieved as unanalysed wholes (even given multiple representations to take care of variability), which then confer communicative benefits to users, an analytic awareness cannot suffice to render the same usefulness. It has been quite well established in L2 acquisition research that input in itself, without interaction, does not lead to acquisition. Thus, corpora offer excellent data to work with, but transforming awareness or knowledge into capacity is as difficult as in language learning generally.

Native-like idiomaticity has frequently been brought up as a target for teaching formulaic language, with or without corpora. Leaving aside questions of imperfect learning, we may not be improving accuracy with corpus use, since there is no evidence that an adult L2 learner's tendency to overgeneralise can be prevented by corpus data. Moreover, fluency in learners may, in principle, follow from spontaneous chunking which is not native-like, and not like that which has been taught in the classroom, but still confer similar advantages to the learner as it does to the native speaker.

Where a corpus provides a radical advantage is in promoting a more relevant perspective on the language to be learnt than more traditional divisions into morphosyntax, lexis, and unanalysable fixed idioms. If the multi-word units of usage are indeed the same as those of storage and access (e.g. Bybee & Hopper 2001), clearly it makes sense to operate primarily with those units in learning and teaching rather than with other units based on traditions of grammatical or lexical analysis.

5. Taking the corpus to the classroom

No teaching method can become an important innovation, whatever its potential, if it does not make its way to the normal classroom where teachers and students can use it as part of their everyday routines, with not too much extra hassle.

Even if we wish to be maximally learner-centred, or construct the learner as a 'researcher', he or she needs skills and guidance in dealing with the kind of data a corpus provides. Noticing things in corpus data is an acquired skill even for linguistically relatively sophisticated learners like L2 majors in university departments (as for example Bernardini's (2000) students). Less sophisticated students need even more tuition in making observations from corpus material. Initially, even making sense of a set of pre-edited concordance lines tends to be hard for many learners, although it soon becomes an easy routine. A much more demanding task is to make meaningful observations on unedited (i.e. unselected) concordances independently, and the hardest of all is learning to ask appropriate questions of the corpus. All this is learnable, but what I want to emphasise here is that corpus skills constitute a learning task in themselves, much in the way that many other subskills of learning do, such as group work skills. Once acquired, they facilitate learning greatly and need not be constantly refreshed.

Before learners can be introduced to good corpus skills, their teachers need to possess them in the first place. The first experiences with EAP teachers interested in using the MICASE corpus point to a considerable practical and attitudinal hurdle between teachers' general willingness to adopt this new approach and their actual taking it on board. One of the main issues seems that using a corpus differs crucially from what teachers are used to, not only technically but above all in terms of thinking about language. If teachers are not familiar with corpus data from their own education, it constitutes a radical departure from their normal ways of looking at language, as evidenced in their spontaneous questions and hopes for a corpus. Teachers' questions and wishes tend to be functional or pragmatic (like what discourse markers can be found in the corpus, how to express disagreement, or what is a good way of beginning or ending a presentation), and hard to transform into searchable strings, even with the help of regular expressions, batch searches, or grammatical annotation. If teachers are uncertain about optimal (or even possible) search procedures, they are likely to lose interest fast or maintain that it does not work for their purposes, or their students. Neither will they be able to pass them on to their students so that the students would feel confident in finding what they are looking for, or asking the kinds of questions that are answerable with corpus methods.

Teachers need awareness of tasks that are sensible and manageable for their pupils. Sometimes the search tasks given to students would be hard for professional linguists (e.g. different uses of very frequent verbs like *think* or *set*), let alone undergraduates or schoolchildren. Impossible tasks will certainly turn any novice against corpora. Working with translators, we noticed (Jääskeläinen & Mauraanen 2000, 2001) that introducing professionals to corpora can be a harder task than initiating undergraduates. Careful initiation to corpus skills – or most of all corpus thinking – for the classroom teacher is crucial before corpora can conquer ordinary learners in schools and universities. Useful suggestions like the checklists for students that Zorzi (2001) presents for showing relevant contextual features for the meanings of discourse markers certainly constitute good pedagogical practice, but do not suffice for corpus introduction to newcomers. What we need is user-friendly program packages, with simple concordancers and good initiation exercises. Making corpora a normal part of teacher education will certainly serve to establish corpora in the classroom, but to speed things up, in-service courses for practising teachers should be taken seriously.

When teachers and learners are given strategic skills for thinking in terms of corpora, they may and will draw their own conclusions from the data, which

may not be descriptions of the kind that would satisfy a linguist. This need not be a problem, since descriptions by teachers and learners themselves needn't be fully accurate, as Aston (2001: 11) points out; learning involves gradual approximation to the target system, so learners may find their descriptions useful, even though they are partial and approximate. Moreover, to be realistic, we must remember that pedagogical grammars have always relied on simplified, approximate rules – if learners and teachers construct such rules on an *ad hoc* basis for themselves, this may just mean an increase in the relevance of such rules for themselves, not necessarily a drastic reduction in accuracy. Most textbook writers are not professional linguists but experienced teachers – so reliability, accuracy and coverage are not likely to suffer enormously in the hands of ordinary teachers and learners.

An issue that has come up in consultations with practising teachers concerns the kind of corpus that is found to be most useful. For EAP teaching, it is not surprising that an academic speech corpus has been welcomed by teachers. Interestingly, they have also expressed a wish that their own students' presentations and discussions would be available as transcripts for closer scrutiny. Moreover, the question of what kind of speech a 'model corpus' should consist of has come up a few times. As English is increasingly used as a lingua franca among non-native speakers, it is becoming more and more obvious that sensible targets and norms of usage cannot be those designed for a native speaker (see, e.g. Knapp & Meierkord 2002; Seidlhofer 2000). For pronunciation, important pioneering work has been done by Jennifer Jenkins (2000), and similar norms should be developed in other domains of language use, based on what McCarthy (2001) calls "the expert user", rather than the native speaker (see also Mauranen 2003). Accepting more relaxed targets is particularly important for speech corpora, because speech cannot be edited after production like writing, and demanding strict native-like accuracy can be more frustrating than helpful to learners. Speech corpora are also more difficult to compile than written corpora, and practitioners are less likely to compile their own.

Given the labour-intensive nature of compiling a spoken corpus, there is little chance of achieving results by encouraging all teachers or learners to create their own *ad hoc* databases, as is quite sensible with written text these days. Therefore, we need speech corpora compiled for pedagogic use by those who have the resources for doing so, and negotiations about what is really needed in the classrooms should be carried out with practitioners in different countries. It is not necessarily the case that the target corpora should be left to the discretion of native speakers.

6. Hands-on

A number of problems in finding suitable L2 expressions arise during class discussions, presentations, and simulations. Skilful corpus users can be sent off to a computer for solutions, but most ordinary users need more controlled procedures. One point of departure is a feedback session on a recorded performance, where students, teachers, or both can comment on troublespots on tape as they occur. A subcorpus of target event types (say, meetings, or seminar presentations) can be then consulted for solutions. For instance, if the MICASE is consulted for expressions of disagreement by searching for **agree**, it soon becomes clear that *agree* is mostly used in positive expressions, but it also fairly often expresses partial agreement and is followed by *but...*, whereas *I disagree* is rare.

Alternatively, two or three transcripts of target situations may be taken up before looking at the corpus, and suitable-looking expressions can then be searched to check on meaning, frequency, typical context, etc. This procedure helps answer questions of a functional nature, such as how to elicit questions (most commonly *Any questions?*), how to initiate one (frequently *I have a question* in the US context), or how to begin or end a presentation, a meeting, a discussion... Searches for question forms such as *may I*, *can I* and *could I...* show similarities in being used for managing the discourse situation, but differences in frequency and some uses (*could I* is not often followed by a reply). Typical replies can also be found by searching for questions (*may I (ask)...* is nearly always followed by *sure*).

Students in an EAP situation also sometimes like to borrow an MD recorder and make recordings of situations of their own choice – like international student groups preparing a task, or lectures from their courses. These can be played back for classroom elaboration, including corpus work. The corpus can of course also serve as the point of departure, and after discussing uses and patterns of a particular phenomenon, students can observe instances on tape. This stage can also reveal gaps in the understanding derived from corpus examples, like judging the degree of politeness in different pronunciation variants (see, Lindemann & Mauranen 2001).

7. Conclusion

Corpora are particularly well suited for observing repeated sequences and patterning in language, independently of traditional linguistic categorisation. This

is one of their major advantages in pedagogic application. It looks like multi-word units of some kind are those that we basically use for language storage and access, and if this is so, it gives a motive for providing learners with data that foregrounds these units in the target language.

In spoken language, learners get less help from standard pedagogic descriptions than they do for writing, and therefore often need to work out the use of linguistic features for themselves. This is clearly a point where corpora can help. For noticing patterning in speech, it is helpful to be able to freeze a large number of instances for observation, since it may be very difficult to pay attention to such recurrences in ongoing interaction. At the same time, the need to cope with a large mass of language data resembles the learner's real-life tasks.

The main strength of a speech corpus is in presenting linguistic material which is high in authenticity as a representation of the target language. Yet transforming awareness or knowledge into capacity is a problem which can hardly be solved with corpora alone. As in any pedagogic context, the ultimate learning outcome is dependent on several interacting factors, above all the opportunities of applying the incipient knowledge and skills to meaningful tasks. A corpus cannot offer a panacea to language learning problems. But it can offer a rich database, both for learners and for those who need it as a resource for preparing spoken presentations.

Many technological innovations have had a revolutionary effect on L2 learning – the tape recorder, the xerox machine, the video – all devices which have had a considerable impact on linguistic research as well. Interestingly, all of these have made their impact by making language recordable, by arresting the process and turning it temporarily into a product which can be scrutinised with time and from many angles. Inevitably, such procedures abstract language away from its ongoing flow – but this is the normal prerequisite for observational focus. So clearly the computer and access to corpora is no different from other recording devices in this respect. And like them, it will find a permanent place at the centre of language learning materials.

For enabling learners to benefit from speech corpora, it would be crucial to compile such corpora and make them generally available for learner use, since it is unrealistic to expect every teacher or school to do this for themselves. An excellent example for EAP teaching is the MICASE corpus, which is freely available to every user. On-line services, which are available for some large corpora, remain rather expensive for ordinary learners, and educational institutions tend not to be very wealthy anywhere in the world. A need for inexpensive speech databases is quite clear. Something like that is available in the BNC Sampler (or the International version of the British National Corpus),

but they are quite limited in terms of variety – international learners are not primarily in need of British models, but a sensible range of more international varieties, including non-native expert use. Corpora for learners should be compiled in consultation with classroom practitioners and people with educational expertise from different parts of the world.

References

- Aston, G. (1995). Corpora in language pedagogy: Matching theory and practice. In G. Cook & B. Seidlhofer (Eds.), *Theory and Practice in Applied Linguistics* (pp. 257–270). Oxford: Oxford University Press.
- Aston, G. (2001). Learning with corpora: An overview. In G. Aston (Ed.), *Learning with Corpora* (pp. 7–45). Bologna: CLUEB.
- Aston, G. & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.
- Bernardini, S. (2000). *Competence, Capacity, Corpora*. Bologna: CLUEB.
- Biber, D., R. Reppen, V. Clark, & J. Walter (2001). Representing spoken language in university settings: The design and construction of the spoken component of the T2K-SWAL Corpus. In R. C. Simpson & J. M. Swales (Eds.), *Corpus Linguistics in North America* (pp. 48–57). Ann Arbor, MI: University of Michigan Press.
- Brazil, D. (1995). *A Grammar of Speech*. Oxford: Oxford University Press.
- Burdine, S. (2001). The lexical phrase as pedagogical tool: Teaching disagreement strategies in ESL. In R. C. Simpson & J. M. Swales (Eds.), *Corpus Linguistics in North America* (pp. 195–210). Ann Arbor, MI: Michigan University Press.
- Bybee, J. & P. Hopper (2001). Introduction to frequency and the emergence of linguistic structure. In J. Bybee & P. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 1–24). Amsterdam: John Benjamins.
- Ellis, N. C. (1996). Sequencing in SLA. Phonological memory, chunking, and points of order. *SSLA*, 18, 91–126.
- Erman, B. & B. Warren (2000). The idiom principle and the open choice principle. *Text*, 20(1), 87–120.
- Jääskeläinen, R. & A. Mauranen (2000). Development of a corpus for the timber industry. Project SPIRIT. Unpublished research report, Savonlinna School of Translation Studies, University of Joensuu.
- Jääskeläinen, R. & A. Mauranen (2001). Kääntäjät ja kieliteknologia – kokemuksiä työelämästä (= Translators and language technology – experience from work). In M. Charles & P. Hiidenmaa (Eds.), *Tietotyön yhteiskunta – kielten valtakunta* [AfinLA Yearbook 2001] (pp. 358–370). Jyväskylä: AFinla.
- Jenkins, J. (2000). *The Phonology of English as an International Language*. Oxford: Oxford University Press.
- Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (Eds.), *Classroom Concordancing* [ELR Journal 4] (pp. 1–16).

- Knapp, K. & C. Meierkord (Eds.). (2002). *Lingua Franca Communication*. Frankfurt: Peter Lang.
- Lawson, A. (2001). Rethinking French grammar for pedagogy: The contribution of spoken corpora. In R. C. Simpson & J. M. Swales (Eds.), *Corpus Linguistics in North America* (pp. 179–194). Ann Arbor, MI: University of Michigan Press.
- Lightbown, P. (2000). Classroom SLA research and second language teaching. *Applied Linguistics*, 21(4), 431–462.
- Lindemann, S. & A. Mauranen (2001). “It’s just real messy”: The occurrence and function of *just* in a corpus of academic speech. *English for Specific Purposes*, 20(1), 459–476.
- Mauranen, A. (2003). The corpus of English as lingua franca in academic settings. *TESOL Quarterly*, 37, 513–527.
- McCarthy, M. (1998). *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.
- McCarthy, M. (2001). *Issues in Applied Linguistics*. Cambridge: Cambridge University Press.
- McLaughlin, B. (1987). *Theories of Second-Language Learning*. London: Arnold.
- Mparutsa, C., A. Love, & A. Morrison (1991). Bringing concord to the ESP classroom. In T. Johns & P. King (Eds.), *Classroom Concordancing* [ELR Journal 4] (pp. 15–134).
- Nattinger, J. R. & J. S. DeCarrico (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 129–158.
- Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review*, 14, 357–385.
- Seidlhofer, B. (2000). Mind the gap: English as a mother tongue vs. English as a lingua franca. *Vienna English Working Papers*, 9, 51–68.
- Simpson, R. C., S. L. Briggs, J. Ovens, & J. M. Swales (1999). The Michigan corpus of academic spoken English. Ann Arbor, MI: The Regents of the University of Michigan.
- Swales, J. (2001). Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (Ed.), *Academic Discourse* (pp. 153–167). London: Longman.
- Widdowson, H. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21(1), 3–25.
- Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching*, 32, 213–231.
- Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, 21(4), 463–489.
- Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.
- Wray, A. & M. R. Perkins (2000). The functions of formulaic language: an integrated model. *Language and Communication*, 20(1), 1–28.
- Zorzi, D. (2001). The pedagogic use of spoken corpora: Introducing corpus concordancing in the classroom. In G. Aston (Ed.), *Learning with Corpora* (pp. 85–107). Bologna: CLUEB.

Spoken – an example

The use of concordancing in the teaching of Portuguese

Luísa Alice Santos Pereira

University of Lisbon

As an example of a large and recently-established spoken and written corpus, and what can be done with it, this paper describes resource-building of the Portuguese language at the University of Lisbon, and some possibilities envisaged for applications such as language teaching. Portuguese is one of the most widespread languages of the world, with the fifth largest group of native speakers. The corpus linguistics team at the Centro de Linguística da Universidade de Lisboa (CLUL) has been accumulating resources for some years, and has now made them available to the profession. One of their most ambitious publications is a set of 4 CD-ROMs, “Português Falado”, containing large samples of spoken Portuguese from the many countries where this language is in daily use. These samples are presented with text-to-sound alignment. Several examples are presented from the written corpus, showing the kind of information that is only obtainable from a corpus, and which is of great value to language learners and teachers, as well as to other professional users of language data. The differing frequencies of forms and lemmas are an important feature of an inflected language, and the collocation profiles of near-synonyms are directly usable in the classroom. The paper provides information about the corpus and its availability at CLUL.

Introduction

The *Corpus de Referência do Português Contemporâneo* (CRPC), now with 201 million words, is a corpus consisting of samples of Portuguese discourse, both spoken and written, from all its national varieties, and those from Goa, Macao and East Timor. It is being developed in ways laid out in Appendix 1.

This corpus is more and more in demand as the basis of linguistic research projects such as studies of different linguistic areas, theoretical and applied, lexicon construction and lexicographic work.

Linguistic resources

At the *Centro de Linguística da Universidade de Lisboa* (CLUL) a growing portfolio of materials is being developed. The present state of the work is set out in Appendix 2.

From the CRPC, two of the principal resources that we have made are the *Léxico Multifuncional Computorizado do Português Contemporâneo* and *Português Falado*.

The first resource is a general Lexicon of Portuguese, which contains 26,980 lexical entries and associated forms (140,976) with grammatical (morphosyntactical) and quantitative information. The lexicon consists of the lexical entries that have a frequency of 6 or more in the corpus, followed by all the associated forms (inflected forms and compounds).

This Lexicon is based on a 16.2 million word contemporary Portuguese sub-corpus, written (15.35 M words) and spoken (0.86 M words), extracted from the CRPC. The sub-corpus was designed according to the principles generally established and the international recommendations on the dimension and design of general purpose linguistic corpora used for lexical extraction.

The *Léxico Multifuncional Computorizado do Português Contemporâneo* is a very useful product as a basis for different applications such as dictionaries, translation and NLP. It is also a useful tool with which teachers can explore the language. It is now available on-line in PDF format sorted in alphabetical order and in decreasing frequency order at the site of the CLUL (www.clul.ul.pt/). Tables 1 and 2 show the layout of the lexicon and Table 3 gives explanations of the respective symbols.

The second resource is *Português Falado*, the most recent compilation of CLUL/Camões Institut. It consists of authentic spoken texts and it aims mainly to develop the capacity of Portuguese language understanding and production between foreign students of medium or high level Portuguese studies. The materials now published contribute to the observation and analysis of spoken Portuguese in its geographical varieties. Thus, they are also very useful for teachers, translators, interpreters and researchers in general.

The 86 texts made available on CD_ROM are examples of spoken Portuguese varieties from Portugal (30), Brazil (20), Angola (5), Cape Vert (5), Guinea-Bissau (5), Mozambique (5), S. Tomé e Príncipe (5), Macao (5), East-Timor (3) and Goa (3) as presented in Figure 1.

These texts were recorded in several communicative situations from informal conversations to more formal discourse, such as radio programmes, for instance. The recordings were transcribed using normal orthographic conven-

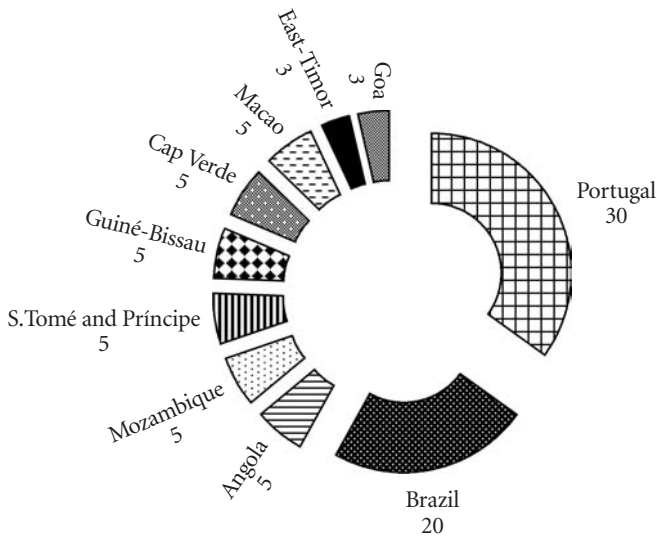


Figure 1. Geographical distribution of spoken text types

tions, and the transcriptions were aligned with corresponding points on the sound wave. To make it easier for the to user to hear the recording and simultaneously read the corresponding transcription on the computer screen, a coloured light runs over the transcription of the sequence which is being

Table 1. Alphabetical order of the lemmas

@ maca (N)	■□□□□	macabro (A)	●●○○○○
maca (N)	●○○○○○	macabros (A)	●○○○○○
macas (N)	●○○○○○		
@ maça (N)	■□□□□	@ macaco (A)	■□□□□
maça (N)	●○○○○○	macaco (A)	●○○○○○
maças (N)	●○○○○○		
@ maçã (N)	■■□□□	@ macaco (N)	■■□□□
maçã (N)	●●○○○○	macaca (N)	●○○○○○
maças (N)	●●○○○○	macacas (N)	○○○○○○
maçãzinhas (N)	○○○○○○	macaco (N)	●●○○○○
		macacos (N)	●●○○○○
@ macabro (A)	■■□□□	macaquinha (N)	○○○○○○
macabra (A)	●○○○○○	macaquinho (N)	●○○○○○
macabras (A)	●○○○○○	macaquinhos (N)	○○○○○○
		macaquitos(N)	○○○○○○

Table 2. Decreasing frequency order of the lemmas

@ funerário (A)	■ ■ □ □ □ □	@ raciocínio(N)	■ ■ □ □ □ □
funerários (A)	● ● ○ ○ ○ ○	raciocínio (N)	● ● ○ ○ ○ ○
funerárias (A)	● ① ○ ○ ○ ○	raciocínios (N)	● ① ○ ○ ○ ○
funerário (A)	● ① ○ ○ ○ ○		
funerária (A)	● ① ○ ○ ○ ○	@ russo (N)	■ ■ □ □ □ □
		russos (N)	● ● ○ ○ ○ ○
@ invisível (A)	■ ■ □ □ □ □	russo (N)	● ① ○ ○ ○ ○
invisível (A)	● ● ○ ○ ○ ○		
invisíveis (A)	● ● ○ ○ ○ ○	russo-japonesa (N)	○ ○ ○ ○ ○ ○
@ maçã (N)	■ ■ □ □ □ □	@ severo (A)	■ ■ □ □ □ □
maçã (N)	● ● ○ ○ ○ ○	severo (A)	● ● ○ ○ ○ ○
maças (N)	● ● ○ ○ ○ ○	severa (A)	● ① ○ ○ ○ ○
maçazinhas (N)	○ ○ ○ ○ ○ ○	severas (A)	● ① ○ ○ ○ ○
		severos (A)	● ① ○ ○ ○ ○
		severíssima (A)	○ ○ ○ ○ ○ ○

Table 3. Explanation of symbols in Tables 1 and 2

Frequency bands ($\log_{10}/2$):			
Lemmas:		Forms:	
6–10	■ □ □ □ □ □	1–5	○ ○ ○ ○ ○ ○
11–31	■ □ □ □ □ □	6–10	● ① ○ ○ ○ ○ ○
32–100	■ ■ □ □ □ □	11–31	● ○ ○ ○ ○ ○
101–316	■ ■ □ □ □ □	32–100	● ① ○ ○ ○ ○ ○
317–1000	■ ■ ■ □ □ □	101–316	● ● ○ ○ ○ ○ ○
1001–3162	■ ■ ■ □ □ □	317–1000	● ● ① ○ ○ ○ ○
3163–10000	■ ■ ■ ■ □ □	1001–3162	● ● ● ○ ○ ○ ○
10001–31622	■ ■ ■ ■ □ □	3163–10000	● ● ● ① ○ ○ ○
31623–100000	■ ■ ■ ■ ■ □	10001–31622	● ● ● ● ○ ○ ○
100001–316227	■ ■ ■ ■ ■ □	31623–100000	● ● ● ● ① ○ ○
316228–1000000	■ ■ ■ ■ ■ ■	100001–316227	■ ■ ■ □ □ □ □
1000001–3162277	■ ■ ■ ■ ■ ■	316228–1000000	● ● ● ● ● ①
Codes:			
Noun	N	Abbreviation	X
Verb	V	Acronimous/Sigla	G
Adjective	A	Symbol_	B
Pronoun	P	Se medio-passive	U
Article	T	Element of group	L
Adverb	R	Enphasys Particle	E
Preposition	S	Element out of order	_d
Conjunction	C	Non-conventional writing	*
Numeral	M	Contraction	+
Interjection	I	Head of lemma*	@
Foreign word	F		

*Head of lemma reconstituted because it did not occur in the corpus

listened to. The user can control what he or she is listening to, can repeat sequences or jump to parts of the text (Gonçalves & Veloso 2000). It is published on 4 CD-ROMs available at CLUL (fbacelar.nascimento@clul.ul.pt) or at the *Instituto Camões* (ded@instituto-camoes.pt).

Resources for corpus-based teaching

This paper focuses on these two products not only because they are the most recent ones available from CLUL, but because of their usefulness in corpus-based teaching. However, the potential of a corpus is endless and teachers can find in it significant resources for new approaches to language teaching.

As an example using frequency information alone, the comparative results of occurrences of the verb *deduzir*, observed in 6 CRPC sub-corpora of comparable dimensions (c. 1 M words) are presented in Tables 4 and 5 below. The analysis aims at observing the frequencies of one lemma and its forms, and also the occurrence of its syntactic, semantic and discourse uses. In this way, comparative tables of the verb *deduzir* in different discourse contexts, both spoken (SPK.) and written are shown: journalistic (JOURN.), literary (LIT.), didactic (DID.), economic (ECON.) and juridical (JUR.) (cf. Bacelar do Nascimento 2001a).

Evidence from concordances can show important distributional patterns at both the syntactic and the semantic level.

It can be seen that in some cases the syntactic places are filled with repeated specific lexical items, and with the same word forms with different meanings. An example for the verb *deduzir* is given, in which the juridical discourse is highlighted. The repeated selection of forms of this word shows the importance of specialized corpora, and this information is essential when preparing courses for students of languages for specific purposes. Table 6 sets out the distribution.

Table 4. Global frequencies of the verb *deduzir* in the analysed corpora

CORPORA*	SPK.	JOURN.	LIT.	DID.	ECON.	JUR.
Frequency	6	9	7	57	40	242

*SPK.: spoken; JOURN.: journals; LIT.: literary; DID.: didactics; ECON.: economics; JUR. law.

Table 5. Distribution of forms of the verb *deduzir* in the analysed corpora

Forms of <i>deduzir</i>	Distribution per corpus*					
	SPK.	JOURN.	LIT.	DID.	ECON.	JUR.
deduz		2		8		9
deduz-se	1	2	1	3		1
deduza						1
deduzem				2		1
deduzi		1	1			
deduzia			1			
deduziam	1					
deduzida		2		6	2	48
deduzidas				1	3	11
deduzido	1		1	1	6	56
deduzidos					8	
deduzimos				4		12
deduzindo				2	5	6
deduzir	1	2	1	21	13	35
deduzir-se					1	1
deduzirá						1
deduziram						9
deduzirem	1					
deduziria				1		
deduzirmos				2		
deduziste				1		
deduziu			1	5	1	50
deduziu-se						1
deduzo	1		1			
Totals	6	9	7	57	40	242

*SPK.: spoken; JOURN.: journals; LIT.: literary; DID.: didactics; ECON.: economics; JUR.: law.

Table 6. The verb *deduzir* in 6 corpora of $\pm 1\text{M}$ words each

Deduzir \cong to deduce: syntactic structure and frequency of occurrence in each corpus

Syntactic structure	SPK.	JOURN.	LIT.	DID.	ECON.	JUR.
$V \begin{Bmatrix} N \\ F \\ \text{Que } F \end{Bmatrix}$ (de, por, a partir de,... N)	4	7	5	5	10	20

Examples:

as carências são muitas e	deduz-se	das suas palavras que também as (SPK)
Nestas circunstâncias, é legítimo	deduzir	que não se trata de um projec (JOURN)
Querem evitar a tropa,	deduziu	ele (LIT)
As conclusões aí	deduzidas	não são aplicáveis imediatamente (DID)
estabilização dos depósitos a prazo leva a	deduzir	uma progressão mais moderada (ECON)
demais circunstâncias do caso concreto,	deduz-se,	com suficiente clareza, que a ind (JUR)

Table 7. Examples of verb form distribution

Verb	Frequency		Spoken Verbal tense	Form	Written		Form	
	Spoken	Written			Verbal tense	Form		
Arreigar-se	7	28	Past participle	arreigado	3	Past participle	arreigada	11
				arreigados	2		arreigado	6
				arreigado	2		Arreigado	6
							arreigadas	2
						Infinitive	arreigar-s	1
						Pret.	arreigava-se	1
						Imperfeito		
Graduar	3	75	Past participle	graduado	2	Past participle	graduado	28
				graduados	1		graduados	21
						graduada		18
						Infinitive	graduar	3
						Simple	graduou	2
						past		
Sofisticar	3	203	Past participle	sofisticada	1	Past participle	sofisticados	60
				sofisticado	1		sofisticado	56
				sofisticados	1		sofisticada	51
							sofisticadas	35
						Infinitive	sofisticar	1

sofisticar in the Portuguese corpus (Table 7), we can see that they are used only in the past participle in spoken discourse and that the written discourse has also a strong tendency to use them mainly in the past participle. Thus the corpus shows morphological gaps about which we get no information from grammars or dictionaries; both of these sources repeat, usually, the same traditional lists of defective forms and they do not make reference to emergent uses (Bacelar do Nascimento 2000a).

Concordances and collocations can also be very useful in shedding new light on the shades of meaning given by near-synonyms. The adjectives *célebre*, *famoso* and *notável* are presented in Portuguese dictionaries as synonyms. But, the concordances and collocates, extracted from a 12 M word corpus and organized according to the Mutual Information (MI) (Church & Hanks 1990), revealed very distinct lexical patterns.

Table 8. Associative patterns of three “near-synonyms”: *Célebre* – *Famoso* – *Notável*: collocates in decreasing order of Mutual Information (MI)

*** FT 454 <i>CÉLEBRE</i> ***			*** FT 686 <i>FAMOSO</i> ***			*** FT 433 <i>NOTÁVEL</i> ***		
Collocates	MI*	Pair**	Collocates	MI*	Pair**	Collocates	MI*	Pair**
tristemente	8.825	10	tornar	7.043	14	conjunto	6.641	6
criminoso	8.320	4	nome	6.876	11	qualidade	6.213	6
ficar	6.203	14	colecção	6.429	4	verdadeiramente	6.184	6
frase	5.891	7	ficar	5.311	9	esforço	5.575	8
tornar	5.846	12	americano	5.253	4	obra	5.044	7
autor	5.034	6	gente	4.262	5	época	4.840	6
mais	3.695	37	sua	4.259	18	exemplo	3.889	5
tão	3.561	10	de	4.068	433	mais	3.849	48
ser	3.394	45	ser	3.953	67	ser	3.836	68
de	3.259	325	mais	3.720	77	fazer	3.792	7
sua	3.186	11	menos	3.630	4	muito	3.769	16
em	3.025	73	grupo	3.513	4	trabalho	3.622	7
seus	2.973	6	por	3.501	55	ter	3.580	8
por	2.901	31	tão	3.467	8	de	3.393	254
muito	2.531	10	seu	3.395	22	com	2.480	29
dia	2.449	4	casa	3.236	4	e	2.388	77
já	2.046	6	já	3.087	23	em	2.092	56
que	1.847	42	em	3.079	64	para	1.902	25
e	1.816	46	muito	2.660	9			
com	1.639	14	como	2.489	16			
			a	2.450	253			
			e	2.441	119			
			com	2.122	30			
			não	1.990	36			

*MI: Mutual Information

**Pair: frequency of occurrence of the pair

In Table 8, frequent collocates are shown in decreasing order of MI. Table 9 gives the noun collocates for each of the adjectives. Table 10 concerns adverb collocates (Bacelar do Nascimento 2000b).

It can be seen that the lexical co-occurrences are quite different. On the one hand, each adjective shows that it is used with relatively specific nouns and adverbs. On the other hand, the information provided shows that those nouns and adverbs are not inter-exchangeable.

When considering the noun collocates, it can be seen that *frase célebre* is used but not *frase notável*, *notável qualidade* but not *famosa qualidade* and *nome famoso* but not *nome célebre*. Relatively to the collocate adverbs, *célebre* collocates with *tristemente*, which has a semantical negative weight, and *notável* collocates with the more emphatic *verdadeiramente*.

Table 9. Associative patterns of three “near-synonyms”: *Célebre* – *Famoso* – *Notável*: noun collocates

<i>célebre</i> nouns	<i>famoso</i>	<i>notável</i>
criminoso	nome	conjunto
frase	coleção	qualidade
autor	americano	esforço
dia	gente	obra
	grupo	época
	casa	exemplo
		trabalho

Table 10. Associative patterns of three “near-synonyms”: *Célebre* – *Famoso* – *Notável*: adverb collocates

<i>célebre</i> adverbs	<i>famoso</i>	<i>notável</i>
tristemente	mais	verdadeiramente
muito	menos	mais
mais	tão	muito
tão	já	
já	muito	
	não	

Conclusion

From the evidence presented above it can be seen that the analysis of natural materials is of relevance and importance to language teachers and students. The possibility of exploring the language that they are studying, either as mother tongue, or second or foreign language, can be very useful and motivating for the learner. The use of concordancing, and, in particular, collocation analysis provides teachers with very good data for study and for the design of materials for classroom work, either for preparing presentations or in the organising of exercises. Similarly, for students it offers a new approach to understanding several different aspects of the language they are studying, such as meaning and semantic disambiguation, real differences between near-synonyms, morphosyntactic aspects, and information on the uses of words in natural contexts and terms in specialized discourse.

References

- Baayen, R. H. (1998). Lexis, word frequencies and text types. In *IV-V Jornades de Corpus Lingüistics 1996–1997* (pp. 87–102). Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
- Bacelar do Nascimento, M. F. (2000a). Corpus de référence du portugais contemporain. In M. Bilger (Ed.), *Corpus. Méthodologie et applications linguistiques* (pp. 25–29). Paris: Honoré Champion.
- Bacelar do Nascimento, M. F. (2000b). Exemples de combinaisons établies pour l'écrit et pour l'oral à Lisbonne. In M. Bilger (Ed.), *Corpus. Méthodologie et applications linguistiques* (pp. 237–261). Paris: Honoré Champion.
- Bacelar do Nascimento, M. F. (2003). O papel dos corpora especializados na criação de bases terminológicas. In I. Castro & I. Duarte (Eds.), *Razões e Emoção. Miscelânea de estudos em Homenagem a Maria Helena Mira Mateus Volz* (pp. 167–179). Lisboa: Imprensa Nacional-Casa da Moeda.
- Bacelar do Nascimento, M. F. (2001a). Fenómenos de Lexicalização no Português Contemporâneo, AATSP, San Francisco, U.S.A. (unpublished paper).
- Bacelar do Nascimento, M. F. (2001b). Para um banco de dados do português falado e escrito. Corpora linguísticos: desenvolvimentos e aplicações. In *1st International Meeting of AILP – Internacional Association of Portuguese Linguistics*, Lisbon, Faculdade de Letras da Universidade de Lisboa, October 8th, (unpublished paper).
- Church & Hanks (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.
- Gonçalves, J. B. & Veloso, R. (2000). Spoken Portuguese: Geographic and Social Varieties. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, & G. Stainhaouer (Eds.), *LREC2000 Second International Conference on Language Resources Evaluation Volume II* (pp. 905–908). Athens, Greece: ELRA.
- Sinclair, J. McH. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Appendix 1.

CENTRO DE LINGÜÍSTICA UNIVERSIDADE DE LISBOA (CIUL)
REFERENCE CORPUS OF CONTEMPORARY PORTUGUESE (CRPC)

201 487 845 words (April, 2002)

Text type:

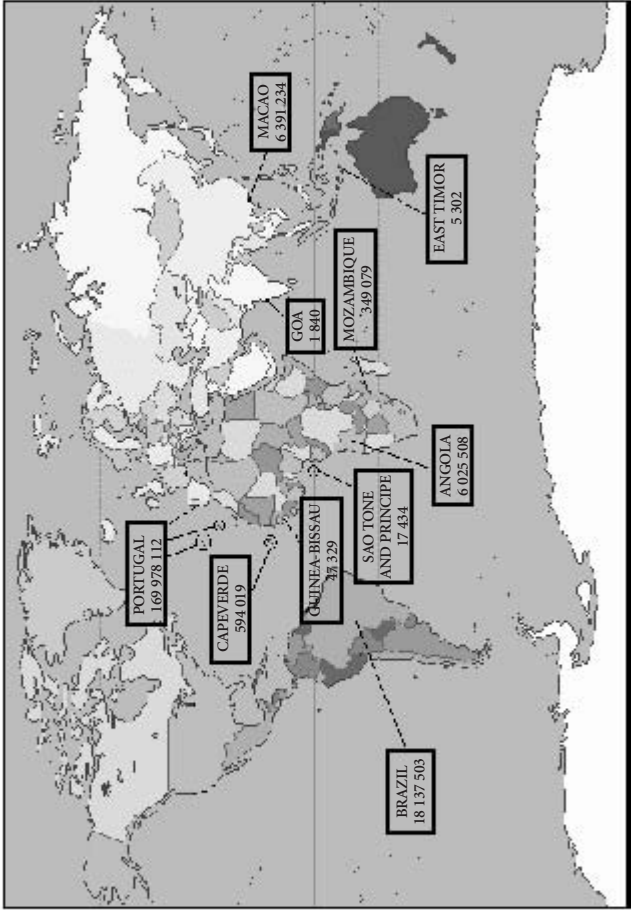
WRITTEN
198 991 203
SPOKEN
2 496 642

Timespan:

BEFORE 1900
1 092 087
1901–1970
2 772 576
AFTER 1970
197 623 182

Written
sources:

NEWSPAPERS
142 420 838
BOOKS
23 512 641
TEC./SCIENTIF.:
11 882 077
LIT.: 9 662 421
DID.: 1 968 143
DOCUMENTS
20 155 210
PARLIAMENT:
16 861 242
COURT: 3 293 968
PERIODICALS
8 491 083
MAGAZINES:
5 990 015
JOURNALS:
2 501 068
VARIA
3 904 756
LEAFLETS
343 483
LETTERS
163 192



Appendix 2.

Available products, issued from CLUL or in partnership, based on the *Corpus de Referência do Português Contemporâneo* (CRPC)

Published transcriptions (P.F.) of Portuguese spoken Corpus	CLUL	http://www.clul.ul.pt
<i>Léxico Multifuncional Computorizado do Português Contemporâneo</i>	partnership	http://www.clul.ul.pt
<i>Português Falado</i> – Records with transcription alignment	partnership	4 CD-ROM
3 million words of the PAROLE corpus, 250,000 of which are morphosyntactically tagged with human desambiguation	partnership	ELRA catalogue – http://www.elda.fr
PAROLE Lexicon with 20,000 entries morphosyntactically tagged and syntactically described	partnership	ELRA catalogue – http://www.elda.fr
3,000 units from PAROLE Lexicon semantically characterised (SIMPLE program) – multilingual	partnership	http://www.ub.es/gilcub/SIMPLE/simple.html
“Corpus compartilhado VARPORT”	partnership	
“Recursos linguísticos para o português”	partnership	http://www.clul.ul.pt
Corpus REDIP	partnership	
C-ORAL-ROM 4 romanic languages comparable spoken Corpus	partnership	

Appendix 3.

Concordance of *deduzir* from the juridical corpus

ntes não tomaram qualquer atitude, pois <i>deduziram</i>	acusação autónoma do Ministério Públi
XXX, esposa do arguido, que havia <i>deduzido</i>	acusação contra XXX pelo
Ministério Público na comarca de Lisboa <i>deduziu</i>	acusação contra os arguidos XXX
vilhã, o Magistrado do Ministério Público <i>deduziu</i>	acusação contra XXX,
em processo comum, o Ministério Público <i>deduziu</i>	acusação contra os arguidos XXX
Ulado por XXX. 4) Foi contra ele <i>deduzida</i>	acusação em processo comum e com in
ferecendo provas, requerendo diligências, <i>deduzir</i>	acusação independente da do Ministéri
de intervenção, pois podem os assistentes <i>deduzir</i>	acusação independentemente da do Mii
n. 3 do Código de Processo Penal e não <i>deduziram</i>	acusação, nem pedido civil. Efectuado
filhado pelo Ministério Público desde que <i>deduziu</i>	acusação. O crime de homicídio priileg
o desta acusação os assistentes podiam ter <i>deduzido</i>	acusação pelos factos acusados pelo M
o do processo acima referido e no qual foi <i>deduzida</i>	acusação por homicídio negligente cont

utilização pública, e ter-se desinteressado de *deduzir*
 o corria seus termos, só depois tendo sido *deduzida*
 riação e XXX *deduziu*
 do se tenha em atenção que, podendo ser *deduzidos*
 XXX, veio *deduzir*
 , n. 2, do Código de Processo Civil, tendo *deduzido*
 citação, é perfeitamente legítimo que ele *deduza*
 4. Perante estes factos, o recorrente veio *deduzir*
 XXX, *deduzir*
 o seu direito. Em nenhuma delas poderão *deduzir*
 move a XXX *deduzir*
 e XXX *deduziram*
 Rca de Guimarães, XXX *deduziu*
 sede na Avenida 5 de Outubro, em Lisboa, *deduziu*
 ilelo, Lda. ” contra XXX, este *deduziu*
 XXX, casada com o XXX, *deduziu*
 XXX tivessem *deduzido*
 filha menor XXX, *deduziu*
 XXX, também com os sinais dos autos, *deduziu*
 do Ministério Público, e podiam ainda ter *deduzido*
 XXX e a Autora, contra quem foi *deduzido*
 a XXX e que nele foi *deduzido*
 ulo, os herdeiros da falecida XXX vieram *deduzir*
 avado -, todos do Código Penal de 1995). *Deduziram*
 m à acusação do Ministério Público, nem *deduziram*
 idos, se constituíram assistentes. E ambos *deduziram*
 se constituíram assistentes nos autos, nem *deduziram*
 lha 190) , perfilhou a acusação pública e *deduziu*
 Código Penal. O ofendido XXX *deduziu*
 “Dado que XXX *deduziu*
 em que disse aderir à acusação pública e *deduziu*
 a na sua petição inicial. O R. contestou e *deduziu*
 ntestou por excepção e por impugnação e *deduziu*
 elada no acórdão recorrido, de terem sido *deduzidos*
 Código Penal. Contra ele foram também *deduzidos*
 mo o poder de dispor da relação jurídica *deduzida*

acusação própria. Finalmente, o direito
 acusação pública. Com a notificação de
 embargos a execução movida pelo XX
 embargos à execução em que a matéria
 embargos à acção executiva que a esta
 embargos com fundamento em que a exe
 embargos de terceiro para discutir a natu
 embargos de terceiro para defender a sua
 embargos de terceiro pedindo o levanta
 embargos de terceiro (artigo 1037 do Có
 embargos de executado, alegando, por u
 embargos de terceiro contra XXX
 embargos de executado por apenso à exe
 embargos de terceiro na execução para e
 embargos de executado, invocando, no e
 embargos de terceiro com fundamento e
 Pedido cível, foram os arguidos também
 Pedido cível contra ambos os arguidos, p
 Pedido cível contra aquele arguido pedin
 Pedido *civil* de indemnização, o qual por
 pedido de indemnização cível, enquanto
 pedido de indemnização cível pelos herd
 pedido de indemnização cível nos termo
 pedido de indemnização *civil* contra o ar
 pedido de indemnização *civil*, pelo que n
 pedido de indemnização *civil*: XXX,
 pedido de indemnização *civil*. 23. Tal co
 pedido de indemnização *civil* contra o XX
 pedido de indemnização *civil* contra o arg
 pedido de indemnização *civil*, está por ess
 pedido de indemnização cível. Mas o Mer
 pedido reconvenicional, replicando o A. n
 pedido reconvenicional, pedindo a conden
 Pedidos cíveis entretanto decididos ou de
 Pedidos de indemnização *civil* pelos famil
 em juízo cabe, em geral, aos respectivos s

Note: Proper nouns were replaced by XXX.

Learner corpora

Learner corpora and their potential for language teaching

Nadja Nesselhauf*

University of Basel

Learner corpora are a fairly recent phenomenon, which has not been systematically explored yet. This paper is intended as a survey of the field, with special attention to the potential of learner corpora in the area of language teaching. In the first part, a definition of learner corpora and a survey of the state of the art in the field are attempted, the general potential and limitations of learner corpora pointed out, and fruitful ways forward in the compilation of such corpora discussed. In the second part, the relation of learner corpora and language teaching is investigated. The so far only slight impact of results from learner corpus studies on pedagogic material is outlined and possible further impact discussed. The potential use of learner corpora in a data-driven learning approach is also explored.

1. Introduction

Until very recently, discussions on the role of corpora in language teaching usually focused on native speaker corpora (e.g. Leech 1997; McEnery & Wilson 1997). Learner corpora, i.e. systematic computerized collections of texts produced by language learners, were at most considered to play a peripheral role. That they have the potential for far more than merely a peripheral role in language teaching is, however, starting to be recognized (e.g. Aston 2000). In this paper, this potential will be investigated in more detail.

Hardly anyone will doubt any longer that native speaker corpora are indeed useful for the improvement of language teaching. They are useful mainly because they can reveal – better than native speaker intuition – what native speakers of the language in question typically write or say (either in general or in a certain situation / in a certain text type). For language teaching, however, it is not only essential to know what native speakers typically say, but also what

the typical difficulties of the learners of a certain language, or rather of certain groups of learners of this language, are. Since, as has been amply demonstrated by critics of the Contrastive Analysis Hypothesis, it is not sufficient to compare the learners' L1 with the target language to identify areas of difficulty, the best way to find out what these difficulties are is to analyse the language produced by a certain group of learners and compare it with the language produced by native speakers.

For both learner corpora and native speaker corpora, two types of application to language teaching can be distinguished (cf. Figure 1).

The most common application of corpora to language teaching so far – sometimes called the ‘COBUILD’ or the ‘Birmingham’ approach – has been to apply results from native speaker corpus analyses to the improvement of pedagogic material,¹ especially reference material, by making it correspond more closely to typical native speaker use. The second major type of application of native speaker corpora to language teaching, often referred to as ‘data-driven learning’, is to use corpora more directly in the classroom, by having students either analyse the corpus itself or examples from the corpus prepared by the teacher. (The broken line between data-driven learning and pedagogic material in Figure 1 indicates that these two approaches are not entirely separable, as pedagogic material can also incorporate corpus examples for the students to analyse.)

Both these approaches can be used for learner corpora as well, the difference being that learner corpora are often best used in combination with native speaker corpora. Learner corpora can be applied to pedagogic material in two different ways. The more direct, and probably more important way is to use a learner corpus to identify what is particularly difficult for a certain group of learners and to put special emphasis on these points in the different materials.² The second, more indirect and more problematic way is to derive insights about second language acquisition (for example about developmental

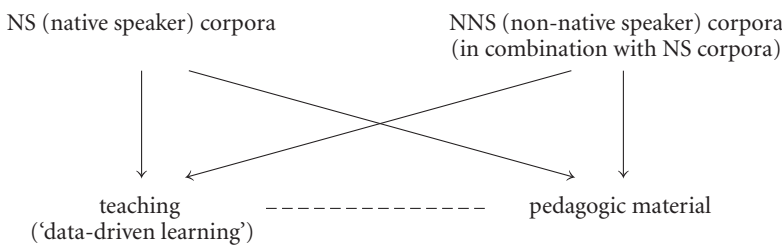


Figure 1. Connections between corpora and language teaching

sequences) from learner corpus analyses and to draw implications for teaching from these insights. So far, only the first of these ways is starting to be used. The third possible application of learner corpora to language teaching, namely using learner corpora or data from learner corpora directly in the classroom, has also only rarely been attempted so far. But although learner corpora have to be used more prudently than native speaker corpora for data-driven learning, the former also have some potential in the classroom, as will be shown later.

The above figure is, of course, not meant to imply that the connections indicated are the only connections between corpora and language teaching. Other connections exist and other types of corpora, such as parallel or comparable bilingual corpora or textbook corpora, also have a role to play in the improvement of materials and/or in the classroom; these connections will however not be considered further in this paper.

Learner corpora are a fairly recent phenomenon, which has not yet been systematically described and which offers great scope for improvement itself. Therefore, in the first part of this paper, I will look at the state of the art and the most promising ways forward in the field of learner corpora as well as at their general potential and limitations. In the second part of the paper, I will look at the current role of learner corpora in language teaching, at their so far largely unused potential in this field, and at ways in which this potential can best be used.

2. Learner corpora

2.1 Learner corpora – the state of the art

Learner corpora were provisionally defined above as systematic computerized collections of texts produced by learners. In order to discuss the state of the art in the field, some more precision and explanation is required. A criterion that should be added to the definition above is that the text collection should not be intended merely for use in one particular study (or a limited number of studies) but for more general use. ‘Systematic’ is taken to mean that the texts included in the corpus were selected on the basis of a number of – mostly external – criteria (e.g. learner level(s), the learners’ L1(s)) and that the selection is representative and balanced. As opposed to definitions of native speaker corpora, ‘texts’ is not modified by ‘naturally occurring’ in the above definition. The reason for this is that in a foreign language environment it is extremely difficult

to find naturally occurring texts in its strict sense. A scale of naturalness in text production might look as follows:

fully natural – product of teaching process – controlled tasks – scripted

In a foreign language environment, what comes closest to naturally occurring texts in its strict sense are texts that are produced for pedagogical reasons and texts that are elicited for the corpus but that use procedures exerting very little control. Typically, therefore, either free compositions produced for a certain course or free compositions or oral interviews produced for the corpus are used for learner corpora. Collections of types of data that have been elicited with procedures exerting more control on the texts produced, such as compositions guided by pictures or student translations, are usually not considered learner corpora. Since the distinction between more and less controlled is, naturally, not clear-cut, such collections might be considered peripheral types of learner corpora. The problems of defining the term 'learner' additionally contribute to making the definition of learner corpora somewhat fuzzy. Typically, 'learner' is used to refer either to somebody learning a foreign language or to a foreigner learning the language in a country where it is spoken natively; somewhat less typically 'learner' is used to refer to students living and learning the language in a country where it is not spoken natively but used officially as a second language (and/or as language of education and/or administration, cf. Granger 1996a). The term is only very rarely used for adult speakers in such countries. However, in countries in which the status of the language in question is somewhere between foreign and second language (for example English in Hong Kong), 'learner' is commonly used for students and also sometimes for adult speakers. The distinction, therefore, between more typical 'learner' corpora and other types of non-native speaker corpora (such as, for example, the East African subcorpus of the International Corpus of English) is not clear-cut.³ Like the concept of native speaker corpora (McEnery & Wilson 2001: 29ff.), the concept of learner corpora is thus prototypical, i.e. there are more and less typical learner corpora. In the rest of this paper, the focus will be on the more typical learner corpora.

Learner corpora, in this more typical sense, have only been in existence for a relatively short time: with very few exceptions, the compilation of learner corpora did not begin until the 1990s.⁴ The idea as such, i.e. to collect learner language, is not new, however. Especially in the late 1960s and 1970s, when error analysis was en vogue, many collections of learner language were created. But in contrast to learner corpora, these collections were usually used only as a repository for errors and not exploited as corpora, so that the text collec-

tion itself was usually discarded after the errors had been extracted. Besides, these collections were not computerized and rarely systematic (cf. Granger 1998b: 5f.).

Despite the fact that learner corpus compilation is a fairly new activity, quite a number of learner corpora already exist or have at least been started. Almost all learner corpora that have been compiled are corpora of learner English, usually consisting of written language and of data from learners of one L1-group. Probably the biggest learner corpus to date is the Hong Kong University of Science and Technology (HKUST) Learner Corpus, which currently contains about 25 million words and is still growing. It consists of different academic text types in English produced by Chinese undergraduate students. Examples of other big written corpora of learner English of one L1 group are the TeleNex Student Corpus, which currently contains about 3 million words of compositions written by secondary school students from Hong Kong, the Chinese Learner English Corpus (CLEC), currently containing 1.2 million words of compositions of secondary school and university students, and the Uppsala Student English Project (USE) with about 1 million words of different types of essays written by Swedish undergraduate students of English. One of the few existing learner corpora consisting of specialized but non-academic language to date is the Learner Business Letters Corpus (BLC), which contains about 200,000 words of business letters written by L1 Japanese business people. A further, much smaller group of corpora contains texts written by learners with different language backgrounds. Two of these corpora are what could be called commercial corpora as they have been compiled and are exploited largely for commercial purposes. These are the Longman Learner Corpus (LLC), which contains about 10 million words from learners of about 160 different language backgrounds, and the Cambridge University Press (CUP) Learner Corpus, also containing about 10 million words. Probably the only existing sizeable non-commercial learner corpus containing data from learners with different L1s is the International Corpus of Learner English (ICLE), currently consisting of around 2 million words and comprising argumentative essays written by university students of English with 14 different L1 backgrounds.

For spoken learner corpora, only a few projects exist so far. Currently the biggest projects are the Standard Speaking Test (SST) Corpus, which aims at collecting 1 million words of interviews with Japanese learners of English, and the Louvain International Database of Spoken English Interlanguage (LINDSEI), which contains interviews with advanced learners of different L1s and also aims at 1 million words. Corpora, both written and spoken, for languages other than English are also rare and so far usually quite small, one of the

biggest being the French Interlanguage Database (FRIDA), currently at 200,000 words. A list of the bigger learner corpora that currently exist can be found in the appendix, together with references to sources of further information about them.

The criteria that can be and have been used for the compilation of learner corpora partly coincide with the criteria used for native speaker corpus compilation and partly apply specifically to learner corpora. The most important of these criteria probably are – like for native speaker corpora – language, medium and text type(s), and – specific to learner corpora – level(s) of learners, L1(s) of learners, type of language acquisition (i.e. instructed and/or naturalistic), and task setting (in the case of free compositions for example timed/untimed writing, use of reference tools allowed or not). At the moment, in addition to the fact that most corpora consist of written learner English and contain data from learners with the same L1, the level of learners in learner corpora is mostly advanced or intermediate, the text types included are mainly general essays or academic text types (such as research papers), the learners are mostly foreign learners, and the first languages of the learners are predominantly European or Asian languages. Since most of the learner corpus projects have been started only recently, most of the corpora are not complete yet. In a number of projects (e.g. HKUST and ICLE) native speaker control corpora have been or are being compiled in addition to the learner corpora. In many projects, there are also plans to tag the data, usually for parts-of-speech and errors, but so far, tagging for most corpora is either incomplete or has not been started at all.

2.2 Potential and limitations of learner corpora

One of the greatest advantages of learner corpora in comparison to other types of learner language collections is that the texts are computerized. Although it is not usually advisable to perform fully automatic analyses on learner corpora (because of the great number of misspellings and especially other kinds of mistakes, which even careful tagging often cannot capture), the computer can still be of great assistance with otherwise very laborious tasks. Even more important perhaps is the fact that computerized data can be distributed more widely, so that results are more easily comparable and also more easily verifiable than if each researcher (or each small group of researchers) uses a different set of data for their analyses. The latter has been and still largely is the practice in the fields of second language acquisition and language teaching, so that it is often unclear to which of the many differing factors differences between results can be attributed. Another advantage is that with learner corpora, real production

data (or at least data coming close to it) is analysed, while so far many investigations into learner language have been based on more experimental data (such as multiple choice tasks or grammaticality judgement tasks). Whereas experimental data can be of use if the learners' more abstract knowledge of a language feature is investigated, for many purposes (including the improvement of language teaching) it is of greater interest to find out what learners can produce spontaneously. Owing to the great gap between the abstract knowledge and the actual performance of language learners, drawing conclusions about what a learner can produce spontaneously is difficult on the basis of experimental data.

Furthermore, learner corpora make more comprehensive studies possible. While experimental data allows investigations into only a few specific aspects of learner language at a time, with learner corpora many aspects can be investigated at once, and more general questions such as the relative frequency of different types of mistakes can be addressed. In addition, it is not necessary to approach corpus data with a hypothesis, so that new aspects of learner language can be discovered. Aspects of pragmatics and discourse, including communication strategies, can also be studied more easily with production data.

Owing to the fact that a learner corpus by definition is systematic, i.e. it has been compiled on the basis of a number of criteria, the influence of various factors on learner language can also be analysed – provided that the corpus or the corpora being used are designed accordingly. Any aspect of learner language can then be investigated with respect to the learners' proficiency level, their L1, the medium, text type, the learning environment in which the language was acquired (i.e. naturalistic and/or instructed), the age and sex of the learners, the years of acquisition, the influence of L3s and any other information that the corpus provides. With a comparable native speaker corpus, over- and underuse (i.e. which features learners use uncommonly often and uncommonly rarely compared to native speakers) can be studied in addition to mistakes and correct forms. With a comparable L1 corpus, the extent to which the learners' difficulties (and non-difficulties) are dependent on their L1 can be investigated.

In spite of the great potential of learner corpora, they naturally also have some limitations. The receptive abilities of learners cannot be investigated, for example, and questions such as how certain the learners are about the acceptability of what they are producing cannot be answered. In addition, if a word or a pattern does not occur in a text, there is no way of finding out whether the learner knows it or not, so that especially rare phenomena might be better studied experimentally. A more detailed investigation into the role of certain learner characteristics, such as language aptitude or motivation, is also difficult with

a learner corpus (though not impossible in principle). Similarly, an analysis of the role of input and interaction or of certain teaching methods on acquisition, might be better and more easily carried out with a more experimental procedure. In general, the best method of analysis depends on what is to be analysed, and in many cases (like for the analysis of native speaker language, cf. e.g. Fillmore 1992; de Mönnink 1999), the best method for learner language analysis is likely to be a combination of corpus analysis and an experimental approach.

2.3 Ways forward

Whether the full potential of learner corpora can be used, crucially depends on the availability of well-designed corpora. But despite the fact that a number of such corpora already exist, there is still great scope for further corpora and for improvement of the existing ones. If the list of existing learner corpora is compared with the list of important criteria (cf. Section 2.1), it immediately becomes obvious that, although the field of learner corpora is growing quickly, the corpora that have been compiled so far are only a beginning. Even if only English – the language for which most learner corpora have been compiled – is considered, many gaps are apparent. For many L1s, there are no corpora yet. If there are, they are often restricted to one medium, to one or only a few text types, and to one level of proficiency. What would be desirable, therefore, besides corpora for more L1s, is corpora containing data from learners of different proficiency levels and corpora of different media and text types. A step in the right direction with respect to proficiency levels is CLEC, for instance, which contains a substantial amount of learner language from each of five different proficiency levels.⁵ With respect to creating spoken learner corpora, LINDSEI indicates a way forward because it is designed so as to be easily comparable with an already existing written corpus (the language it records is of learners of the same level and the same L1s as in ICLE). As to text types, every text type, or more generally, every type of language that certain groups of learners need to produce would be useful in a learner corpus, from informal conversation to highly specialized language. So far, with the exception of academic language, more specialized language (such as business English or the language of specific subject areas) has been almost entirely neglected in learner corpora, although learning and teaching LSP is gaining more and more importance and could certainly benefit from access to learner data.

Having said all that, it has to be added that corpus compilation, not to mention corpus annotation, is of course a laborious task: “the compilation of a corpus (with proper attention to quality, design criteria and so on) always takes

twice as long as one thought, and sometimes ten times as much effort" (Leech 1998: xvii). Since this is even more true for learner than for native speaker corpora, individual efforts will not suffice if real progress is to be made in the field. Progress will crucially depend on collaboration and data-sharing.

The prevalent tendency in learner language analysis (and not only in error analysis) has been to collect learner writing, sometimes substantial amounts, to use it for one study or possibly a few, and to then discard it (as in Kroll 1990; Gitsaki 1999 and many others). This usually means a great effort on the part of the researcher – but an effort that is largely wasted, because the potential of the text collection is only used to a very limited degree. Today's learner corpora mark a step forward since they are used for a number, usually even a large number, of studies. Nevertheless, the prevalent tendency with respect to learner corpora at the moment seems to be that one researcher, or one university department, compiles a corpus. Although such efforts have produced a number of good corpora, corpora with even more potential could be compiled with more co-operation. Through co-operation, bigger and better corpora can be designed, i.e. corpora that allow investigation into a wider variety of phenomena as well as into more than one parameter at a time. Through international co-operation, corpora including different L1s (such as ICLE) can be compiled. If countries in which the language in question is spoken natively co-operate, native speaker control corpora can be compiled as well as corpora including foreign and second language learners of the same L1, which would allow investigation into differences between instructed and naturalistic language acquisition. Co-operation does not necessarily mean setting up new projects – it can and should also mean extending already existing corpora, thus using already existing resources even better. To give just one example, if USE were extended with other L1s, a corpus of considerable size with different L1s would be the result.⁶

Finally, even well-designed corpora are of not much use if they are not widely available for research. Whereas most compilers naturally want to use their own data first and many do not want to give away for free what cost them a lot of time, money and effort, there does not seem any reason why corpora should not be made available after some time and/or for a fee (by publishing a CD-ROM, creating access over the Internet, or via one of the existing data-distribution organisations⁷). At the moment, however, wide availability of learner corpora is still a rarity, and one can only hope that many compilers follow the example set by those of ICLE, the BLC, the JPU Corpus (a Hungarian EFL corpus), and the Corpus of English by Japanese Learners (cf. appendix), which are among the few corpora that have been or soon will be made available.

3. Learner corpora and language teaching

3.1 Learner corpus studies

Of the two approaches to using learner corpora for language teaching outlined in the introductory section, the one that holds most potential is the use of learner corpora to improve pedagogic material. The basis for realizing this potential is the analysis of learner corpora. This section will look at the question to what extent the learner corpus studies that have been carried out so far constitute such a basis.

Studies analysing learner corpus data are now rapidly increasing in number. The majority of learner corpus studies published so far have been carried out on the basis of ICLE subcorpora and therefore look at advanced learner argumentative writing. Almost all the major areas of language structure have been studied to some degree: syntax (e.g. complement clauses: Biber & Reppen 1998; tenses: Granger 1999), lexis (e.g. high-frequency verbs: Ringbom 1998a), phraseology (e.g. recurrent word combinations: Milton & Freeman 1996; formulae: De Cock 1998), and discourse (e.g. connectors: Altenberg & Tapper 1998). Morphology and orthography have received practically no attention so far, and phonology cannot be analysed with a typical learner corpus.⁸ A few more comprehensive studies simultaneously investigating different areas of language have also been carried out (e.g. the relation between different error types: Dagneaux et al. 1998). Only a few of the studies have been primarily concerned with questions of second language acquisition (e.g. Tono 1998). Most of the other studies either claim or at least imply that they attempt to make a contribution to language teaching.

Some more general results with implications for teaching recur in the analyses that have been carried out so far. First, in a number of studies, the learners' first language has been found to have an even greater influence on several aspects of learner language than has commonly been assumed. Investigating advanced French-speaking learners of English, for example, Granger (1999: 198) found that a large number of tense errors are transfer-related. The same type of learner was also found to overuse amplifiers that have a direct and frequently used equivalent in their L1 (such as *completely* or *totally*, Granger 1998d: 148ff.). Similarly, investigating Chinese-speaking learners of English, Man-Lai (1994: 163) found that the use of delexical verbs is often modelled on L1. A second recurrent result has been that learner writing is often less qualified than native speaker writing. Flowerdew (2000: 150ff.), for example, reports an underuse of hedging devices, such as certain verbs (e.g. *seem*, *sug-*

gest, indicate) and phrases (e.g. *it would appear*). In a similar vein, Milton & Hyland (1999: 152ff.) report that learners use certainty markers (such as *obviously, certainly*) much more frequently than native speaker students. A third observation that recurs in a number of studies is that written learner language is more speech-like in many respects than comparable written native-speaker language.⁹ Altenberg & Tapper (1998: 86ff.), for example, found that more formal conjuncts (such as *therefore, thus, however* and *yet*) were underused by learners, whereas more informal ones (such as *but* or *still*) were overused. Granger & Rayson (1998) demonstrated that underuse of more formal and overuse of more informal words is a phenomenon concerning almost all word classes. They observed, for example, that more general nouns such as *people* and *things* were overused as well as more informal adverbs such as *also, only, very* and *so*, whereas more formal *-ly*-adverbs such as *importantly* or *ultimately* were underused. An underuse of more formal constructions in learner language was also repeatedly reported, for example for passives (Granger 1997a) and participle clauses (Granger 1997b). In addition to these more general insights, many of the studies that have been carried out so far have produced results on more specific aspects of learner language, which are of interest for language teaching. To give just one example, Lorenz, in his study of adjective intensification in learner language, found that one of the reasons why learner adjective intensification often sounds non-native-like is that learners often intensify adjectives in thematic position, while native speakers mostly intensify adjectives in rhematic position (1998: 61ff.). The unnaturalness of the intensification in *I thought that my **absolutely authentic** Rock music should hit the charts in seconds*, for instance, is largely due to its occurrence in thematic position (1998: 62).

Despite a number of useful results, learner corpus analysis has clearly only just begun, and there is still much room for improvement. Firstly, at the moment there is a wide variety of disconnected studies, usually concentrating on a few single words or uses of words, and there are hardly any studies that look at a phenomenon in more depth (one of the few exceptions so far is Lorenz's work on adjective intensification mentioned above; cf. Lorenz 1998, 1999). This is largely, but not exclusively, due to the newness of the field. Learner corpus research should now start to bring forth more comprehensive studies and also to better co-ordinate its efforts (ideally on the basis of more co-ordinated learner corpus compilation as envisaged in the previous section). Secondly, many, if not the majority, of learner corpus studies so far have concentrated on phenomena that can easily be studied automatically. Almost all studies look either at certain individual words, at continuous word sequences, or at other fea-

tures that can be easily extracted from the corpus (such as direct questions). In addition, many studies are exclusively or primarily quantitative. Particularly popular objects of analysis are tag sequences (e.g. de Haan 1997, 1998; Aarts & Granger 1998; Tono 2000) and word frequencies (e.g. de Haan 1997; Kaszubski 2000; Ringbom 1998a, 1998b). While such studies can be interesting starting points for further quantitative analyses, they do not usually in themselves contribute much to language learner analysis, let alone to language teaching. If progress is to be made, it is imperative that this current stage is left behind and that more qualitative analyses are carried out. A precondition for this is that it is realized more widely that learner corpus analysis necessarily involves a greater amount of manual work than native speaker corpus analysis. It seems misguided to assume that the main or even the only point of having computerized learner corpora is "to get at the phenomena under investigation automatically, with a marginal amount of manual work" (Virtanen 1997: 308). As stated before, the main point rather seems to be to *facilitate* analyses (cf. also for example Petch-Tyson 2000: 44), and to make the data more widely accessible. Improvement in learner corpus annotation, which is often called for in quantitative studies, will also only marginally reduce the necessary manual work. The only way in which learner corpus analysis can lead to a significant reduction of manual work on the part of the researcher is with an extremely thorough error-annotation. This, however, means that the manual work will merely be shifted from the researcher to the annotator, and that the annotator and not the researcher will then carry out a central part of the analysis.

To sum up, due to the great reliance on automatic analyses and a more general desire to get first results fast, many studies so far have been rather superficial and either very general (e.g. by determining the most frequent words in learner writing) or very specific (by looking at a few single words). What is needed now is more studies, especially more detailed ones, that cover the large middle ground, i.e. studies that investigate certain areas of grammar, lexis or discourse and go beyond single words. Ideally, these studies should start from functions, not from forms. If connectors are looked at, for example, investigating a certain type of connector would be preferable to selecting a few individual connectors and then analysing their use. If forms are the starting point, instances where a certain element should have been used but was not will not be found (unless the corpus is very thoroughly error-tagged, cf. above), and new discoveries are much more unlikely, because in most cases only words or forms that the researcher assumes to be difficult for the learner are analysed. Finally, more studies comparing different levels of learners, different media and text

types are desirable, but this will only be possible on a greater scale once more learner corpora along the lines outlined in the last section become available.

3.2 Learner corpora and pedagogic material

Given the lack of more comprehensive analyses so far, it is not surprising that learner corpus studies have not yet had any remarkable impact on pedagogic material. The only exception are learner dictionaries: for a number of them, systematic learner corpus research has been carried out and the results have been used to compile or improve them. The first dictionary incorporating results from learner corpus analysis was the *Longman Language Activator* (1993); other dictionaries incorporating such results are the *Longman Dictionary of Contemporary English* (1995), the *Longman Essential Activator* (1997), and the *Cambridge International Dictionary of English* (1995). For the *Longman Essential Activator*, for example, the Longman Learner Corpus was used to identify the most common learner errors, which were then listed in so-called 'help-boxes' at the end of the corresponding entry (Gillard & Gadsby 1998: 164). The entry for the verb *mention*, for instance, includes a help-box alerting learners to the fact that whereas both *mention something about* (as in *He mentioned something about a party*) and *mention something* are acceptable, *mention about something* is not. The Longman Learner Corpus was also used for the *Longman Dictionary of Common Errors* (Turton & Heaton 1987). Learner reference grammars, on the other hand, have not been influenced by learner corpus analysis to any significant extent. As far as I am aware, the only grammatical description of English that is based on a learner corpus is that in TeleNex, an internet site for teachers of English in Hong Kong.¹⁰ Besides providing teaching material and giving teachers the chance to ask questions that are answered by other teachers and linguists, TeleNex also includes a grammatical description of English called TeleGram, which emphasises points that are difficult for Chinese learners of English and in which common mistakes are pointed out and explained. For example, if the user looks up "passive voice", the use of the passive voice in English is explained with particular emphasis on the passive of linking verbs, which presents a particular problem for Chinese learners of English, and typical mistakes in this area are presented (cf. Figure 2).

As with grammars, there has been no influence so far on printed teaching material such as textbooks or workbooks, but some suggestions have been made on how textbooks could benefit from learner corpus analysis (e.g. Kaszubski 1998). CALL programs based on investigations of learner corpora have, however, started to appear. One of the pioneering projects in this field

Using passive voice

Students' problems

Passive voice with linking verbs

Students quite often try to use passive voice with linking verbs which cannot be passivised:

Although I was so frightened, I still tried my best to give first aid to the poor manager. For a while the street was remained quiet, but then I could clearly hear the robbers and the policemen exchanging fire again.

(TeleNex Students)

→

*For a while the street **remained** quiet, but then I could clearly hear the robbers and the policemen exchanging fire again.*

Seeing a film in the cinema is cost fifty dollars. (TeleNex Students)

→

*Seeing a film in the cinema **costs** fifty dollars.*

A controversy has been lasted for some time concerning one of the most popular types of film in Hong Kong which deals with local gangsters. (TeleNex Students)

→

*A controversy **has lasted** for some time concerning one of the most popular types of film in Hong Kong which deals with local gangsters.*

Figure 2. Example page from TeleGram

is AutoLANG, an interactive teaching program for advanced Chinese-speaking students of English, which can be used to practise areas that are difficult for this learner group. In this program, the learner is asked to identify and explain common errors that are hidden in texts; help is provided if the learner asks for it (cf. Milton 1998; for a similar project also partly based on learner corpora cf. Osborne 2000). A further, related, area in which learner corpus results are starting to be applied is writing tools. One on-going project aims at adapting English style and grammar checkers for French-speaking users (Granger 1998c). A further project aims at the development of a tool that detects collocational errors in the writing of Chinese learners of English (Shei & Pain 2000). Any more general influence of learner corpora on syllabuses, finally, does not seem to have taken place yet.

It is too early yet to make specific suggestions as to the incorporation of results from learner corpus studies in pedagogic material. But if what is start-

ing to show is borne out by further studies, namely that the L1 influence in many areas is greater than commonly assumed, one of the most important consequences of learner corpus analysis should be that more material is made more L1-sensitive. Once more detailed and more comprehensive results become available, these can then be applied to pedagogic material in several different ways. First, they can help to decide what features should be particularly emphasized in teaching or even lead to the introduction of hitherto neglected elements (such as certain formulaic sequences, for example). A contribution to the selection of what is to be taught can be made not only by identifying difficulties – though this is the most important way – but also by identifying what is particularly useful for learners, especially beginners (i.e. what elements they can use widely and are able to use successfully). Secondly, results from learner corpus studies can also give indications on how to teach certain features. From her analysis of advanced learner tense use, for instance, Granger concludes that tenses need to be taught at discourse-level instead of sentence-level and that, at an advanced level, tenses should partly be looked at contrastively (1999:200). Thirdly, results on developmental sequences can help to determine in what order language features should be taught. If teaching follows the developmental sequence, language acquisition is likely to be better than if it runs counter to this sequence (Ellis 1994:632ff.). Finally, learner corpora can be more directly used to provide examples of typical mistakes and typical cases of overuse and underuse in teaching and in reference materials. To what extent and how (i.e. on what levels, with what kinds of features) this is useful for language learning will have to be investigated.

3.3 Learner corpora and data-driven learning

While it will probably take some time before learner corpora can play the role they deserve in the improvement of pedagogic material, data-driven learning can be attempted straight away by anybody who has access to a learner corpus or is willing to create one. Nevertheless, using learner corpora for data-driven learning has only been suggested by a few researchers (e.g. Granger & Tribble 1998), and so far there have been only a few attempts to actually try it out (Flowerdew 2001; Horvath 2001; Milton & Hyland 1999; Ragan 2001).¹¹ Of course, data-driven learning on the basis of learner corpora is only useful to the extent that focused negative evidence supports language learning. But although it is not known exactly to what degree or how, it seems very likely that negative evidence is useful at least to a certain degree and under certain circumstances (cf. for example Carrol & Swain 1993; Ellis 1994:639ff.). This approach should

therefore certainly not be regarded as a panacea. Nevertheless, it can be reasonably speculated that especially in the case of advanced learners, and especially for forms that have become or are becoming fossilized, focused negative evidence can be a good way to aid language acquisition (Granger 1996b: 5). And in those cases where focused negative evidence is useful, data-driven learning has a number of advantages over merely alerting learners to their mistakes. One of these advantages is that asking learners to look for mistakes, or rather for differences in learner and native speaker language, can increase learner autonomy and train the learners' general ability to notice such differences. In addition, such a procedure might also lead to a more positive attitude towards mistakes, because mistakes are then no longer merely a feature that has to be corrected, but also a feature that can be discovered. As has been reported, learners find it motivating to identify and analyse mistakes (Fan et al. 1999: 187). Data-driven learning with learner data is probably particularly useful for points which have already been covered in the classroom, possibly even repeatedly, but which the learners nevertheless still get wrong. In this way, instead of being told once again that what they are doing is wrong, learners have the opportunity to get something right, namely to identify and explain the mistake in question.

When using learner corpora for data-driven learning, a completely different kind of evidence is dealt with than when using native corpora, so that there are a number of differences between the two kinds of data-driven learning, which need to be pointed out. First of all, since negative evidence naturally is only useful if the learner is aware that it is *negative* evidence, what has been called 'divergent learning' (Leech 1997: 11), i.e. browsing corpora to discover new facts about the language, is out of the question with learner corpora. Secondly, learners always have to be provided with positive evidence in addition to the evidence from the learner corpus. This evidence can come from either a comparable native speaker corpus or from a general native speaker corpus such as the BNC. Thirdly, since there is the danger that the learner will remember the negative but not the positive evidence or that the positive evidence will remain partly undigested (cf. Milton & Hyland 1999: 158), the exposure to negative evidence should be followed by exercises to consolidate the native speaker structure (cf. also Granger 1996b).

Two different types of data can be used by teachers who want to use learner corpora for data-driven learning: teachers can either use an already existing corpus containing language from learners comparable to their group (e.g. as to level and L1), or they can use data from their own group of learners, i.e. create their own learner corpus. Whereas the latter option is naturally more time-consuming (though less so if the students hand in their assignments in

electronic form), it may be inevitable if teaching a group for which no learner corpus exists. Such a self-made corpus has the advantage of being more relevant to the needs of the students and therefore possibly also more motivating (cf. Seidlhofer 2000a: 212ff.). Whether the possible embarrassment over being exposed to one's own errors is stronger than this possible motivation remains to be seen; in any case, the texts included should of course be anonymized.

Independent of the nature of the data used, some areas of language lend themselves more easily to learner corpus analysis in the classroom than others. Probably best suited are co-occurrences of words, especially if the co-occurring words are adjacent, such as prepositions or complementation of verbs, nouns and adjectives. For example, concordance lines from the German subcorpus of ICLE and the corresponding native speaker corpus LOCNESS (the Louvain Corpus of Native English Essays) could be used to make German-speaking learners aware of the fact that complementation of *suggest* with *to*+infinitive occurs frequently in their English, but does not correspond to native-speaker usage (cf. Figure 3 and 4).¹²

women themselves as suggested	above. However,
may protest, or whether we suggest	alternatives to what
suffer most of all. I would suggest	applying some
human problems, this theory suggests	each individual,
to her as the book says, it suggests	"fish" as a potential
magazine "Home and Garden" suggests.	Garden centres
as Bertrand Russell rightly suggests	in his statement.
the book first. Perhaps, so I suggest,	it would be helpful
to individual problems. This suggests	returning to smaller
of the problem. This notion suggests	that all problems
Smith, author of the article, suggested	that all cars should
of our time to make. It suggests	that his own works
The authors of this bill suggest	that there should be
of their voices obviously suggested	that they
takes root in the ideology suggesting	that truth is not
me for anything. I therefore suggest	that we stop discussing
to the notion of truth, I would suggest	to define photography
this won't happen. Russell suggests	to go on holiday.
tendencies, and furthermore suggests	to neglect relevant
air in town. The other could suggest	to her two colleagues
(on the contrary, I would suggest	to give them more

Figure 3. *Suggest(s/ed/ing)* in the German subcorpus of ICLE

AIDS. A program like the one suggested for New York City by
 and this shall be followed by suggested solutions to these
 lost sovereignty. It has been suggested than when the U.K.
 the Lottery and the arguments suggested that as opposed to
 experience of AIDS educators suggests that condom distribution
 would support the theory that suggests that is does not act as
 discovering far less to uncover, suggests that some specialisation
 nations of the union. It has been suggested that the Single Union
 United States. What is absurd is to suggest that the movement has
 recently in the Sun the editorial suggested that there should be
 offered these facts some try to suggest that we eliminate the
 executions. The Dann study also suggests that. This would
 by no means totally blameless. I suggest the human brain could
 There is strong evidence to suggest this causes asthma and
 arguments which these people suggested were numerous. First

Figure 4. *Suggest(s/ed/ing)* in LOCNESS

Wrong uses of individual words can also be a suitable focus of data-driven learning, but only if the wrong use is obvious from the immediate context. A search for *economical* in the German and French ICLE subcorpora, for example, yields many occurrences of the kind **economical problem*, **economical advantages*, **economical reform*, **economical interest*, which can easily be identified as confusions of *economical* with *economic*. In addition to revealing mistakes, the comparison of data from (comparable) learner and native speaker corpora can also reveal an overuse and underuse of particular words. A comparison of learner and native speaker concordance lines of a general word (such as *important*) and more specific words with a similar meaning (such as *critical*, *crucial*, *major*, *serious*, *significant*, *vital*), for example, is likely to reveal that the general word is overused and the more specific words are underused by the learners (Tribble & Granger 1998: 206f.).

What is less suitable for data-driven learning with learner corpora is more general grammatical areas, such as tense or aspect, because searching for them is more difficult and because the immediate context does not necessarily reveal whether a certain instance is correct or not. The use of single words is also difficult to investigate if they occur too rarely or too frequently in the given corpus. Moreover, a given corpus often fails to yield suitable or enough examples of a certain phenomenon, although it is in principle well-suited for data-driven learning. Take the combination **an example for* (as in **The Waste Land is an*

the game. He always tried to be an example for fairness on the field
conscious, because then we set an example for industry and politicians
First of all, new jobs are created, for example for people working in
users of the passed years. Another example for a “devotee” of

Figure 5. *example for* in the German subcorpus of ICLE

say that this ‘idol’ talk sets a bad example for children when growing
Child care work makes a good example for this point. American
criminals in a ways that will set an example for others, and probably

Figure 6. *example for* in LOCNESS

example for a modernist poem), which is a common mistake among German-speaking learners of English (modelled on German *ein Beispiel für*). Since the words are adjacent and the combination is neither particularly rare nor particularly common, it could be assumed that the German subcorpus of ICLE and LOCNESS would provide suitable concordance lines for analysis in the classroom. However, this is not the case, as can be seen in Figures 5 and 6: *example for* occurs four times in the German subcorpus, but also three times in LOCNESS. Moreover, of the four occurrences in the learner corpus, two are correct (line 2 and 3) and only one of the other two could be corrected by simply replacing *for* by *of* (in line 1, *example* should probably additionally be changed to *model*). Although the lines could still be analysed, they are not particularly suitable for the classroom.

This example shows that it is essential for the teacher to thoroughly prepare learner corpus concordancing if it is to be done by the students themselves. With learner language even more than with native speaker language, the students are likely to be confused either by some of the occurrences or by their very number. Therefore, editing concordance lines and providing them on paper might often be the better alternative to letting the students work directly with the corpus (cf. also Granger 1996b). In the case of *suggest* above, for example, it might also be preferable to provide edited lines – unless the learners are experienced in working with concordance lines. As the lines are now, they could be confusing for learners in many respects: there is at least one typographical error (*suggested than* instead of *that* in LOCNESS); one of the occurrences of *to* after *suggest* does not constitute wrong complementation (*could suggest to her two colleagues*); and, as *suggest + ing* only occurs in the learner but not in

the native speaker corpus, learners might even come to the conclusion that this construction is not possible in English.

Assuming the teacher has found suitable examples for the phenomenon to be discussed, the procedure in the classroom might be as follows. First, the teacher presents concordance lines taken from a learner and a native speaker corpus. The students then have to work out the differences between the two, possibly supported by specific questions, especially if the differences are less obvious (for examples see Milton & Hyland 1999:157; Granger & Tribble 1998:202ff.). A variant of this procedure could be to present native speaker usage first and then the learner data. This variant seems particularly useful if the native speaker use is well-known to the learners but the learners' frequent deviance is not (for example in areas of underuse). If the learner data, ideally from the group itself, is presented shortly after the native speaker data, the surprise about the fact that a pattern that seems entirely obvious is often not used or misused in the language of learners could make the insight longer-lasting. Finally, the phenomenon should be practised with further exercises, for example with concordance lines in which the word in question has been edited out. Practising is probably more efficient if several related phenomena are looked at in one session. If only the complementation of *suggest* is looked at, for example, follow-up exercises are likely to be too easy and therefore demotivating. This danger can be avoided if a few other verbs of which the complementation is problematic for the learners (such as *afford*, *accept*, *recommend*) are dealt with at the same time.

Compared to data-driven learning on the basis of native speaker corpora, then, the teacher plays a more important role in data-driven learning on the basis of learner corpora, both with respect to what is investigated and how it is done. But given that a method like the one outlined above is used, data-driven learning with learner corpora seems to have a rather high potential at least for some areas of language. What remains to be investigated is exactly for which areas, for which learners and with what procedures data-driven learning with learner corpora is most efficient.

4. Conclusion

The potential of learner corpora is not restricted to language teaching, and learner corpora are not the only type of corpus with a potential in this field. They also have an important part to play in second language acquisition research, and other corpora, especially native speaker corpora, also have an im-

portant part to play in the improvement of language teaching. However, as I hope to have shown, learner corpora clearly can make a significant contribution to language teaching. Most importantly, they can contribute towards the improvement of pedagogic material through revealing typical difficulties of certain groups of learners. This is of particular relevance for advanced learners, whose difficulties are often rather subtle and therefore not accessible by unsystematic observation. Other ways in which learner corpora can help to improve pedagogic material are more indirect, for example through identifying typical second language acquisition processes or through finding out what words or patterns are particularly useful, especially for certain groups of learners (especially beginners). In the classroom, learner corpora can also be fruitfully used, by providing focused negative evidence and giving learners the opportunity to discover this evidence themselves.

For the application of learner corpora in the classroom itself, only a few preconditions need to be fulfilled: teachers either have to have access to data from learners corresponding to their own group or have to be willing to create a corpus. For the application of learner corpora to the improvement of pedagogic material, more preconditions need to be fulfilled before this potential can be used to any significant extent. More well-designed learner corpora need to be compiled or completed (for different L2s, L1s, text types etc.) and be made available for research. In addition, more qualitative, and more co-ordinated and comprehensive analyses of learner corpora need to be carried out.

Learner corpora have begun to attract the interest of quite a number of researchers; it can even be assumed that “we are on the verge of a learner corpus boom” (Granger 1998: xxii). We also seem to be on the verge of a learner corpus studies boom. If we manage to channel this boom by co-ordinating both learner corpus compilation and learner corpus analysis, and if the improvement of language teaching continues to be one of our aims, the impact of learner corpora on language teaching is bound to be considerable.

Notes

* I would like to thank Ute Römer and John Sinclair for helpful comments on earlier versions of this paper.

1. ‘Pedagogic material’ is used in a very broad sense in this paper, including not only textbooks, grammars and dictionaries, but also for example CALL programs, writing tools and even syllabuses.

2. Other criteria besides degree of difficulty, such as usefulness, should of course also be taken into account.
3. Very recently, compilation of a type of non-native speaker corpus which is called "English as a Lingua Franca (ELF) corpus" has begun, (e.g. Seidlhofer 2000b). ELF corpora aim at recording communication between speakers of English from a variety of first language backgrounds and are compiled for the investigation of common features of ELF. The difference between an ELF corpus and a learner corpus therefore lies primarily in their purpose and not necessarily in their composition.
4. Probably the only exception is the Longman Learner Corpus, which was started in the late 1980s. An example of one of the few less typical learner corpora from the 1980s is the European Science Foundation Second Language Databank, a corpus of immigrant speech of different languages. This corpus cannot be considered typical, since the number of subjects of each L1-L2 pair included is small (four) and some of the data was elicited with a high degree of control, but it meets all the other criteria of a learner corpus, including computerization (cf. Perdue 1993 and <http://www.mpi.nl/world/tg/lapp/esf/esf.html> (26.6.02)).
5. An interesting, but clearly more distant possibility would be to create a truly developmental corpus (i.e. a corpus containing data from the same learners collected over a period of time), and possibly even to add what might be called an 'input corpus', in order to investigate the influence of input to acquisition. Such an input corpus, which naturally would only be possible in countries where the L2 in question is a foreign language, could contain transcribed recordings of all classroom interaction as well as all written texts that the learners have dealt with in their course.
6. This is of course not to say that small learner corpora cannot yield interesting results (cf. for example Fan et al. 1999: 187). Moreover, in certain situations it can be inevitable or useful to compile a corpus individually, especially in order to investigate the language of a group of learners with very specific needs (e.g. with a rare L1 or attempting to acquire a rare L2).
7. Two of the biggest data-distribution organisations are the *Linguistic Data Consortium* (<http://www ldc.upenn.edu>; 26.6.02) or the *European Languages Resources Association* (<http://www.icp.grenet.fr/ELRA/home.html>; 26.6.02).
8. A corpus recording learner phonology is the ISLE Corpus of non-native spoken English (cf. <http://nats-www.informatik.uni-hamburg.de/~isle> (26.6.02)).
9. It seems likely that this is just one aspect of a more general characteristic of learner language, namely a less clear distinction between different levels of formality than in native speaker language.
10. Cf. <http://www.telenex.hku.hk/telec/pmain/opening.htm> (26.6.02) and the paper by Tsui in the present volume.
11. Seidlhofer (2000a) describes a different use of learner corpora in the classroom: she uses a self-made learner corpus to introduce students to corpus linguistics.
12. LOCNESS contains writing from British and American students. Slightly reduced versions of the ICLE subcorpora and LOCNESS are used in this section, each containing about 150,000 words. I would like to thank the coordinator of German ICLE, Gunter Lorenz, as well as the Centre for English Corpus Linguistics at the Université Catholique de Louvain,

Belgium, for integrating me into the ICLE-project at a late stage and for providing me with data from ICLE and LOCNESS.

References

- Aarts, J. & S. Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In S. Granger (Ed.), *Learner English on Computer* (pp. 132–141). London: Longman.
- Altenberg, B. & M. Tapper (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on Computer* (pp. 80–93). London: Longman.
- Aston, G. (2000). Corpora and language teaching. In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 7–17). Frankfurt: Lang.
- Biber, D. & R. Reppen (1998). Comparing native and learner perspectives on English grammar: A study of complement clauses. In S. Granger (Ed.), *Learner English on Computer* (pp. 145–158). London: Longman.
- Cambridge International Dictionary of English* (1995). Cambridge: Cambridge University Press.
- Carroll, S. & M. Swain (1993). Explicit and negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition*, 15, 357–386.
- Dagneaux, E., S. Denness, & S. Granger (1998). Computer-aided error analysis. *System*, 26, 163–174.
- De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, 3(1), 59–80.
- de Haan, P. (1997). An experiment in English learner data analysis. In I. de Mönnink & H. Wekker (Eds.), *Studies in English Language and Teaching: In Honour of Flor Aarts* (pp. 215–229). Amsterdam: Rodopi.
- de Haan, P. (1998). How 'native-like' are advanced learners of English? In A. Renouf (Ed.), *Explorations in Corpus Linguistics* (pp. 55–65). Amsterdam: Rodopi.
- de Mönnink, I. (1999). Combining corpus and experimental data. *International Journal of Corpus Linguistics*, 4(1), 77–111.
- Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
- Fan, M., C. Greaves & M. Warren (1999). Identifying characteristic patterns in students' writing using a corpus of learner data. In R. Berry, B. Asker, & K. Hyland (Eds.), *Language Analysis, Description and Pedagogy* (pp. 147–161). Hong Kong: Language Centre HKUST.
- Fillmore, C. J. (1992). 'Corpus linguistics' or 'computer-aided armchair linguistics'. In J. Svartvik (Ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991* (pp. 35–60). Berlin: de Gruyter.

- Flowerdew, L. (2000). Investigating referential and pragmatic errors in a learner corpus. In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 145–154). Frankfurt: Lang.
- Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, A. Henry, R. Roseberry (Eds.), *Small Corpus Studies and ELT: Theory and Practice* (pp. 363–379). Amsterdam: John Benjamins.
- Gillard, P. & A. Gadsby (1998). 'Using a learners' corpus in compiling ELT dictionaries. In S. Granger (Ed.), *Learner English on Computer* (pp. 159–171). London: Longman.
- Gitsaki, C. (1999). *Second Language Lexical Acquisition: A Study of the Development of Collocational Knowledge*. Bethesda, MD: International Scholars Publications.
- Granger, S. (1996a). Learner English around the world. In S. Greenbaum (Ed.), *Comparing English Worldwide: The International Corpus of English* (pp. 13–24). Oxford: Clarendon.
- Granger, S. (1996b). Exploiting learner corpus data in the classroom: Form-focused instruction and data-driven learning. Paper presented at TALC 1996, Lancaster, 9–12 August 1996.
- Granger, S. (1997a). Automated retrieval of passives from native and learner corpora. *Journal of English Linguistics*, 25(4), 365–374.
- Granger, S. (1997b). On identifying the syntactic and discourse features of participle clauses in academic English: Native and non-native writers compared. In I. de Mönnink & H. Wekker (Eds.), *Studies in English Language and Teaching: In Honour of Flor Aarts* (pp. 185–198). Amsterdam: Rodopi.
- Granger, S. (1998a). Introduction. In S. Granger (Ed.), *Learner English on Computer* (pp. xxi–xxii). London: Longman.
- Granger, S. (1998b). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer* (pp. 3–18). London: Longman.
- Granger, S. (1998c). The computer learner corpus: A testbed for electronic EFL tools. In J. Nerbonne (Ed.), *Linguistic Databases* (pp. 175–188). Stanford: CSLI.
- Granger, S. (1998d). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, Analysis, and Applications* (pp. 145–160). Oxford: Clarendon.
- Granger, S. (1999). Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus. In H. Hasselgård & S. Oksefjell (Eds.), *Out of Corpora: Studies in Honour of Stig Johansson* (pp. 191–202). Amsterdam: Rodopi.
- Granger, S. & P. Rayson (1998). Automatic profiling of learner texts. In S. Granger (Ed.), *Learner English on Computer* (pp. 119–131). London: Longman.
- Granger, S. & C. Tribble (1998). Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on Computer* (pp. 199–209). London: Longman.
- Horváth, J. (2001). *Advanced Writing in English as a Foreign Language: A Corpus-based Study of Processes and Products*. Pécs: Lingua Franca Csoport.
- Kaszubski, P. (1998). Enhancing a writing textbook: A national perspective. In S. Granger (Ed.), *Learner English on Computer* (pp. 172–185). London: Longman.

- Kaszubski, P. (2000). Lexical profiling of English (learner) corpora: Can we measure advancement levels? In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *PALC '99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Łódź, 15–18 April 1999* (pp. 249–286). Frankfurt: Lang.
- Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second Language Writing: Research Insights for the Classroom* (pp. 140–154). Cambridge: Cambridge University Press.
- Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 1–23). London: Longman.
- Leech, G. (1998). Preface. In S. Granger (Ed.), *Learner English on Computer* (pp. xiv–xx). London: Longman.
- Longman Language Activator* (1993). London: Longman.
- Longman Dictionary of Contemporary English* (1995). London: Longman.
- Longman Essential Activator* (1997). London: Longman.
- Lorenz, G. R. (1998). Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification. In S. Granger (Ed.), *Learner English on Computer* (pp. 53–66). London: Longman.
- Lorenz, G. R. (1999). *Adjective Intensification – Learners versus Native Speakers: A Corpus Study of Argumentative Writing*. Amsterdam: Rodopi.
- Man-Lai, A. (1994). Collocational problems amongst ESL learners: A corpus-based study. In L. Flowerdew, A. K. K. Tong (Eds.), *Entering Text* (pp. 157–165). Hong Kong: University of Science and Technology.
- McEnery, T. & A. Wilson (2001). *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press.
- McEnery, T. & A. Wilson (1997). Teaching and language corpora (TALC). *ReCALL*, 9(1), 5–14.
- Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (Ed.), *Learner English on Computer* (pp. 186–198). London: Longman.
- Milton, J. & R. Freeman (1996). Lexical variation in the writing of Chinese learners of English. In C. E. Percy, C. E. Meyer & I. Lancashire (Eds.), *Synchronic Corpus Linguistics. Papers from the Sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16)* (pp. 121–131). Amsterdam: Rodopi.
- Milton, J. & K. Hyland (1999). Assertions in students' academic essays: A comparison of English NS and NNS student writers. In R. Berry, B. Asker, K. Hyland (Eds.), *Language Analysis, Description and Pedagogy* (pp. 147–161). Hong Kong: Language Centre HKUST.
- Osborne, J. (2000). What can students learn from a corpus? Building bridges between data and explanation. In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 165–172). Frankfurt: Lang.
- Perdue, C. (Ed.). (1993). *Adult Language Acquisition: Cross-Linguistic Perspectives*. 2 vols. Cambridge: Cambridge University Press.

- Petch-Tyson, S. (2000). Demonstrative expressions in argumentative discourse: A computer corpus-based comparison of non-native and native English. In S. Botley & A. M. McEnery (Eds.), *Corpus-based and Computational Approaches to Discourse Anaphora* (pp. 43–64). Amsterdam: John Benjamins.
- Ragan, P. H. (2001). Classroom use of a systemic functional small learner corpus. In M. Ghadessy, A. Henry, R. Roseberry (Eds.), *Small Corpus Studies and ELT: Theory and Practice* (pp. 207–236). Amsterdam: John Benjamins.
- Ringbom, H. (1998a). High-frequency verbs in the ICLE corpus. In A. Renouf (Ed.), *Explorations in Corpus Linguistics* (pp. 191–200). Amsterdam: Rodopi.
- Ringbom, H. (1998b). Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In S. Granger (Ed.), *Learner English on Computer* (pp. 41–52). London: Longman.
- Seidlhofer, B. (2000a). Operationalizing intertextuality: Using learner corpora for learning. In L. Burnard & T. McEnery (Eds.), *Rethinking Language Pedagogy from a Corpus Perspective* (pp. 207–224). Frankfurt: Lang.
- Seidlhofer, B. (2000b). Mind the gap: English as a mother tongue vs. English as a lingua franca. *Vienna English Working Papers*, 9(1), 51–68.
- Shei, C. C. & H. Pain (2000). An ESL writer's collocational aid. *Computer Assisted Language Learning*, 13(2), 167–182.
- Tono, Y. (1998). A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. Paper presented at TALC, Oxford, 24–27 July 1998.
- Tono, Y. (2000). A corpus-based analysis of interlanguage development: Analysing POS tag sequences of EFL learner corpora. In B. Lewandowska-Tomaszczyk, P. J. Melia (Eds.), *PALC'99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Łódź, 15–18 April 1999* (pp. 323–340). Frankfurt: Lang.
- Turton, N. D. & J. B. Heaton (1987). *Longman Dictionary of Common Errors*. Harlow: Longman.
- Virtanen, T. (1997). The progressive in NS and NNS student compositions: Evidence from the International Corpus of Learner English. In M. Ljung (Ed.), *Corpus-based Studies in English. Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17). Stockholm, May 15–19, 1996* (pp. 299–309). Amsterdam: Rodopi.
- Yang, H. (2001). Computer analysis of Chinese learner English. Paper presented at “How to Use Corpora in Language Teaching”, Tuscan Word Centre, 15–21 October 2001.

Appendix

Current learner corpora and learner corpus projects

Written learner corpora of English – one L1:

– Hong Kong University of Science and Technology (HKUST) Learner Corpus: ~25 million words, L1 Chinese, undergraduate students, different aca-

demic text types

(cf. Milton and Freeman 1996 and <http://leo.meikai.ac.jp/~tono/> (26.6.02))

– **TeleNex Student Corpus** or **TELEC Secondary Learner Corpus (TSLC)**: ~ 3 million words (aim: 10 million), L1 Chinese, secondary school students of different levels, compositions

(<http://www.telenex.hku.hk/telec/smain/sintro/intro.htm> (26.6.02))

– **Chinese Learner English Corpus (CLEC)**: ~ 1.2 million words, L1 Chinese, different levels (middle school students to senior college English majors), compositions

(cf. Yang 2001)

– **Uppsala Student English Project (USE)**: ~ 1 million words, L1 Swedish, university students of English, essays of different types

(<http://hem2.passagen.se/ylvaberg/useinfo1.htm> (26.6.02))

– **Corpus of English by Japanese Learners (CEJL)**: ~ 1 million words, L1 Japanese, different levels (junior high school to university students), different text types

(cf. <http://www.lb.u-tokai.ac.jp/lcorpus/> (26.6.02))

– **Taiwanese Learner Corpus of English (TLCE)**: ~ 700,000 words, L1 Chinese, university students of English

(cf. <http://leo.meikai.ac.jp/~tono/> (26.6.02))

– **Janus Pannonius University (JPU) Corpus**: ~ 500,000 words, L1 Hungarian, university students, different academic text types

(cf. Horváth 2001 and http://www.geocities.com/jpu_corpus (26.6.02))

– **Polish Learner English Corpus (PLE)**: ~ 500,000 words, L1 Polish, different levels, different text types

(<http://www.uni.lodz.pl/pelcra/intro.htm> (26.6.02))

– **Québec Learner Corpus**: ~ 250,000 words, L1 French, different levels, essays

(http://www.er.uqam.ca/nobel/r21270/cv/QLCorpus/QL_Corpus_files/frame.htm (26.6.02))

– **Learner Business Letters Corpus (BLC)**: ~ 200,000 words, L1 Japanese, business letters

(<http://isweb9.infoseek.co.jp/school/ysomeya/> (26.6.02))

Written learner corpora of English – different L1s:

– **Longman Learner Corpus (LLC)**: ~ 10 million words, ~ 160 different L1s, different levels, different text types

(<http://www.longman-elt.com/dictionaries/corpus/lclearn.html> (26.6.02))

- **Cambridge University Press (CUP) Learner Corpus**: ~ 10 million words, many different L1s, intermediate and advanced learners, exam scripts (http://www.cambridge-efl.org/rs_notes/0001/rs_notes1_6.cfm (26.6.02))
- **International Corpus of Learner English (ICLE)**: ~ 2 million words, 14 different L1s, university students of English, argumentative essays (cf. Granger 1996a and <http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/introduction.html> (26.6.02))

Written learner corpus of French:

- **French Interlanguage Database (FRIDA)**: ~ 200,000 words, different L1s, intermediate learners, different text types (<http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Frida/frida.htm> (26.6.02))

Spoken learner corpora of English:

- **Louvain International Database of Spoken English Interlanguage (LIND-SEI)**: aim: 1 million words, a number of different L1s, university students of English, informal interviews (cf. De Cock 1998 and <http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/introduction.html> (26.6.02))
- **Standard Speaking Test (SST) Corpus**: aim: 1 million words, L1 Japanese, different proficiency levels, interviews (cf. <http://leo.meikai.ac.jp/~tono/> (26.6.02))

Research

Composition

The use of adverbial connectors in Hungarian university students' argumentative essays

Gyula Tankó

Eötvös Loránd University

The problematic nature of the use of connectors in non-native writers' English texts has been revealed and investigated by several research studies. This paper presents a study of the use of adverbial connectors in high-rated argumentative essays written by Hungarian advanced learners of English and contends that a corpus-based data-driven approach to the teaching of adverbial connectors should be adopted. The in-depth study of the corpus of argumentative texts combines corpus linguistics and discourse analytical approaches and reveals that Hungarian writers share the problems of writers with other cultural backgrounds. Hungarian writers tend to use more yet slightly fewer types of adverbial connectors than native speakers. The distribution of adverbial connectors in the essays is uneven, few phrasal equivalents are used, and there is a common tendency to use listing adverbial connectors to mark the superstructure of the text. The results show that Hungarian writers use a large number of resultative and contrastive adverbial connectors correctly to mark the reasoning processes in their texts, have an adequate awareness of the characteristic positions of adverbial connectors in English, and are familiar with the stylistic requirements of the academic register.

1. Introduction

A number of studies investigating coherence report on non-native students' use of connectors¹ in academic texts written in the English language. For example, Mauranen (1993) studied the use of connectors in Finnish writers' texts, Wikbork and Björk (1989) focused on the use of connectors in Swedish university students' texts, Altenberg and Tapper (1998) compared Swedish learners' L1 and English texts and contrasted Swedish learners' English texts with those

written by French students. Each of these studies reveals various types of shortcomings in the use of connectors in EFL writers' texts. The comparative analysis of Finnish and Anglo-American academic texts showed that Finnish writers tended to avoid the use of overt connectors in their English texts. This tendency was attributed to L1 influence, since the emphasis on the indication of textual relations is less marked in Finnish. The misuse (i.e. the incorrect use of connectors) and underuse (i.e. the presence of markedly fewer connectors in the non-native texts than in the native texts) of particular connectors was found to be the cause of the coherence breaks identified in Swedish writers' texts. Furthermore, both Swedish and French learners were reported to lack the register awareness necessary for the appropriate use of connectors in academic writing.

Apparently, the use of connectors is a problematic feature of non-native writers' English language texts. There has been no study conducted yet to investigate whether the use of connectors in formal texts produced by Hungarian native speakers writing in English is marked by similar discrepancies. This paper intends to make a preliminary exploratory step towards filling this gap through the presentation of the results of a small scale study that focuses on the use of adverbial connectors in high-rated argumentative essays written by advanced learners of English.

Although the study presents an in-depth analysis of one feature of a small sample of texts and in this respect pertains to the domain of Discourse Analysis, it relies on Corpus Linguistics in a number of ways. Corpus Linguistics provided the design criteria for the principled compilation of the corpus used in the study, the tools to retrieve the specific information sought, and the observation protocol, i.e. the paradigmatic examination of repeated instances of a linguistic phenomenon in order to draw conclusions on the basis of the similarities observed (e.g. colligations). Furthermore, the concordance-based activity suggested at the end of the study relies heavily on the findings of large scale corpus linguistic studies that investigated the use of connectors in texts produced by natives.

The paper is structured as follows: the second section reviews the stance of writing theorists regarding the role of connectors and discusses the linguistic and methodological factors that can render difficult for ESL and EFL writers the acquisition and appropriate use of connectors. The third section presents the research questions. The fourth section describes the origin and composition of the corpus compiled for the purposes of the investigation as well as the theoretical framework adopted for the study. The fifth section discusses the findings and contrasts them to those reported in a study on native speaker students' argumentative texts. The sixth section is the conclusion, and the seventh

section summarises the implications of the study for teaching and presents a concordance-based classroom activity for the teaching of connectors.

2. A brief overview of issues concerning the teaching of connectors

2.1 The significance of connectors

Works on writing theory and discourse emphasize the importance of connectors in the production of good quality texts and assert that language learners should be coached in their use. McCarthy (1991:50) states that advanced L2 learners' output data can seem unnatural because, unlike native speakers, they cannot employ a variety of appropriate connectors to express cognitive relations and this decreases the comprehensibility of their texts (p. 155). The correct use of connectors is considered important not only for the explicit signalling of connections, but because as a means for the indication of attitudes and emphases they also serve a rhetorical purpose (McCarthy & Carter 1994:50). Cook (1989) states that "language learners need to know both how and when to use them. Their presence or absence in discourse often contributes to style, and some conjunctions can sound very pompous when used inappropriately" (p. 21). Zamel (1983:27) further remarks that the overuse of connectors results in artificial, mechanical prose; therefore, writing instructors have to teach learners both when to use and also when not to use connectors.

McCarthy & Carter advocate that teachers should initiate whole-class discussions on the use of conjunctions, they should familiarise students with lists of conjunctive words and phrases, should encourage them in employing such words in their writing, and finally should sensitize them to the genre sensibility of these words and phrases (1994:50).

There is thus a well-grounded agreement about the importance of connectors. However, teaching learners why, when, and how to use connectors so that their written output approximates the norms of native texts is not an easy undertaking.

2.2 Why are connectors difficult?

The sources of difficulty related to the use of connectors are diverse and rooted in their discourse-organising function, grammatical, semantic and morphological attributes, and also in shortcomings in the techniques employed to teach these devices.

2.2.1 *Linguistic factors*

Halliday and Hasan (1976:226) note that conjunction, the fifth type of cohesive relation, is not as easily classifiable as reference, substitution, or ellipsis as it combines the features of both grammatical and lexical cohesion. They contend that conjunction establishes a relation between meanings rather than grammatical units; it is a semantic relation that instead of simply marking *which* elements are connected conveys information on *how* the elements are connected.

Conjunctive elements are cohesive not in themselves but indirectly, by virtue of their specific meanings; they are not primarily devices for reaching out into the preceding (or following) text, but they express certain meanings which presuppose the presence of other components in the discourse.

Furthermore, the linguistic unit's connectors span can vary from clauses to paragraphs and even longer stretches of discourse (Hatch 1992:266; Quirk et al. 1985:632). Learners thus first need to familiarize themselves with the meaning of the individual connectors, the type of components they normally occur with in discourse, and finally the distance they can span in order to be able to identify the type and expanse of the relation they indicate. However, if learners turn to monolingual dictionary entries for help, including those that illustrate the use of a connector in a sample sentence, they do not always find this kind of information (cf. the informal *anyhow*, p. 44; or the formal *furthermore*, p. 503; and *hence*, p. 583 in the *Oxford Advanced Learner's Dictionary of Current English*).

Another characteristic feature of connectors is, as Hatch (1992:225) and McCarthy (1991:48) point out, that there is no one-to-one correspondence of connectors and their functions. For example, the one word conjunct *then* can be found in Halliday and Hasan's (1976) categorisation in three subcategories: in the *sequential* group within the *temporal* category; and within both the *simple* and *conditional* groups in the *causal* category (p. 242). Similarly, in Quirk et al.'s (1985) classification of adverbial connectors *then* is entered twice under the *listing* category (*enumerative* and *reinforcing*), as well as under the *summative*, *inferential*, and *contrastive* (*antithetic*) categories (p. 634). Other connectors that illustrate the same problem but are more frequent in Hungarian student writing are *on the one hand* and *on the other hand*, which are found in the *enumerative* subcategory under the *listing* conjuncts as well as in the *antithetic* subcategory under the *contrastive* conjuncts in Quirk et al.'s classification (p. 635); or *thus*, which can have both a *summative* and an *appositive* function.

Moreover, even the use of such basic connectors as *and* or *or* may become equivocal because of the dissimilar relations they can connote (Quirk et al., p. 930). Discussing the uses of co-ordinating conjunctions (p. 930), Quirk et al. point out that *and* can connote several relations, namely *consequence/result*, *temporal sequence*, *contrast*, *concession*, *condition*, *similarity*, *addition*, and finally *comment/explanation*. Similarly, *or* can function as a connector that is *exclusive*, *inclusive*, *corrective*, or that expresses a negative condition (p. 932).

A further source of difficulty for language learners is the variety of connectors. A list of conjunctive words or phrases aiming at comprehensiveness is difficult and futile to memorise as regular lexical items, and attempts at the selection of a representative set for instruction have proved ineffective. The range and type of connectors language learning course books or writing books generally focus on varies to a large extent, and most of the time the selection procedure of the connectors included in the teaching material is not informed by empirical evidence (cf. Biber et al. 1999:875–892). Learners thus receive either an exhaustive list of connectors, which is daunting and impractical, or a randomly selected one that does not include the most frequently used connectors that would be of real help for apprentice writers uncertain in their handling of such cohesive devices.

Furthermore, the form connectors can take is diverse and can be confusing to learners: McCarthy (1991:46) lists example sentences in which the same cognitive relation is realised with the use of the one-word, phrasal and clausal forms of the same adverbial connector. Textbooks do not explain the difference between the possible forms, do not supply guidelines concerning the use of the various options and thus leave it to learners to make uninformed and consequently often incorrect choices. Such trial and error approaches to the use of connectors can easily be counterproductive unless students receive expert feedback that reinforces the correct choices and highlights the wrong ones. One of the major problems with the lists of connectors found in most textbooks is, as Zamel (1983:24) remarks, that they prompt the conclusion that the items categorised together are freely interchangeable:

Yet another serious problem is the fact that devices categorized together are not necessarily interchangeable: 'but' and 'however' cannot be substituted for 'on the contrary' or 'on the other hand', although they are often classified together. Even when linking devices in a list do serve similar semantic functions, however, the fact that they may carry different grammatical weight causes other difficulties.

Such lists do not only usually ignore the fact that connectors have different syntactic functions; they also ignore the stylistic value of connectors, and therefore learners do not become aware of their genre sensitivity (cf. Conrad, this volume). Such lists do not provide information on the most frequent positions characteristic of linking devices, or on such secondary yet also important features as the punctuation rules that are to be applied with them.

2.2.2 *Methodological factors*

The major methodological shortcomings in the teaching of connectors revealed by previous research (e.g. Crewe 1990; Zamel 1983) are (a) the inappropriate guiding principles followed in the compilation of categorisations of connectors, (b) the presentation of connectors in lists without their context, and (c) the treatment of connectors primarily as stylistic devices.

Zamel (1983:24), notes that the presentation of a decontextualised list of cohesive devices misleads students who do not realise that the connectors grouped in one category often express different logical relationships that cannot be understood without a context. Moreover, connectors in the same category can have different grammatical functions. She argues that cohesive devices should be categorised on the basis of their grammatical function. Indeed, it is important that learners understand the difference between the syntactic features of co-ordinating, subordinating and adverbial connectors before they learn to differentiate semantically the connectors with the same grammatical function. For practice purposes, Zamel suggests completion, sentence combining, and sentence unscrambling exercises that illustrate the individual meaning as well as semantic and grammatical restrictions of connectors.

An important characteristic of connectors with regard to their grammatical function is the position that they can occupy in a sentence. Decontextualised lists of connectors fail to provide this information. Learners may not realise thus that whereas one particular connector can appear in sentence initial, middle or final position, a different one is restricted to a certain position (cf. Quirk et al. 1985:634; Biber et al. 1999:890–891). Nor can they observe the punctuation rules that vary according to the type of connector and the position it occupies in the sentence.

Considering the issue of connectors from a methodological point of view, Crewe (1990) points out that their overuse and misuse is due to mechanical exercises and the fact that they are not treated as “higher-level discourse units which organise chunks of text in relation to the direction of the argument”(p. 316) but as stylistic devices. He contends that the problem can be ameliorated through limiting the number of connectors used by students to a few under-

standable ones; through explicitation, i.e. the use of the phrasal equivalent of a one word conjunction; and through laying emphasis primarily on the logic of the composition.

In order to promote cohesion in language learner texts, McCarthy recommends both traditional problem-solving methods (e.g. the insertion of connectors into a text) and, notably, those made available by the process approach to writing through redrafting (1991: 156). The teacher can indicate in the feedback given on consecutive drafts if it is necessary to add, remove or change connectors.

The approaches recommended for the teaching of connectors have one key feature in common: directly or indirectly they all stress the importance of context from the very first stage in the acquisition process of connective devices, when the learner can infer from contextual clues the syntactic and semantic forces that produce a relation, to the final production stage, when the learner manipulates written discourse through an informed choice of connectors.

3. Research questions

The term *connector* is generally used to refer to co-ordinating and subordinating conjunctions and adverbial connectors (for extensive discussions on the items that fall into the category of connectors see Biber et al. 1999; Halliday & Hasan 1976, or Quirk et al. 1985). The essay examination scripts written by Hungarian students that have been marked over several semesters have given evidence of erroneous co-ordinating and subordinating conjunction uses only in an insignificant number of instances. The students have been observed to have significantly more problems, however, with the appropriate use of adverbial connectors. The present small-scale study draws on both Corpus Linguistics and Discourse Analysis and proposes to investigate the use of adverbial connectors in high-rated argumentative essays written by second and third year students enrolled in the MA in English Language and Literature course at Eötvös University, Budapest. The investigation aims to (1) make an inventory of the adverbial connectors used in the texts written by the university students, to (2) analyse their use and to (3) compare the results with those reported on the basis of a similar analysis conducted on a native speaker student corpus. It seeks to answer the following questions:

1. What type of adverbial connectors do the texts contain?

2. What is the distribution of the adverbial connectors within individual essays?
3. Which type of semantic relationship is most commonly marked with adverbial connectors?
4. What linguistic units do adverbial connectors span?
5. In what position do adverbial connectors occur most frequently in the texts?
6. Are the adverbial connectors employed stylistically appropriate?
7. How does the non-native speaker use of adverbial connectors compare with the native speaker use?

It is expected that the findings of the study will provide informative insights for writing pedagogy in general, and especially that with their help the problems particular to the connector use of advanced Hungarian EFL writers can be effectively addressed.

4. Methods

4.1 The participants and the corpus

The corpus built for the analysis consists of essays selected from 93 argumentative essays written in an examination by second and third year students attending a five-year degree course in English. The participants were all Hungarian native speakers who live in Hungary and are aged between 20 and 24. They speak advanced English and learn all their core subjects in English. In the first two years of their studies they attended three compulsory writing classes employing the process approach: one focusing on basic genres (narration, description, argumentation), one on academic writing, and one on advanced argumentative writing.

The essays were written on the basis of a short printed prompt in an essay examination. The examinees were allowed to use a monolingual dictionary as a reference tool and were expected to write formal English texts of approximately 500 words. Only those students' essays were incorporated into the corpus who consented that their texts can be used for research purposes, that is 82.79% of the tested population. The design criteria of the corpus are based on Granger (1998). Table 1 summarises the main features of the corpus.

Each essay was marked by three independent graders on the basis of standardised marking criteria and the final scores were calculated on the basis of

Table 1. The features of the Hungarian Corpus

Language		Learner	
Medium	written	Age	20–24
Genre	argumentative essay	Mother tongue	Hungarian
Topic	free education	Region (country of provenance)	Hungary
Technicality	formal (semi-academic text ²)	Proficiency level	advanced
Task setting	timed examination (use of monolingual English dictionary allowed)	Learning context	EFL

the three graders' scores (inter rater reliability = 0.87). The maximum number of points an essay could be awarded was 24. For the investigation only those essays were selected that scored 18 points and above, which reduced the number of eligible essays to 27. As a last step in the selection process, from these essays only those 21 were chosen that scored the maximum number of points on the *cohesion/coherence* criterion (i.e. 3). The essays are of an average length of 490 words, the full corpus contains 9,969 words.

4.2 Terminology and analytical decisions

In this study the term adverbial connector is used to refer to adverbials that Quirk et al. call conjuncts (1985:631). On the basis of their semantic functions Quirk et al. distinguish between seven types of adverbial connectors: *listing*, *summative*, *appositive*, *resultive*, *inferential*, *contrastive*, and *transitional*. Several of these categories have subdivisions (Figure 1).

The three types of adverbial connectors to be collected from the corpus for analysis were simple adverbs (*first*, *next*, *further*), adverbs that Halliday and Hasan (1976:231) call compound adverbs (e.g. *subsequently*, *therefore*, *nevertheless*), and adverbial prepositional phrases (*as a result*, *in addition*). Any adverbial connector in any position within a sentence was considered a cohesive device and included in the analysis.

The terms overuse and underuse are defined in this paper on the basis of native English use of connectors. Thus, overuse describes instances when the non-native writers use significantly more connectors than the native writers, whereas underuse refers to those cases when the non-native writers use significantly fewer connectors than the native writers. The present study considers native speaker usage of connectors as a reasonable target and interprets the findings accordingly.

1. Listing:	– enumerative	
	– additive:	– equative
		– reinforcing
2. Summative		
3. Appositional		
4. Resultive		
5. Inferential		
6. Contrastive:	– reformulatory	
	– replacive	
	– antithetic	
	– concessive	
7. Transitional:	– discorsal	
	– temporal	

Figure 1. The seven types of adverbial connectors distinguished by Quirk et al. (1985)

In order to produce more than a descriptive account of the use of adverbial connectors in high rated argumentative essays written by Hungarian students of English, the findings that result from the analysis of the Hungarian Corpus are compared whenever there is a possibility to those reported in a similar study conducted by Altenberg and Tapper (1998). Altenberg and Tapper compiled a corpus of essays (app. 50,000 words) written by native English speakers (henceforth referred to as the Native Corpus). The essays are argumentative and about one thousand words long. The authors give the number of tokens per 10,000 words, which makes possible a comparison between the Native Corpus and the non-native Hungarian Corpus.

Altenberg and Tapper investigated the same type of adverbial connectors as the present study does with the exception of one category they labelled *corroborative* (p. 84). They include in this category items such as *actually*, *in fact*, *of course*, or *indeed*, which they claim to have a cohesive role. Since Altenberg and Tapper include this category into their analysis and report the ratio of corroborative items together with adverbial connectors found in the Native Corpus, the Hungarian Corpus was also searched for corroboratives. No corroborative items were found, the corpus contains one instance of *indeed*, whose function is to add emphasis, it neither introduces a new point, nor shifts the focus of the discussion.

So after all, there are indeed suitable and well-paid jobs for intellectuals holding a diploma in humanities. (Hun E67)

Therefore, corroboratives will not be discussed with the exception of one comment in Section 5.6.

5. Results and discussion

5.1 Adverbial connectors in the Hungarian Corpus

The number of adverbial connectors the Hungarian students used in their essays is given in Table 2, and for comparison the rate per 10,000 words of the Native Corpus is also shown. From this we can see that Hungarian students used slightly fewer adverbial connector types more frequently than the native speaker writers. The difference may reflect the emphasis that the writing courses the Hungarian students attend place on the responsibility of the writer for the clarity of their academic texts and on the importance of coherence and cohesion.

The top ten adverbial connectors are listed in Table 3. Only three of the ten most frequently used adverbials are identical in the two corpora. The Hungarian students used the same contrastive connector (*however*) as the native speakers. The high frequency of the two resultive adverbial connectors *therefore* and *thus* in the two corpora can be explained with the text type. The syllogistic sequence *premise 1 + premise 2 therefore conclusion* is typical of argumentative texts and both groups of students most probably applied this logic in their reasoning.

The difference in the rank order of the two sets of connectors is disrupted by the reinforcing listing adverbial connector *also* that the majority of Hungarian writers used repeatedly in their texts. From the fourth item in the list the two ranks are completely dissimilar, two noteworthy differences being that the Hungarian students employed a larger number of compound adverbs, and very few phrasal equivalents of adverbial connectors. It appears that the writers' main concern was to highlight the fact that their texts list and contrast ideas and to emphasise their conclusions.

Table 2. Comparative frequencies of adverbial connectors in the Native Corpus and the Hungarian Corpus

	Native Corpus	Hungarian Corpus
Tokens/10,000 words	95	202
Types	48	42

Table 3. The ten most frequent adverbial connectors in the Native Corpus and the Hungarian Corpus

Native	n per 10,000	Non-native	n per 10,000
<i>however</i>	25.4	<i>however</i>	30
<i>therefore</i>	11.4	<i>also</i>	26
<i>thus</i>	8.7	<i>therefore</i>	20
<i>for example</i>	7.5	<i>thus</i>	13
<i>so</i>	6.9	<i>furthermore</i>	10
<i>of course</i>	4.6	<i>moreover</i>	8
<i>in fact</i>	3.3	<i>secondly</i>	8
<i>that is</i>	2.9	<i>though</i>	7
<i>yet</i>	2.7	<i>in addition</i>	6
<i>indeed</i>	2.5	<i>first of all</i>	5

5.2 The distribution of adverbial connectors in the Hungarian Corpus

It is to be noted that the distribution of the adverbial connectors in the individual essays shows that there are significant differences between the Hungarian texts: the number of adverbial connectors used in a text ranges from 1 to 15 (Table 4 shows the distribution for frequencies of three or more). The texts in the Native Corpus, which are twice the length of the Hungarian texts, were reported to contain 1 to 25 adverbial connectors. Taking the difference in length into account, those Hungarian texts that contain 10 to 15 adverbial connectors (i.e. 10–15 tokens/app. 500 words) can be said to contain nearly the same number as the texts produced by the native speakers who used more than 20 adverbial connectors (i.e. 20–25 tokens/app. 1000 words). However, the types of adverbial connectors identified show that the majority of the texts that have the largest number of adverbial connectors have a low type count, which shows frequent repetition.

Instead of using a set of different adverbial connectors, some Hungarian students used a few “pet” adverbial connectors repeatedly. Table 5 shows the distribution of the nine most frequent adverbial connectors (the tenth most frequent belongs to a small subset of connectors that all contain the element *first* and will be discussed later). To mention only the most outstanding cases of “pet” adverbials, *however*, *therefore*, and *thus* are connectors which appear five times in texts 67, 50, and 19. These instances illustrate an overuse of adverbial connectors that has been revealed through the analysis of individual texts.

Apart from the ten most frequent adverbial connectors, Hungarian writers also have a marked preference for those connectors that denote superstructural

Table 4. The distribution of adverbial connectors in the Hungarian Corpus

Text No	Token	Type	Ratio
19	15	10	1.5
50	15	9	1.6
56	15	11	1.3
70	15	11	1.3
55	13	11	1.1
46	12	8	1.5
52	13	9	1.4
64	11	10	1.1
69	11	9	1.2
20	9	9	1
21	9	7	1.2
27	9	7	1.2
67	9	5	1.8
63	8	6	1.3
02	7	6	1.1
14	6	6	1
28	6	6	1
47	6	5	1.2
13	5	5	1
75	5	3	1.6
07	3	3	1

units within a text. Thought units in written discourse are organised into paragraphs that fulfil specific functions in the introduction, body or concluding section of a text. There are a set of adverbial connectors, such as *first*, *second*, and *third*, that are used to indicate the sequence in which the thought units follow. Hungarian writers use quite a large number of these adverbial connectors to mark the sequence in which they organise their ideas in the body of the essay (Table 6). There are fourteen texts in the corpus that feature a listing connector containing the word *first*, ten texts contain *second* or *secondly*, and six texts have the concluding connectors *lastly*, and *finally*. The writers also prefer resultive adverbial connectors such as *therefore*, *thus*, or *consequently* to indicate a deductive logical relationship, and summative adverbial connectors such as *in conclusion*, *to conclude*, or *to sum up* that serve to mark overtly the function of the concluding paragraph. These adverbs most probably appeal to students because of their unambiguous discourse organising qualities that are easy to understand. Nevertheless, their overuse results in a text that reads exceedingly structured and artificial.

Table 5. The distribution of the nine most frequent adverbial connectors in the Hungarian Corpus

Conj. adverb	n	Text																				
		02	07	13	14	19	20	21	27	28	46	47	50	52	55	56	63	64	67	69	70	75
<i>however</i>	30	2	1	1	–	2	–	–	3	1	3	3	–	2	1	2	–	1	5	1	1	1
<i>also</i>	26	1	–	–	–	1	1	2	–	–	3	1	–	3	2	2	2	2	–	3	2	1
<i>therefore</i>	20	–	–	1	–	–	–	–	–	1	1	1	5	3	1	1	1	–	–	1	2	2
<i>thus</i>	13	–	1	–	1	5	1	–	–	–	–	–	–	–	–	1	–	–	–	1	2	1
<i>furthermore</i>	10	–	–	1	1	–	–	1	1	–	–	–	–	–	1	1	2	–	–	–	2	–
<i>moreover</i>	8	–	–	–	1	1	–	–	1	–	–	–	1	–	–	2	1	1	–	–	–	–
<i>secondly</i>	8	–	–	–	1	–	–	1	–	1	–	–	1	1	1	–	–	1	1	–	–	–
<i>though</i>	7	–	–	–	–	–	–	2	–	–	–	–	3	–	–	–	–	–	–	1	1	–
<i>in addition</i>	6	–	–	–	–	–	–	–	1	–	–	–	–	1	2	–	–	–	1	–	1	–
Total	128	3	2	3	4	9	2	6	6	3	7	5	10	10	8	9	6	5	7	7	11	5

Table 6. The distribution of the most frequent listing adverbial connectors in the Hungarian Corpus

Adverbial connector	n
<i>firstly</i>	4
<i>first</i>	1
<i>first and foremost</i>	4
<i>first of all</i>	5
<i>second</i>	2
<i>secondly</i>	8
<i>third</i>	1
<i>in addition</i>	1
<i>furthermore</i>	4
<i>finally</i>	4
<i>lastly</i>	2
<i>in conclusion</i>	4
<i>in short</i>	2
<i>to conclude</i>	1
<i>to sum up</i>	3
Total	46

5.3 The most common types of semantic relationships in the Hungarian Corpus

Table 7 shows that the Hungarian students marked about eight times more listing relations than the native speaker writers. There are a high number of enumerative (*first, second, third*) and additive (*also, furthermore, moreover*) adverbial connectors in their texts. The second most frequent type of relationship marked in both corpora is the resultive, but the Hungarian students use fewer resultive adverbial connectors than the natives. The frequency of contrastive relations is high in both corpora, yet except for a similarly low preference for the overt marking of transitional relations, the distribution of the other types of relations is markedly different.

The writing of Hungarian students is characterised by the presentation of a highly structured contrastive set of ideas arranged cumulatively. The listing relations suggest that their reasoning is constructed with the enumeration of arguments for a certain standpoint, and the contrastive relations indicate that the texts contain frequent references to the opposing standpoint. The high number of resultive adverbial connectors suggests that the underlying logic applied in the argumentative texts is deduction. While the native speakers seem to emphasise the cause-effect relations with the overt use of significantly more

Table 7. The semantic relationships marked in the Native Corpus and the Hungarian Corpus

Native	n/10,000	%	Non-native	n/10,000	%
contrastive	33.2	34.95	listing	92.0	45.54
resultive	28.0	29.47	resultive	44.0	21.78
listing	11.0	11.58	contrastive	40.0	19.80
appositive	10.5	11.05	summative	13.0	6.43
corroborative	10.3	10.84	appositive	9.0	4.45
summative	2.0	2.11	inferential	3.0	1.48
transitional	0	0	transitional	1.0	0.49
inferential	–	–	corroborative	0	0
Total	95	100		202	100

resultive connecting items, the Hungarian students lay much more emphasis on highlighting that it is the quantity, the sum total of their arguments that decisively justifies their standpoint.

The following sample paragraph from the Hungarian Corpus serves as an illustration for this type of text development:

Politicians and experts arguing for the right to higher education usually do so on theoretical grounds: they claim that there exists a fundamental right to education for all human beings. *On the one hand, however*, primary and secondary school education would still remain accessible to all citizens; *hence*, the fundamental right to education would be satisfied. *On the other hand*, potential students who have not mastered the skills and synthesized the knowledge essential for higher education should not have the right to continue their studies there. *Furthermore*, this step would ensure that quality work takes place at universities. *In short*, although there may be a right to education, it does not follow that it also includes tertiary tuition as well. (Hun E70)

listing/enumerative, &
contrastive/concessive
resultive
listing/enumerative

listing/reinforcing
summative

5.4 The span of the relations marked by adverbial connectors

As Quirk et al. (1985:632) note, connectors establish a relation between linguistic units that can be the components of a phrase, a sentence, a paragraph or an even longer stretch of text. The adverbial connectors used by the Hungarian writers can be divided in two broad categories on the basis of the linguistic

units they span: those that mark the superstructural subdivisions of the texts; and those that are used within the superstructural elements.

The first category consists of listing adverbial connectors, both of the enumerative and additive types, and of summative adverbial connectors. These connectors are used as the first element of the body and concluding paragraphs and signal the relationship between the paragraphs in the essays. The enumerative adverbial connectors (*first, second, third*) included in the first sentence of body paragraphs indicate overtly the sequence between consecutive paragraphs. They show furthermore the place of an individual paragraph in the overall structure of the essay. There are some instances where the standard pattern (*first, second, third; firstly, secondly, thirdly*) contains reinforcing adverbial connectors like *in addition*, or *furthermore* and concludes with enumeratives like *finally*, or *lastly*. Whereas the reinforcing adverbial connectors are also frequently used within paragraphs, there is only one text in the Hungarian Corpus in which enumerative connectors organize ideas within a paragraph:

What is more, this system would affect students in a positive way: *firstly*, they would become more conscious and more involved in their field of study, because they would consider it as something they purchased and have financial interest in; *secondly*, students would not lengthen the time of their studies unnecessarily because of the threat of paying higher fees. (Hun E19)

The summative adverbial connectors (*in conclusion, to sum up*) are used to introduce concluding paragraphs. Like their listing counterparts, they formulate overtly the rhetorical function of a paragraph type whose role, in the case of the conclusion, is to bring the essay to an end usually through the synthesis of the main ideas discussed in it. For this reason, these adverbial connectors establish a relation with the title, introduction and all the body paragraphs of an essay.

Since their function is straightforward, students can easily master the use of listing adverbial connectors and they overuse them probably for the same reason. As can be seen in Table 7, nearly half of the adverbial connectors that Hungarian writers used come from the listing category (45.54%).

The adverbial connectors in the second broad category establish relations within paragraphs and smaller linguistic units. This category comprises the reinforcing adverbs that also function as superstructural markers and have been discussed above as well as adverbs from the other types of classes (Table 8).

There are several instances in the Hungarian Corpus in which adverbial connectors mark a relationship that connects more complex linguistic units than two sentences. In the following excerpt the resultive adverbial connector

Table 8. The span of relationships marked by adverbial connectors

Connection type	Adverbial connectors
components of a clause	<i>as a result, also, for example, thus</i>
two clauses	<i>consequently, furthermore, hence, moreover, therefore, thus</i>
two sentences	<i>again, also, besides, furthermore, however, in this case, moreover, therefore, thus</i>
parts of a paragraph	<i>consequently, furthermore, however, in either case, meanwhile, moreover, nevertheless, on the one hand/on the other hand, similarly, so, therefore, thus, to start with</i>
two paragraphs	<i>besides, furthermore, however, moreover</i>

accordingly indicates that the writer arrives at the conclusion on the basis of the information provided in the preceding sentences.

In this particular case, a parallel between a college degree and money is too apparent not to be recognized. Money is valuable because people have to work for it, consequently, no one can have as much money as he wants. The value of a degree lies in that there are only a few people in every country of the world who have one. *Accordingly*, if everyone were granted a degree, then it would become nothing more than a mere sheet of paper that one can hang on the wall. (Hun E64)

The writer aims to convince the reader with an argument by analogy: if the reader summarises and accepts what has been presented in the first part of the paragraph, he or she should also accept the conclusion.

In a few cases the relationship marked by the contrastive adverbial connector *however* spans more than one paragraph. This occurs in the case of refutations that are constructed so that they refer back to the preceding arguments presented in the essay. The most frequent position for a refutation in the studied essays is the last body paragraph. There are a few instances in which the connection between two paragraphs is marked by listing adverbial connectors (e.g. *furthermore, therefore*).

The Hungarian writers whose texts were analysed use adverbial connectors to mark relationships between a variety of linguistic units. They overuse those adverbial connectors that mark structural features in a text and most often indicate the relationship between consecutive sentences or sections of a paragraph. There are, however, significant differences between the individual writers in the practice of using adverbial connectors to mark relationships

overtly, as well as in the type of adverbial connectors used. On the basis of the overall scores awarded for the individual texts, those students' essays were rated highly who marked such relationships overtly with listing and summative adverbial connectors that span larger stretches of text than a sentence.

5.5 The position of adverbial connectors in the texts produced by Hungarian writers

Quirk et al. (1985:643) state that the normal position of most adverbial connectors is initial, and some are actually restricted to this position; some adverbial connectors whose meaning cannot be misinterpreted occur in medial position; and a few adverbial connectors appear in final position (Table 9). The predictions of Quirk et al. are born out by the findings of Biber et al. (1999), who state that "...in academic prose, the most common position for linking adverbials is initial" (p. 890) and that "medial positions account for the second highest proportion of occurrences; final position is rare" (p. 891).

The Hungarian students follow closely the distribution of adverbial connectors as described by Quirk et al. (1985) (Table 10). The most frequent position for adverbial connectors in the Hungarian Corpus is the initial position followed by about half as many instances of adverbial connectors used in medial position. The three instances of sentence final adverbial connectors are the appositive *for example* (n=1), and *for instance* (n=2). Students are taught the features of formal and informal language use in writing classes and it can be argued on the basis of the low number of final position adverbs found in their texts that they are aware that this position is not characteristic of formal written texts. In a few instances, the writers chose to link two clauses with the adverbial connectors *consequently* (n=1), *thus* (n=2), *moreover* (n=1), and *furthermore* (n=2). Of the six instances the punctuation of the sentence is incorrect in only one case, in all the other cases the adverbial connector linking two clauses is preceded by a semicolon and followed by a comma.

Money is valuable because people have to work for it, consequently, no one can have as much money as he wants. (Hun E64)

The single instance of a clause final adverbial connector is *though*. The low number of such instances (n=1) is again most probably indicative that students are aware that the final position is more typical of spoken discourse and thus when they use *though* they place it in medial position (n=6).

The frequencies of the positions of adverbial connectors compares surprisingly well with the Native Corpus, as shown in Table 11.

Table 9. Typical positions of the adverbial connectors in English

Position	Adverbial connectors
Restricted to initial position	<i>again, besides, yet, still, what is more, so, else, hence</i>
Frequent in medial position	<i>however, nevertheless, in other words, on the contrary</i>
Frequent in final position	<i>in other words, anyhow, anyway, though</i>

Table 10. The positions of adverbial connectors in the Hungarian Corpus according to type

Adverbial connector type	SI	M	SF	CI	CF
Listing	59	30	0	3	0
Resultive	27	14	0	3	0
Contrastive	25	14	0	0	1
Summative	13	0	0	0	0
Appositive	2	4	3	0	0
Inferential	3	0	0	0	0
Transitional	0	1	0	0	0
Total (n=202)	129	63	3	6	1
Total %	63.86	31.18	1.48	2.97	0.99

SI = sentence initial; M = sentence medial; SF = sentence final; CI = clause initial; CF = clause final

Table 11. The position of adverbial connectors in the Native Corpus and the Hungarian Corpus

Native Corpus			Hungarian Corpus		
	n	%		n	%
Sentence initial	53	66.25	Sentence/clause initial	135	66.83
Sentence medial	25	31.25	Sentence medial	64	31.68
Sentence final	2	2.5	Sentence/clause final	3	1.48
Total	80	100	Total	202	100

It seems that when Hungarian writers opt to mark a relationship between two linguistic units overtly, they put the adverbial connectors in the same positions as the native writers. These results may also be indicative of the fact that students do observe and accurately reproduce linguistic phenomena that they encounter frequently (e.g. by reading) and that have a regular pattern.

5.6 Register awareness in the use of adverbial connectors in the Hungarian Corpus

Quirk et al. (1985:634) add register information to the adverbial connectors they include in their comprehensive categorisation. For example, they mark *furthermore* or *thus* as formal and *for a start* or *so* as informal. The instances of adverbial connector use in the corpus are in the majority of cases formal. The very few instances of informal use are in the case of the resultive adverbial connector *so* (n=2), and the sentence final use of the appositive *for example/instance* (n=3), which is a position more frequent in spoken discourse. The analysis of the Hungarian texts shows that the students are aware of the register characteristics of these linking devices. This explains the absence of corroboratives (*of course, actually, in fact, surely*), which are characteristic of spoken, often informal, discourse.

6. Conclusion

As a result of the analysis of the use of adverbial connectors in a corpus consisting of argumentative essays written by Hungarian students, it can be stated that some noteworthy differences and similarities have been revealed. The dissimilarities are of typological and distributional nature. The Hungarian writers' texts have been shown to feature more adverbial connector tokens than the native students'; albeit, it was also revealed that the Hungarian writers use nearly as many adverbial connector types as the natives. A possible explanation for this comparative overuse of adverbial connectors is that the Hungarian language, similarly to the Finnish language, does not require the overt marking of relations between linguistic units of the text. As a result of this, the teachers of academic English writing generally put more emphasis on the explicit teaching of adverbial connectors to Hungarian students.

Nevertheless, the overuse of adverbial connectors is quite as prevalent as their underuse as can be seen when the individual essays are ranked on the basis of the number of adverbial connector tokens they contain. A cline becomes visible with essays containing only a small number of adverbial connectors on one end and with others that have a large number of adverbial connectors on the other end. Both seem to be problematic cases, yet the ones with more adverbial connectors display more irregularities. For example, some writers develop an affection for a particular type of adverbial connector and use it disturbingly frequently even in short texts. There is a low frequency of phrasal equivalents

of adverbial connectors in spite of the fact that these equivalents are often more transparent lexical items and easier to use. The texts contain a very large number of listing adverbial connectors that mark the superstructure of the essay. These are most probably intended to increase the coherence and cohesion of the text, there are even instances when an attempt is made to break the regular pattern and with it its monotony and cliché-like effect, but the result is often a text that is artificial and in which the coherent flow of ideas is rather hindered by the adverbial connectors.

The presence of a large number of resultive and contrastive adverbial connectors in the texts written by Hungarian students cannot be considered surprising. These adverbial connectors are typical of argumentative discourse (rhetoric textbooks refer to them as premise and conclusion indicators), they signify the typical reasoning processes that these texts encapsulate. Their presence demonstrates that the writers are familiar with the lexical realisations of the rhetorical features of argumentation and that in their texts they are probably engaged in the reasoning processes that underlie these rhetorical features.

What is commendable in the use of adverbial connectors in the Hungarian writers' texts is the awareness of the characteristic positions of adverbial connectors in English. Similarly, the writers' register awareness appears to be appropriate for the production of formal discourse: there were hardly any instances of informal connector use identified in the essays with either a low or a high number of adverbial connectors.

The insights gained with this study can be improved with further analyses. A comparative study of texts written by expert academic writers would probably reveal whether the overuse of adverbial connectors shown in the comparison with the native student essays should indeed be attributed only to the stressed emphasis put on the teaching of these devices in writing classes, or also to the large number of academic texts written by expert native writers which the Hungarian students read in the course of their studies. Further insights could be gained with the in-depth analyses of those texts that contain few adverbial connectors, as their absence may be not only the sign of underuse of overt connectors but also of the fact that the coherent flow of ideas has been ensured with the skilful use of other cohesive devices.

7. Implications for teaching

The process of acquisition of adverbial connectors is most effective if it is both teacher and student controlled. However, the ratio of the control is paramount

and with the help of data gained from corpora it can be modified so that the students' role in the learning process becomes increasingly substantial. Whereas there are some aspects of the process that the teacher cannot hand over to the students, the students can discover the majority of the characteristics of adverbial connectors for themselves.

The teacher should supply a reliable and thorough introduction to adverbial connectors. Information on the variety of adverbial connectors and their frequency in various spoken and written text types can be given on the basis of such sources (e.g. Biber et al. 1999: 875–892) that rely on corpus evidence: it is Corpus Linguistics studies that provide the most reliable empirical evidence on the use of adverbial connectors. The teacher can furthermore give valuable feedback concerning the number of adverbial connectors used in student texts as well as make explicit, relevant and therefore effective comments based on particular instances taken from student texts concerning the question of when to use and when not to use adverbial connectors.

Given that the students have direct access to corpora either through the Internet or on CD-ROMs, or to KWIC (key word in context) concordances featuring adverbial connectors, they can systematically collect with the help of paradigmatic presentation of repeated patterns such information about adverbial connectors as their meaning and the cognitive relation they express, their grammatical function, their genre sensitivity, the linguistic units they span, and the various forms the same adverbial connector can have. Furthermore, the observation of repeated occurrences of an adverbial connector furnishes information on the positions it can occupy and the punctuation typical to these positions.

Such a data-driven approach to learning about adverbial connectors can be effective in improving the types of deficiencies observed in the corpus built from the texts of Hungarian writers. Being a heuristic and empirical approach that allows the students to control the focus of their investigation, corpus evidence based tasks are reliable, interesting, motivating and more informative than a dictionary entry or a random list of adverbial connectors presented in a course book. The knowledge students gain in such a practical manner can lead to the disappearance of “pet” adverbial connectors through the informed exploitation of the variety of adverbial connectors available. It can, furthermore, directly improve the coherence of student texts through the informed exploitation of the functions of adverbial connectors.

It is to be noted, however, that the presence, the frequency, and the distribution of connectors in a particular text cannot be considered the ultimate indicator of text quality. A text that contains an acceptable number of stylistic-

cally appropriate connectors applied in the right positions can still be devoid of either logic or content.

7.1 A concordance-based classroom activity on adverbial connectors

This activity is suitable for intermediate level ESL or EFL students and above. The students are provided with lists of connectors categorised on the basis of their grammatical functions. After the grammatical functions are explained to the students, they form small groups and either access a corpus available on the Internet or receive KWIC concordance sheets with the one word and phrasal equivalents of the same adverbial connector. Students then are told that with the help of the information obtained from their Internet searches or what they find on their concordance sheets they have to work out

- the function (& meaning) of the adverbial connector
- the position(s) of the adverbial connector in the sentence
- the punctuation marks used with the adverbial connector relative to the position it occupies
- the type of text the sentences come from (This is the easiest in the case of a corpus accessed over the World Wide Web. For example, the results page returned by the search engine of the British National Corpus provides a link to the description of the source of every single displayed entry.)

and to note down any regularity that they have observed. Each group then summarises their findings for the others and thus the observed regularities can be collected by a note taker at the board.

Notes

1. In this paper, the term ‘connector’ will be used with the exception of specific references to particular types of connectors as a general term that refers to co-ordinating conjunctions, subordinating conjunctions, and adverbial connectors.
2. The texts share the key features of academic texts in that they “present a reasoned argument, evaluate evidence, and draw appropriate conclusions” (Jordan 1997: 13), but do not contain any references or quotes and do not feature sections typical of scientific research articles (i.e. review of the literature, methods, limitations).

References

- Altenberg, B. & M. Tapper (1998). The use of adverbial connectors in advanced Swedish learner's written English. In S. Granger (Ed.), *Learner English on Computer* (pp. 80–93). Harlow: Addison Wesley Longman.
- Biber, D., S. Johansson, G. Leech, S. Conrad, & E. Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.
- Conrad, S. (2004). Corpus linguistics, language variation, and language teaching. In J. McH. Sinclair (Ed.), *How to use Corpora in Language Teaching* [Studies in Corpus Linguistics 12] (pp. 67–85). Amsterdam: John Benjamins.
- Cook, G. (1989). *Discourse*. Oxford: Oxford University Press.
- Crewe, W. J. (1990). The illogic of logical connectors. *ELT Journal*, 44, 316–325.
- Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer* (pp. 3–18). Harlow: Addison Wesley Longman.
- Halliday, M. A. K. & R. Hasan (1976). *Cohesion in English*. London: Longman.
- Hatch, E. (1992). *Discourse and Language Education*. Cambridge: Cambridge University Press.
- Jordan, R. R. (1997). *English for Academic Purposes: A Guide and Resource Book for Teachers*. Cambridge: Cambridge University Press.
- Mauranen, A. (1993). *Cultural Difference in Academic Rhetoric: A Textlinguistic Study*. Frankfurt: Peter Lang.
- McCarthy, M. (1991). *Discourse Analysis for Language Teachers*. Cambridge: Cambridge University Press.
- McCarthy, M. & R. Carter (1994). *Language as Discourse: Perspectives for Language Teaching*. London: Longman.
- Oxford Advanced Learner's Dictionary of Current English* (4th ed.). 1989. Oxford: Oxford University Press.
- Quirk, R., S. Greenbaum, G. Leech, & J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- Wikborg, E. & L. Björk (1989). *Sammanhang i text. En empirisk undersökning och skrivpedagogiska konsekvenser.* [Coherence in text. An empirical study with pedagogical consequences for composition teaching.] Uppsala: Hallgren and Fallgren.
- Zamel, V. (1983). Teaching those missing links in writing. *ELT Journal*, 37, 22–29.

Textbooks

A corpus-driven approach to modal auxiliaries and their didactics

Ute Römer

University of Hanover

The paper presents an example of the indirect use of corpora in language pedagogy. It centres on a comparative analysis of modal auxiliaries, their distribution, meanings, and contexts, in spoken British English corpus data and in selected texts from EFL textbooks. A focus lies on the differences observed between authentic English as used in natural communicative situations and the kind of synthetic English that pupils are often confronted with in the classroom. It is argued that, if taken seriously, corpus evidence can contribute to an improvement of teaching materials and that it is essential, especially in pedagogical contexts, to pay more attention to frequent phenomena and typical patterns of used language.

1. Introduction

Modal auxiliaries constitute one of the grammatical problem areas in teaching English as a foreign language to German learners and probably also to learners of other nationalities. This observation forms the basis of the research project reported on in this paper, which deals with the use of modals in spoken English and in English language teaching.¹

The purpose of this paper is to summarise the results of an investigation which centred on a corpus-driven analysis of the nine central modal verbs as listed in Quirk et al. (1985:137): *can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, and *must*, plus the modal *ought to* in contemporary spoken English and in one of the major German textbook series used in the EFL classroom.² The leading questions were: “Is the English taught at German schools identical to the English which is used by native speakers?” and “How extensively does the grammar of ‘school’ English differ from authentic spoken English?”

2. Modals in spoken British English (BNC analysis)

Trying to find answers to the questions above and starting from the assumption that pupils should learn a type of English which is really used and understood by native speakers nowadays, first of all data from the 10-million-word spoken part of the British National Corpus (BNC) was collected and analysed.³ The results of this corpus analysis were meant to show how often, in which contexts, and in which meanings the different polysemous modals are used in spoken British English. One reason for choosing exclusively spoken material was the fact that modal auxiliaries occur more frequently in spoken than in written English (cf. Quirk et al. 1985:136). However, the main reason was the pre-eminence of spoken language in English lessons as demanded in the *Richtlinien* for teaching English as a foreign language in Germany.⁴

Using SARA's (the BNC concordance program's) part-of-speech (POS) query option, queries on the different forms (i.e. positive, full negative, and contracted negative) of the ten modal verbs mentioned above were carried out. The query builder allows the researcher to combine different types of corpus searches, such as POS queries and SGML queries. In the present study an SGML query had to be used to make sure that exclusively spoken texts were searched (cf. Aston & Burnard 1998:100–108). For the SGML element <CATREF> (category reference) the attribute *SPOKEN_TYPE* with its values *dialogue* and *monologue* was selected. For each verb form a random set of 200 concordance lines was downloaded for further manual analysis. The “random set” box is one of the options available in SARA's “download hits” window. It is also possible to save the initial *n* solutions, all the solutions found, or only one solution per text (cf. Aston & Burnard 1998:66–67). The main reason for choosing the random set option here was to achieve a maximum distribution of downloaded concordance lines over all the texts in the spoken part of the BNC.

2.1 Frequency analysis

With the concordancer SARA it was also possible to get frequency information about the verb forms in question. Frequencies can be very important as they show us which words or structures are central in a language. Thus they can help with decisions about what to include in teaching materials and what not. On the basis of frequency data it is possible to see which modals are the most important ones and should thus be dealt with first in EFL teaching. Figure 1 shows the frequency distribution of the central modal auxiliaries (plus *ought to*) including negative forms in the spoken part of the BNC. As can be seen in

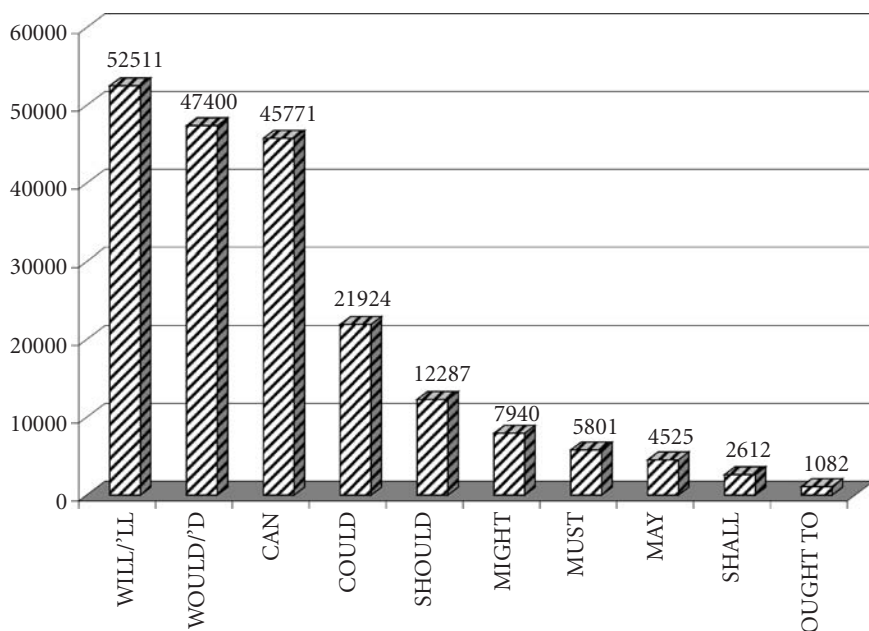


Figure 1. Frequency of modals in BNC spoken

this diagram, the three most frequent modals by far are *will/'ll*, *would/'d*, and *can* each with more than 45,000 occurrences in the spoken part of the BNC, followed by the modal *could* with 21,924 occurrences. The frequencies of the remaining modals are much lower, ranging from 12,287 occurrences (*should*) to 1,082 occurrences (*ought to*).

2.2 Different meanings analysis

Having collected the frequency data, the saved data sets (200 concordance lines for each verb form) were analysed manually with regard to different meanings and co-occurrences; i.e. the syntactic and semantic surroundings of the verbs in question were examined. The results of the different meanings approach were summarised in diagrams of the following kind (Figure 2), indicating for each modal the distribution of its different functions in spoken English.

For the modal auxiliary *can* three different meanings were found in the corpus material with the following frequency distribution: 36% ability, 31.5% possibility and 23.5% permission. The following are example sentences

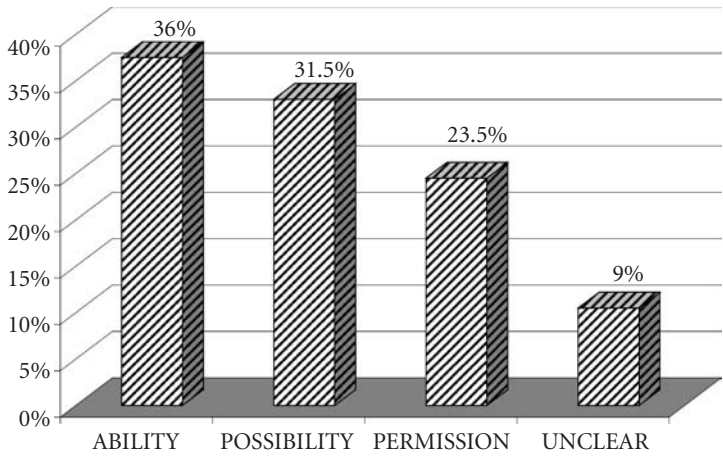


Figure 2. Different meanings of *can* (BNC spoken)

from the BNC in which *can* expresses an ability (1), a possibility (2), and a permission (3).

- (1) *and when it gets to the chasing teddy bears you've got to run as fast as you can, so you'd better move out of the way.* (BNC, KBW, 16264)
- (2) *Yeah, the whole sentence is, is a constituent itself er it and can be a constituent of larger sentences obviously.* (BNC, HE0, 187)
- (3) *Can I have an apple please?* (BNC, FLY, 355)

In addition, there was a considerable number of indeterminate or unclear cases, where it was impossible to make a decision about the function of the modal, either because the sentence was fragmentary or because there simply was not enough context, as for example in (4) and (5).

- (4) *I just offered to erm in Tech class can you?* (BNC, KPG, 5404)
- (5) *can it in those rooms with the dogs* (BNC, KE6, 10457)

The percentages of the different meanings distribution for *can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, *ought to* and *must* can be found in Table 1 below.

Table 1. Different meanings distribution of modals (BNC spoken)

	ability	possibility	permission	hypothet. meaning	prediction	volition	obligation/ advice	inference/ deduction	unclear
<i>can</i>	36%	31.5%	23.5%						9%
<i>could</i>	34%	41.5%	3.5%	14.5%					6.5%
<i>may</i>		83%	13%						4%
<i>might</i>		95%	3.5%						1.5%
<i>will</i>					87.5%	7.75%			4.75%
<i>would</i>				28.5%	50.5%	15.5%			5.5%
<i>shall</i>					31%	65%			4%
<i>should</i>				30%			62.5%		7.5%
<i>ought to</i>				16%			79%		5%
<i>must</i>							52%	39%	9%

2.3 Co-occurrence analysis

Some crucial observations could also be made in the analysis of co-occurrences of the modal verbs. Among the features examined were negations, and the occurrence of the different modals in questions, set phrases, if-clauses, and passive constructions. In the following, some of the most interesting findings are listed.

The highest percentages of negations were found with *can* (27.8%) and *could* (17.6%). Contracted forms (e.g. *can't*, 94.25%) are in all cases much more frequent than full forms (e.g. *cannot*, 5.75%). In his empirical study on modal verbs Mindt experienced a similar tendency but found much higher figures for *can* (40%) and *could* (32%) in negative contexts (Mindt 1995: 176). An explanation for these differences may lie in the different types of corpora used in the two studies. From Mindt's descriptions it does not become clear, however, what exactly his corpus consists of.

Another observation that could be made is that *shall* is used very frequently in questions (36.5% of the sentences examined), e.g. in

- (6) *Well shall I tell you what you were going to ask?* (BNC, HMP, 112)

This finding is in accordance with the results of Mindt's analysis where *shall* tops the list of modals in interrogative contexts (1995: 177).

In 85% of the analysed concordance lines the modal *shall* is accompanied by a first person subject ("I" or "we"):

- (7) *We shall see him says John, and we shall be like him.* (BNC, J8Y, 336)

Compared with the other modals *should* is more often (in 10% of the cases) found in passive constructions, as in the following BNC example:

- (8) *But but first of all I would like to say the officers of the agencies really should be congratulated on.* (BNC, J9D, 145)

May is quite frequent in if-clauses (19%). In this context the occurrence of the modal in the fixed expression “if I *may*” is worth mentioning. This phrase was found in 26.3% of all if-clauses with *may*, as for example in

- (9) *If I may come back Mr Chairman an and er express a view on behalf of Darcy.* (BNC, J42, 105)

3. Modals in EFL teaching (textbook analysis)

The second major part of this investigation is an analysis of the treatment of modal auxiliaries in *Learning English Green Line* (Vols. 1–6), a German textbook series widely used in the EFL classroom in grammar schools, and in the *Learning English Grundgrammatik*, an introductory grammar German pupils are supposed to work with and/or use as a reference grammar.⁵ As an electronic version of the textbook series was not available and the analysis thus had to be carried out manually, the six volumes of *Green Line* could not be analysed completely and were not regarded as a pedagogical corpus. Instead, several texts (32 altogether) from those textbook units which mentioned one or more modal auxiliaries in their grammar sections were examined thoroughly. The 32 texts were treated as a sample of EFL textbook language – the kind of language pupils are exposed to in the EFL classroom – enabling a comparison of textbook English with authentic language material collected from a general corpus.

The major aim of this textbook analysis was to find out whether the use of modals in *Green Line* was an accurate representation of the actual language use, i.e. of the occurrence of modal verbs in the spoken part of the BNC.⁶ Basically the same types of investigations were carried out as in the corpus analysis. A frequency count including an examination of the order in which the modals are introduced in *Green Line* (and which may reveal something about their prominence in grammar teaching) was followed by a different meanings and a co-occurrence analysis. In addition to several units from the textbooks (Vols. 1–6) all *Green Line* grammar sections and the *Learning English Grundgrammatik* were included in the analysis.

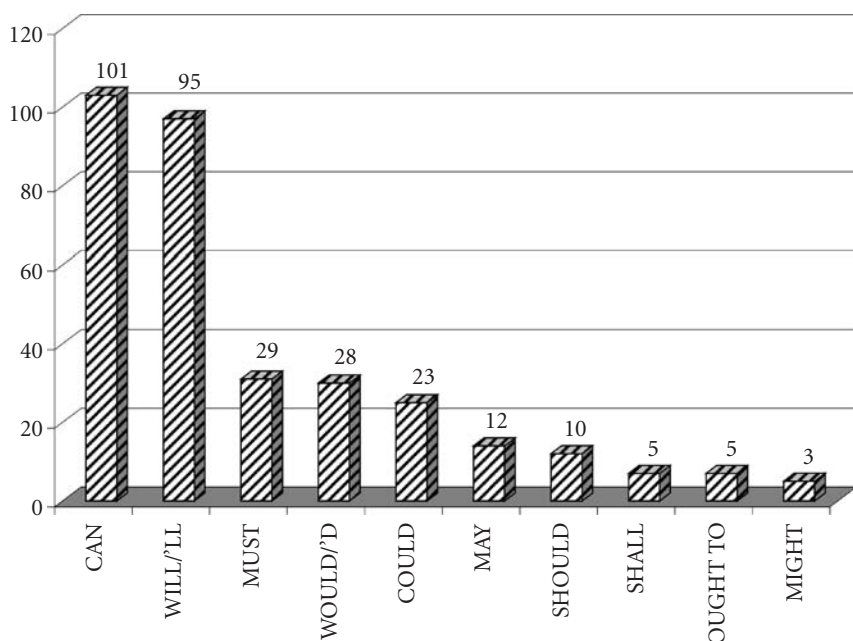


Figure 3. Frequency of modals in *Green Line*

3.1 Frequency analysis

Figure 3 shows the results of the frequency counts of the analysed texts from *Green Line* 1–6.

As we can see in this diagram, there is a huge frequency gap between *can* and *will/'ll* on the one hand and the other eight modals on the other hand. Thus in the textbook texts I found 101 occurrences of *can* and 95 occurrences of *will* and *'ll* but only between 3 and 29 instances of *could*, *would'd*, *may*, *might*, *shall*, *should*, *ought to*, and *must*.

3.2 Different meanings analysis

The results of the different meanings approach were collected in diagrams comparable to those that were used in the BNC data evaluation. In the diagram for *can* shown below (Figure 4) it becomes clear that in *Green Line* the modal is most frequently used to express an ability (52.5% of the cases). The meanings “possibility” and “permission” (24.7% and 22.8%) seem to be less important.

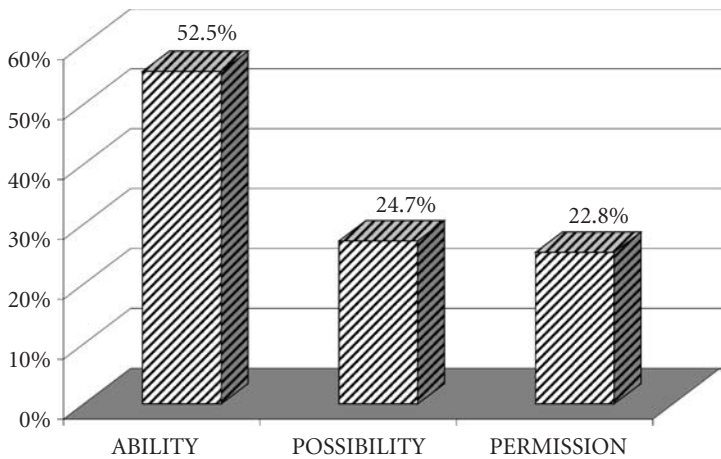


Figure 4. Different meanings of *can* (Green Line)

Analyses of that kind were also conducted for the other central modals and *ought to*. The percentages for all verbs under investigation have been collected in Table 2 below. Since we are dealing with invented examples here in which the sentence contexts are always constructed unambiguously and as there was always enough context available, there are no semantically indeterminate cases to be found in the textbook texts.

Table 2. Different meanings distribution of modals (Green Line)

	ability	possibility	permission	hypothet. meaning	prediction	volition	obligation/ advice	inference/ deduction	unclear
can	52.5%	24.7%	22.8%						
could	78.3%	13%		8.7%					
may		58.3%	41.7%						
might		100%							
will					82.1%	17.9%			
would				39.3%	28.6%	32.1%			
shall						100%			
should				20%			80%		
ought to				20%			80%		
must							93.1%	6.9%	

3.3 Co-occurrence analysis

Interesting findings from the co-occurrence examination of *can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, *must*, and *ought to* in *Green Line* texts are the following ones:

Very high incidence of negation is found with *can* (36.7%), *may* (33.3%), *could* (21.7%), and *must* (20.7%). On the other hand, there are no negative forms of *might*, *shall*, and *ought to*. The modal *shall* is found exclusively in questions (100%). *Could* (30.4%) and *may* (25%) also show rather high percentages of questions. *May* does not occur in if-clauses; e.g. there is not a single instance of the set phrase “if I *may*” in any of the textbook texts. Very high percentages of if-clauses are found with *would* (35.7%) and *might* (33.3%), and the modal *shall* is always used with a first person singular subject (100%).

4. Comparison: The use of modals in “real” English and in “school” English

The third part of this investigation consists of a comparison of the results of corpus analysis (BNC spoken) and textbook analysis (*Green Line*). Differences of the findings were pointed out again with regard to frequencies, different meanings and co-occurrences. This comparison made it clear that there are huge discrepancies between the use of modal auxiliaries in authentic English and in the English taught in German schools.

The frequency distribution of the modals in *Green Line*, for instance, differs quite a lot from the one found in the spoken part of the BNC. As we can see in Figure 5, the modals *will*/*'ll*, *can*, and *must* are overused in *Green Line* while there is an underuse of *would*/*'d*, *could*, *should*, and *might*. This underuse is especially significant in the case of *would*/*'d*. In the BNC the modal (including its contracted form *'d*) is the second most frequent one with 23.48%, whereas in the textbook series it only comes in fifth place (relative frequency: 9%).

More differences can be found if we compare the diagrams showing the different meanings distribution of each modal verb for the spoken part of the BNC and for the textbooks. For *can* and *could* expressing an ability for instance the percentages in *Green Line* (52.5% and 78.3%) are much higher than in the BNC (36% and 34%). In the sentences from the BNC *could* more frequently expresses a possibility (in 41.5% of the cases) than an ability. Concerning *may* we get a much higher share of the permission meaning in *Green Line* (41.7%) than in spoken English (13%), although the modal is mainly used to convey

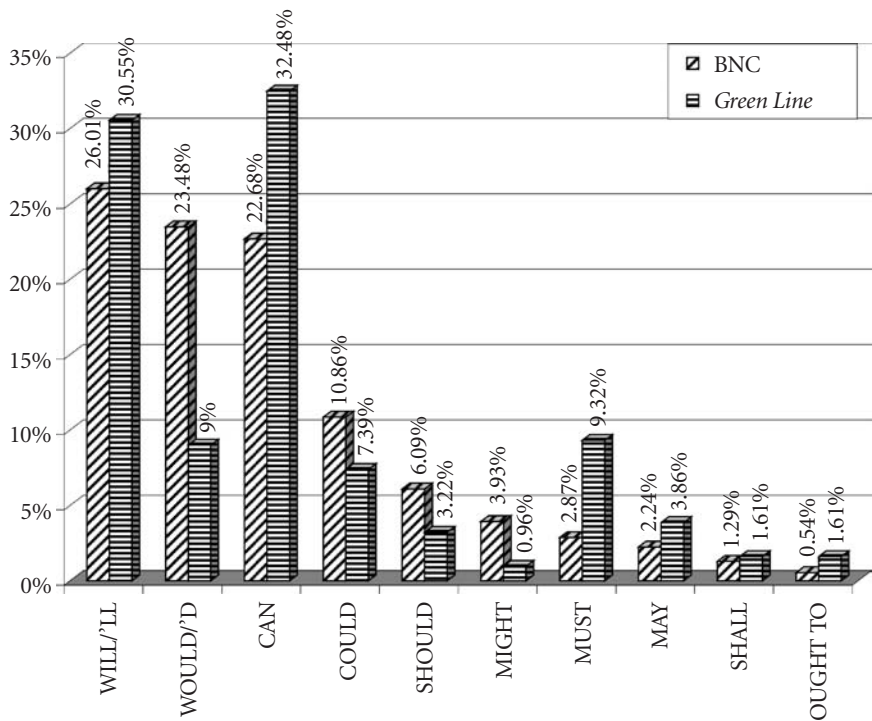


Figure 5. Relative frequencies of modals in BNC and *Green Line*

the meaning “possibility” in actual language use (83%). In the textbooks *might* is exclusively used to express a possibility. Even if this is also the most frequent meaning in the BNC data, there are some instances of the modal where it is used to ask for permission (3.5%). *Will* and *would* are less frequently used in *Green Line* in their prediction meaning than they are in “real” English. Another striking phenomenon is that *shall* is never used to make predictions in the textbooks and school grammars although this is an important meaning of the modal in the BNC (with 31%). There is another mismatch between “school” English and “real” English concerning *must*. Although the modal expresses an inference/deduction in 39% of the BNC concordance lines analysed, this meaning is only expressed in very few sentences in *Green Line* (6.9%).

Beside the differences related to the polysemy of the modal verbs, a couple of interesting observations could be made concerning the modals’ syntactic surroundings in corpus and textbook data. On the whole, we can say that the percentages of modal negation are much higher in *Green Line* than in spoken English. It is striking, however, that some modals (*might*, *shall*, *ought to*) are not

negated at all in the textbook texts analysed. There are also some differences with respect to the percentages of questions. In *Green Line* *might*, *must*, and *ought to* never occur in questions, whereas *could*, *may*, *will*, *shall*, and *should* are much more frequently used in questions in the textbooks than in the spoken part of the BNC. In the case of *shall* the textbooks even imply that this modal is exclusively used in questions despite the fact that 63.5% of the BNC concordance lines are statements. The shares of if-clauses in *Green Line* as compared to the BNC data are much too high concerning *might* and *would*, but the other modals are too rarely, or even not at all, used in if-clauses. Another mismatch between corpus and textbooks is the non-occurrence of set phrases and word clusters like “if I *may*”, “*must* admit”, or “*might* as well” in the latter.

5. Suggestions for the improvement of teaching materials

From these and other discrepancies between corpus and textbook data some consequences can be drawn and suggestions for an improvement of teaching materials on the basis of the findings from this corpus-driven approach can be made. The central questions are: “How can we come closer to achieving a high degree of authenticity in English language teaching as called for in the *Richtlinien*?” and “How can we reach the aim of teaching pupils an English which is comparable to native speaker English?”.

Assuming that the collection of EFL textbook texts used in the present study indicates the kind of English prioritised in English language teaching in German schools, a couple of changes concerning the use of modal verbs might be helpful to make the English that is taught more natural and native-like. First of all, I would suggest changing the order in which the modals are introduced from

can → *must* → *may* → *could* → *would*/'d → *should* → *will*/'ll → *shall* →
ought to → *might*

to

will/'ll → *would*/'d → *can* → *could* → *should* → *might* → *must* → *may*
→ *shall* → *ought to*,

an order which is based on corpus findings. In my opinion, other things being equal the more frequent verbs (i.e. the more important verbs, at least from a communicative point of view) should be introduced at an earlier stage in the learning process than the less frequent ones. Secondly, I consider it impor-

tant, if we want to enable pupils to communicate successfully, not to leave out some of the different meanings modal auxiliaries can have, e.g. the permission meaning of *might* and *could*, and to stress the inference/deduction meaning of *must*. To achieve a higher degree of authenticity, we might want to use similar proportions of the different senses of a polysemous verb in the English used in schools as found in the English used in real-life situations. Hence it would be a move in the right direction to present *can*, *could*, and *may* more frequently in contexts where they express possibilities.

As most of the modals were found to be used too frequently in negative contexts in *Green Line* texts, this overuse ought to be avoided if possible. This is not supposed to mean that some of the modals' negative forms are to be excluded from the teaching materials. It is, however, important to mention how often each verb occurs in affirmative and negative contexts.

Other changes that might lead to a higher degree of authenticity and thus to an improvement of textbooks like *Green Line* are to use *might* and *would* less frequently but the other modals more often in if-clauses, to mention the fact that *shall* does not only occur with a first person singular subject, and to avoid using *shall* exclusively in questions as it also occurs in statements. It may also be worth mentioning that some of the modals could be presented more frequently in tag-questions and in set expressions like "if I *may*" or "*might* as well" and that *should* and *must* could be used more often in passive constructions in the textbooks.

As many of the examples taken from *Green Line* sound rather unnatural, I would like to stress the importance of banishing invented sentences from textbooks and suggest to use preferably authentic material from a corpus instead. This approach to base English language teaching on real examples taken from corpora and to expose pupils to natural language was already formulated by Dave Willis in *The Lexical Syllabus*.⁷ Willis gives a description of the *Collins COBUILD English Course*, a lexically-based course making extensive use of spontaneously produced examples and stresses the importance of an "exposure to authentic language materials" (Willis 1990:46). Another supporter of the authenticity principle is John Sinclair who advises language teachers to "[p]resent real examples only" and considers it "most unwise to offer examples which are unattested, or to make major changes to actual instances" if example sentences are supposed to serve as models of English language usage (Sinclair 1997:31).

Moreover it could be considered an improvement to present the modals as a group rather than treating them separately. In this context the differences to full verbs such as the so-called "NICE properties" ought to be stressed to

give pupils a clearer picture of how modal verbs are used differently from other verbs.⁸ Finally I would like to suggest to focus more on the connection between past-tense-modals and politeness, an important concept which is still very much neglected in the EFL classroom.

6. Conclusion

As we have been able to observe, the results of the analysis make it clear that corpus-driven approaches to language learning and teaching can be very helpful for teachers and schoolbook publishers and that, to cite Dieter Mindt, “corpus-based studies of grammar (...) can do much to bring the teaching of English into accordance with actual language use” (Mindt 1997:50). The way the topic of “modal auxiliaries” is treated in English lessons in German grammar schools and the way *can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, *ought to*, and *must* are presented in teaching materials differ considerably from the use of those verbs in contemporary spoken British English. At the expense of quite frequent and important aspects (e.g. certain modal meanings) which are underrepresented or sometimes even left out completely some minor and less important features of usage are over-emphasised in the textbooks. I fully agree with McEnery and Wilson, when they say “... non-empirically based teaching materials can be positively misleading and [...] corpus studies should be used to inform the production of materials, so that the more common choices of usage are given more attention than those which are less common” (McEnery & Wilson 2001:120). This postulate suggests that a lot of corpus-driven work still has to be done to reach the aim of enabling both pupils and teachers to learn and teach an English which is more authentic and closer to that of native speakers.

Notes

1. A more detailed account of the investigations reported on in the present paper can be found in Römer 1999.
2. For the convenience of the reader the modals under investigation (i.e. *can*, *could*, *may*, *might*, *will*, *would*, *shall*, *should*, *ought to*, and *must*) will always be italicised.
3. The BNC is a fully part-of-speech-tagged corpus of over 100,000,000 words of both written and spoken British English. The spoken subcorpus makes up 10% of the whole corpus

and contains e.g. interviews, lectures, radio programmes, and everyday conversations. (cf. Aston & Burnard 1998: 31–36)

4. The *Richtlinien* serve as guide-lines which tell teachers what they are supposed to teach and how they are supposed to teach it. They are similar to the National Curriculum in Great Britain.

5. At the time they use the six volumes of *Green Line*, German pupils are between 10 and 16 years old. For most of the pupils English is the first foreign language.

6. According to one of the editors, *Green Line* is committed to a close representation of today's English (cf. Tegethoff 1984: L5).

7. The "lexical syllabus" was first described by A. Renouf and J. Sinclair in their 1988 article "A lexical syllabus for language learning" (in R. Carter & M. McCarthy (Eds.) *Vocabulary and Language Teaching*. London: Longman) and then further explained in Willis' book.

8. NICE is an acronym which stands for negation, inversion, code, emphasis (cf. Coates 1983: 4). Full verbs and modal verbs differ considerably with respect to the NICE properties.

References

- Aston, G. & L. Burnard (1998). *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.
- Bald, W.-D. (1991). Modal auxiliaries: Form and function in texts. In C. Uhlig & R. Zimmermann (Eds.), *Anglistentag 1990. Proceedings* (pp. 348–361). Tübingen: Niemeyer.
- Beile, W., A. Beile-Bowes, R. Hellyer-Jones, & P. Lampater (Eds.). (1984 [1989]). *Learning English. Green Line 1* [–6]. *Unterrichtswerk für Gymnasien*. Stuttgart: Klett.
- Coates, J. (1983). *The Semantics of the Modal Auxiliaries*. London: Croom Helm.
- Hermerén, L. (1978). *On Modality in English. A Study of the Semantics of the Modals*. Lund: CWK Gleerup.
- Leech, G. & J. Coates (1980). Semantic indeterminacy and the modals. In S. Greenbaum, G. Leech, & J. Svartvik (Eds.), *Studies in English Linguistics: For Randolph Quirk* (pp. 79–90). London: Longman.
- McEnery, A. M. & A. Wilson (2001). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Mindt, D. (1987). *Sprache – Grammatik – Unterrichtsgrammatik*. Frankfurt: Diesterweg.
- Mindt, D. (1995). *An Empirical Grammar of the English Verb: Modal Verbs*. Berlin: Cornelsen.
- Mindt, D. (1996). A corpus-based empirical grammar of English modal verbs. In C. E. Percy, C. F. Meyer, & I. Lancashire (Eds.), *Synchronic Corpus Linguistics. Papers from the Sixteenth International Conference on English Language Research on Computerized Corpora (ICAME 16)* (pp. 133–141). Amsterdam: Rodopi.
- Mindt, D. (1997). Corpora and the teaching of English in Germany. In A. Wichmann, S. Fligelstone, A. M. McEnery, & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 40–50). London: Longman.

- Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (Ed.). (1993). *Richtlinien und Lehrpläne für das Gymnasium – Sekundarstufe I – in Nordrhein-Westfalen*. Frechen: Ritterbach.
- Mitchell, K. W. (1988). Modals. In W.-D. Bald (Ed.), *Kernprobleme der englischen Grammatik – Sprachliche Fakten und ihre Vermittlung* (pp. 173–192). Berlin: Langenscheidt-Longman.
- Palmer, F. R. (1990). *Modality and the English Modals*. London: Longman.
- Quirk, R., S. Greenbaum, G. Leech, & J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- Römer, U. (1999). Das System englischer Modalverben und seine Stellung im Unterricht. Unpublished MA thesis, English Department, University of Cologne, Germany.
- Sinclair, J. (1997). Corpus evidence in language description. In A. Wichmann, S. Fligelstone, A. M. McNery, & G. Knowles (Eds.), *Teaching and Language Corpora* (pp. 27–39). London: Longman.
- Tegethoff, E. (Ed.). (1984). *Learning English. Green Line 1. Lehrerbuch*. Stuttgart: Klett.
- Ungerer, F., P. Pasch, P. Lampater, & R. Hellyer-Jones (1989). *Learning English Grundgrammatik. Ausgabe für Gymnasien*. Stuttgart: Klett.
- Willis, D. (1990). *The Lexical Syllabus. A New Approach to Language Teaching*. London: Harper Collins.

Resources – Computing

Basic processing

Software for corpus access and analysis

Michael Barlow

University of Auckland

This chapter describes the use of software to extract wordlists, collocates, collocations, and concordances from a text. Each of these data types can be seen as different views of a corpus and each provides different kinds of information about the form and content of texts. In some cases, it is profitable to examine words with all context removed, as in a wordlist or word frequency list. In other cases, the frequent collocates in the proximate context of keywords, such as *husband* and *wife*, yield important information on recurrent formal patterns and also on semantic, and even cultural, associations of the keyword. For other analytical purposes, a larger contextual unit, perhaps the sentence or paragraph, is necessary to fully understand the function of a word or phrase. The availability of corpora in a digital format allows for multiple (and concurrent) transformations of a text, thereby providing the analyst with multiple perspectives on the text. As described in this chapter, each of these perspectives has particular advantages for highlighting distinct aspects of the usage of words and larger units.

There are now available a wide variety of software programs, PERL scripts, and UNIX utilities providing access to machine-readable texts. Rather than describe the operation of particular programs and scripts, I will focus on some fundamental questions related to corpus access, in particular the relationship between corpus access and corpus analysis, showing how different views of a corpus lead to the highlighting of some patterns in the texts and the obscuring of others. By considering a range of different ways of analysing a text, we can gain a clearer appreciation of the role of corpus analysis software in mediating our view of a text and we can discover the kinds of text transformation that fit best with different purposes in corpus analysis.

The benefits of converting text into a digital form come from the accessibility that digital formats provide and from the potential for interchange and transmission of data. In addition, the representation of a text in digital format

allows for unlimited duplication without any loss of quality. Consequently it is easy to transform texts in many different ways and also to obtain multiple, simultaneous, perspectives on a corpus. The availability of alternative views of a single data source is, perhaps, taken for granted these days, even though it was not so long ago that text analysis involved reading through a text and marking particular items of interest with coloured pens.

1. Starting with the text

Access to a text, in terms of viewing the text, need not involve radical transformations. However, since we are viewing a digital representation of the text rather than a text in a newspaper or book, we should keep in mind that a digital text is, itself, a transformed text. The transformations involved, such as a change in font type, the loss of page boundaries, or changes in the representation of paragraphs, may well be unimportant for many kinds of linguistic analysis, but we should not forget that what looks like an ordinary rendition of a text has, in fact, been changed quite radically in the process of digitising it.

The small text below in Figure 1 is a representation of a letter to the editor in *The Times* newspaper. This text contains a fair amount of mark-up, here enclosed within angle brackets, making some of the changes from the original readily apparent and concomitantly making the letter itself harder to read. In the electronic version of the text, some of the annotations are designed to make

```
<Article>
<PageNumber>15</PageNumber>
<Source>The Times</Source>
<Date>01 January 1996</Date>
<Headline>Ashtrays a la Carlyle</headline>
<byline></byline>
<SectionOfPaper>Letters to the editor</SectionOfPaper>
<Story><Group>
From Mrs A. C. Whitmore
Sir, I was interested to read of the unusual ashtrays at The Carlyle, New York (letter, December 21). One thing puzzles me. How does one successfully empty and clean an ashtray which is glued to a table? Is there some exclusively American technique?
Yours faithfully,
BERYL P. WHITMORE,
29 Davies Avenue,
Roundhay, Leeds, West Yorkshire.
December 21.
</Group></Story></Article>
```

Figure 1. Letter to the editor of *The Times*

explicit the provenance of the text. The annotations, or tags, at the beginning of the extract provide information about the text, and coincidentally make clear the distinction between the digital version and the original. Some information from the original has been lost and some information has been added, or at least rearranged, such as the reference to the page number and date.

2. Creating a wordlist

Probably the most radical transformation of a text used in linguistic analyses is to, in effect, rip it apart to produce a wordlist. (An even more extreme modification would involve breaking the text into individual characters, a change that might be more useful for cryptographic purposes than for linguistic investigations.) Creating a wordlist involves snipping the text at particular places, typically the spaces between words, and listing and counting the resulting tokens. Again, it is important to be aware of the existence of choices made in performing this transformation, as well as of the uses that can be made of the transformed text. While modern programs typically allow us to simply click on a button to create a wordlist from our text, it is instructive to consider how we might go about this task step by step. Our first thought might be to use spaces (and line returns) as markers of word boundaries, but applying this procedure to our sample text would mean that the first word would be `<Article>` and the second word `<PageNumber>15</PageNumber>`. This might lead us to add punctuation as a marker of word boundaries, and we can make further distinctions in order to differentiate among different kinds of punctuation and also distinguish symbols that are taken to be part of a word (such as a hyphen) from those that define a word boundary (e.g., brackets and forward slash).

Having settled on a suitable set of word delimiters, we would still be left with *PageNumber* and other tag names in the wordlist, which we presumably do not want. As corpus mark-up becomes more and more complex, it becomes increasingly important to be able to distinguish between words and annotations so that the information contained in the tags can be excluded from the wordlist if desired.

The fact that written English separates words with spaces makes the creation of a wordlist much easier for English and for other European languages than for languages such as Japanese and Chinese. However, the ease with which a wordlist can be created should not lead to the conclusion that the definition of a word in English is straightforward. As noted above, it is necessary to decide whether *it's* is one word or two. In other words, is the apostrophe to be treated

as a word boundary? And then if we add the apostrophe to the set of word delimiters, we must decide whether to process *don't* in a special way to avoid the division into *don* and *t*. Similarly, we might want to consider whether phrases such as *of course* (cf. *because*) and *in spite of*, or even *a la*, which occurs in the heading of the letter to the editor shown in Figure 1, should be treated as single lexical units. In practice, most investigators make do with a basic approach to wordlists, defining spaces and punctuation as word boundaries in all cases, and then searching separately for phrases such as *in spite of*, although some tagsets may treat a phrase such as *in spite of* as a single unit. In practical terms, taking spaces and punctuation as word delimiters works quite well, but we should be aware of the possibility that decisions made on practical grounds can have an insidious effect on the way that we think about the structure of language.

Transforming a text into a list of words removes the context for individual words, which means that all of the linear context, i.e., all syntagmatic information, is lost. Consequently, it is not possible to use a wordlist to analyse how the information in the text is presented, but we can obtain some idea of the content of the text from an unstructured wordlist alone.

The most frequent words in any corpus, the function words *the*, *of*, *and*, etc., cannot be used to identify text content. And indeed these words are sometimes filtered out of the word list by using a stoplist. A stoplist typically consists of the most common words or closed class categories of words such as determiners and prepositions.

The first fifteen lines of the wordlist for the sample text are shown in Figure 2. Tag names have been omitted and case-differences have been ignored. The list shows us that the text is related in some way to 'ashtrays,' which is a distinctive word mentioned twice.

4	the
3	a
3	to
2	21
2	ashtrays
2	december
2	is
2	one
1	01
1	15
1	1996
1	29
1	american
1	an
1	and

Figure 2. A frequency-ordered wordlist

In general, wordlists are created for larger corpora and their distinctiveness is assessed either by inspection or by comparison with other corpora, perhaps a large reference corpus.

3. Adding context

We can take a less disruptive perspective on the corpus by adding a small amount of context. Each word in the text has a syntagmatic context; it occurs with preceding and following words, with the simplest window of context being a span of one or more of these preceding and following words. This leads naturally to the “keyword in context” or KWIC display, typically associated with concordancers, in which the keyword is displayed in the centre of the concordance line. This simple form of display has two main advantages. First it is not necessary to read through the concordance line searching for the keyword, and secondly, lining up multiple instances of the keyword in the same location allows textual patterns based on the context of the keyword to stand out visually, as we will see below.

Context words are displayed regardless of the tightness of their connection with the keyword. In other words, no account is taken of intervening boundaries such as sentences or paragraphs or section breaks. This is not normally problematic since accidental cooccurrence will typically be overwhelmed by the more frequent patterns, but some disruptive noise can be introduced when looking at low frequency patterns.

One way of transforming a text is by producing a concordance for every word or, perhaps, every content word. The first part of this type of concordance (from *a* to *empty*), based on the text of the letter to the editor, is shown in Figure 3. This kind of display, where the text is transformed such that each word is displayed along with its context, illustrates very clearly both the duplication of an original text and also the notion of corpus access as corpus transformation since the original text has been thoroughly modified to present this view.

The next and perhaps most useful type of transformation is the listing of all occurrences of a single item. Searching a corpus for a single item and its contexts reveals patterns of cooccurrence which can give us some idea about the usage of the item, as well as perhaps telling us something about its cognitive associations.

Let us observe a concordance of the word *thing*, which is quite common in both written and spoken texts, in order to see how changes in the size and form of the context affect the patterns that emerge.

1.	... 1996 Ashtrays	[[a]] la Carlyle ...
2.	... glued to	[[a]] table? Is ...
3.	... some exclusively	[[American]] technique? ...
4.	... and clean	[[an]] ashtray which ...
5.	... successfully empty	[[and]] clean an ...
6.	... clean an	[[ashtray]] which is ...
7.	... January 1996	[[Ashtrays]] a la ...
8.	... the unusual	[[ashtrays]] at The ...
9.	... unusual ashtrays	[[at]] The Carlyle, ...
10.	... 29 Davies	[[Avenue]], Roundhay, ...
11.	... Yours faithfully,	[[BERYL]] P. WHITMORE, ...
12.	... a la	[[Carlyle]] Letters to ...
13.	... at The	[[Carlyle]], New York ...
14.	... empty and	[[clean]] an ashtray ...
15.	... WHITMORE, 29	[[Davies]] Avenue, ...
16.	... York (letter,	[[December]] 21). One ...
17.	... West Yorkshire.	[[December]] 21. ...
18.	... me. How	[[does]] one successfully ...
19.	... to the	[[editor]] From Mrs ...
20.	... one successfully	[[empty]] and clean ...

Figure 3. Concordance of each word in a text

If we look at the use of *thing* in the letter to the editor in Figure 1, we see that it occurs just once and is preceded by the word *one* and followed by the word *puzzles*. Since there is only one instance of *thing*, it is impossible to assess the strength of the association between *thing* and the neighbouring words, *one* and *puzzles*. Clearly, we need to examine multiple instances of a word in its context to judge whether a particular word combination is a common cooccurrence or something more unusual.

Performing a search for *thing* in *The Times* newspaper corpus with a restricted context of one word on each side, we notice some patterns emerging as certain words naturally cluster around *thing* recurrently, as can be seen in the small sample of concordance lines in Figure 4. These few instances of *thing* are ordered as they are in the text itself. In other words, line 1 contains the first instance of *thing*; line 2, the second, and so on.

Taking these twelve examples together, we might find suggestive the fact that *thing as*, *one thing*, *right thing* and *whole thing* all occur twice. Line 3 contains the usage of *one thing puzzles* from our letter corpus, and *one thing* also shows up again in line 5.

Scanning down the list of 3-word sequences in Figure 4, we see an important change in perspective. We have moved from a completely linear, syntag-

1. Every **thing** now
2. bad **thing**. Yours
- 3 One **thing** puzzles
4. right **thing** to
5. One **thing** is
6. bullying **thing**. It
7. a **thing** or
8. whole **thing** as
9. of **thing** government
- 10 a **thing** as
11. right **thing**: putting
12. whole **thing** is

Figure 4. Concordance of *thing*

matic, view of a text (shown in Figure 1) to more of a paradigmatic view of the lexical item *thing*, which can only be observed if we have multiple instances of the target word. Figure 4 still reflects some aspects of the syntagmatic dimension due to the fact that the twelve items are arranged in the order in which they appear in the text, and, of course, within each line we have the three-word syntagmatic sequence. To obtain a clearer paradigmatic view, we need many more examples and also a better arrangement of those examples.

The lexical associations or patterns involving *thing* stand out visually when all the lines containing *thing* are re-ordered such that they are alphabetised in terms of the following word, as in Figure 5. There is a second component to this ordering. For those lines in which *thing* is followed by the same word, the lines are sorted, secondarily, in alphabetical order of the word preceding *thing*. Thus we have a 1st Right, 1st Left sorting pattern.

The selection in Figure 5 shows just twelve lines (a different twelve lines from those shown in Figure 4), but in practice we would typically be working with results numbering in the hundreds or thousands of instances.

Alternatively, we can sort the results primarily on the alphabetical order of the preceding word, and secondarily on the following word, giving a 1st Left, 1st Right sort order, as shown in Figure 6.

The change in observable patterns moving from Figure 4 to Figures 5 and 6 is quite marked and we are able to see different patterns emerging as we rearrange the data. It is also evident from these simple examples that the patterns we observe are greatly influenced by the frame we impose on the data. In this case the frame can be described as the search word ± 1 word (in alphabetical order).

1. one **thing**. A
2. the **thing**. A
3. unique **thing** a
4. a **thing** about
5. a **thing** about
6. amazing **thing** about
7. astonishing **thing** about
8. basic **thing** about
9. best **thing** about
10. best **thing** about
11. best **thing** about
12. best **thing** about

Figure 5. Concordance of *thing* sorted 1st Right, 1st Left

1. 20th-century **thing** to
2. a **thing**), commentary
3. a **thing**." Blair
4. a **thing**!" And
5. a **thing** about
6. a **thing** about
7. a **thing** as
8. a **thing** as
9. a **thing** as
10. a **thing** as
- 11 a **thing** as
12. a **thing** as

Figure 6. Concordance of *thing* sorted 1st Left, 1st Right

Changing the size or nature of the context leads to different results. This is perhaps an obvious fact, but the consequence is that unless you are developing your own scripts to manipulate texts, your analysis will be limited by the choices available in the software you use. A well-designed software program makes it appear that the patterns in the data are simply being revealed for the user, but we have to remember that text analysis programs not only highlight, but also limit, the kinds of patterns that can be extracted from a text.

3.1 Collocates and collocations

An alternative to scanning through a large number of concordance lines is to work from a collocate frequency table, which shows, for the whole corpus, the frequency of words within a particular span of the search word. A collocate fre-

quency table based on *thing* (Table 1) shows most clearly collocations involving *one thing*, *whole thing*, *only thing*, and other recurrent modifiers. The data also suggests other patterns to look for such as those involving *is* and *to* following *thing*, as well as patterns based on *such* and *sort*, which appear in the 2nd left column (i.e., two words before the search term).

Why are linguists so interested in collocates? One reason comes from the idea that collocations are the building blocks of language and are, in some sense, fundamental units of language in use and that frequency of occurrence has a direct impact on the organisation of grammar (Barlow & Kemmer 2000; Bybee & Hopper 2001). In addition, the frequency of context words surrounding a keyword is used, within the Firthian-Sinclair tradition (Firth 1957; Sinclair 1991), to provide clues as to the nature of the search word. The accumulation of contextual patterns provides good empirical evidence on various aspects of the meaning of a word.

As a brief illustration, we can look at the collocates of *husband* and *wife*, using *The Times* newspaper, 1995–1997, in order to throw some light on the ways in which husbands and wives are viewed, at least through the prism of a British newspaper in the mid-nineties.

First, by simply counting the instances of the words *husband* and *wife*, we discover that there are 9,197 instances of the word *husband* and 17,795 instances of *wife*, showing that *wife* is almost twice as frequent as *husband*. Bigamy and polygamy aside, one would expect the number of actual husbands and wives to be equal, and yet there is this large discrepancy in mentions, which

Table 1. Collocates of *thing*

2nd Left		1st Left		1st Right		2nd Right	
freq	token	freq	token	freq	token	freq	token
666	the	203	one	183	is	176	the
279	The	117	whole	153	to	87	a
184	a	116	a	131	that	68	that
87	most	109	of	127	about	52	to
81	sort	100	only	101	I	49	have
69	no	95	the	80	in	44	do
58	is	89	same	80	you	36	can
35	such	83	first	71	as	34	is
31	147	75	good	71	148	31	it
30	only	72	last	63	for	30	146
27	kind	67	best	56	and	28	I
21	an	65	real	51	of	28	would
19	very	59	important	35	we	22	be

Table 2. Modifiers of *husband* and *wife*

Modifiers of <i>wife</i>	Percentage of total occurrences (<i>wife</i>)	Modifiers of <i>husband</i>	Percentage of total occurrences (<i>husband</i>)
his wife	48.2	her husband	49.8
my wife	8.3	my husband	10.0
the wife	7.0	the husband	6.7
a wife	2.5	a husband	3.7

leads to some interesting questions: To what extent are the newspaper articles about men rather than women, and are women more likely to be referred to via their roles as wives? If women are, in fact, more likely to be referred to as wives than men are referred to as husbands, then we might expect the phrase *his wife* to occur more often than *her husband* (taking into account the frequency difference in the occurrence of *husband* and *wife*).

Looking at the data shown in Table 2, we see that this expectation is not borne out. The use of pronouns and articles is remarkably similar for *wife* and *husband*, with about half the instances of *husband/wife* being preceded by *her/his*. On the other hand, there are some differences in percentages, such as the slightly greater percentage of *my husband* compared to *my wife*. Such differences are suggestive and call for closer investigation and a search for explanation.

For further insight, we can examine the words in the corpus that precede *husband*, excluding the grammatical words *my*, *the*, etc. The result is the following list, which is presented in decreasing order of frequency: *former*, *first*, *late*, *estranged*, *second*, *future*, *new*, *jealous*, *dead*, *violent*, *devoted*, *good*, *drunken*, *rich*, *third*, *French*, *American*, *English*, *elderly*, *dying*, *British*, *faithless*, *alcoholic*, *wonderful*, *unfaithful*, *daughter's*, *friend's*, *philandering*, *sick*, and *architect*. The equivalent list for *wife* is: *second*, *first*, *former*, *estranged*, *young*, *third*, *new*, *pregnant*, *future*, *farmer's*, *Minister's*, *fourth*, *long-suffering*, *minister's*, *leader's*, *Tory*, *beautiful*, *English*, *divorced*, *common-law*, *American*, *late*, *battered*, *vicar's*, *MP's*, *political*, *President's*, *devoted*, *doctor's*, *perfect*, *French*, *Blair's*, *actress*, *politician's*, *pretty*, *good*, *murdered*, *non-working*, and *working*.

There are various aspects of these lists that are worthy of comment, both in terms of similarities and differences. One might note the differences in the adjectives used in the two lists: the ones preceding *husband* in many cases refer to negative traits, while the adjectives associated with *wife* are more likely to be either positive traits for a wife, or modifiers portraying the wife as a victim.

Further, we find a notable contrast in these lists between the kinds of possessive nouns that typically occur with each of the two words, as in *daughter's*

or *friend's husband* versus *farmer's /Minister's/leader's/MP's/President's/doctor's/ Blair's/politician's wife*, where the possessive nouns preceding *husband* refer to social or familial relations, while those preceding *wife* refer to men in terms of their position or role in society, usually a prominent one.

This brief illustration of observing and comparing collocates of contrasting terms provides some sense of how using something as simple as sorting and counting software can yield new insights into language usage and suggest avenues for further investigation into aspects of culture and society.

4. Wider context

So far we have looked at a particular search word and examined the common words that cluster around it. A related, but slightly different, perspective involves the identification of collocations containing the search word that are three words or longer. Table 3 shows the most common three-word sequences containing *thing* in a newspaper corpus.

Table 3. Three-word collocations based on *thing*

115	5.2%	the whole thing
100	4.5%	the only thing
88	4.0%	the same thing
81	3.7%	sort of thing
69	3.1%	the last thing
69	3.1%	the first thing
64	2.9%	the real thing
60	2.7%	the best thing
55	2.5%	thing is that
54	2.4%	no such thing
48	2.1%	a good thing
40	1.8%	the right thing
40	1.8%	such thing as
39	1.7%	is one thing
38	1.7%	a thing of
35	1.6%	such a thing
35	1.6%	only thing that
33	1.5%	thing of the
32	1.4%	thing in the
31	1.4%	the one thing
31	1.4%	important thing is
30	1.3%	thing to do

Table 4. Five-word collocations based on *thing*

28	is no such thing as	5	the only thing you can
16	no such thing as a	5	knows a thing or two about
12	the most important thing is	4	is the last thing on
11	it is one thing to	4	the most important thing was
11	the most unusual thing you	4	the most remarkable thing about
10	is the first thing you	4	but the great thing is
10	most unusual thing you have	4	not the only thing that
10	the first thing you would	4	the first thing you notice
10	the important thing is that	4	the nearest thing to a
10	the first thing you would do	4	the same thing is happening
10	most unusual thing you have done	4	such a thing as the
9	is such a thing as	4	such a thing as a
9	is a thing of the	4	no such thing as a free
8	most important thing is to	4	be a thing of the past
7	the great thing is that	4	the great thing is that they
7	is a thing of the past	3	the best thing to happen to
6	the sort of thing that	3	the important thing is that the
6	s no such thing as	3	the nearest thing we have to
6	the best thing to do	3	no such thing as bad publicity
6	the whole thing is a	3	the hardest thing i have ever
6	be a thing of the	3	knew a thing or two about
6	say one thing and do	3	the first thing you notice is
6	the whole thing has been	3	the best thing to do is
6	say one thing and do another	3	the best thing to have happened
5	the kind of thing that	3	a funny thing happened on the
5	knows a thing or two	3	the important thing is that we
5	the important thing is to		

We observe that almost a quarter of the instances of *thing* in the corpus occur in the top six collocations listed in Table 3, confirming the fact that viewing the text through this particular three-word window reveals some robust patterns, which were not apparent in the collocate frequency table shown in Table 1. Shifting perspective again, we can look at five-word collocations involving *thing*, as shown in Table 4. Here, not surprisingly, the number of repeated patterns is smaller and we find the occurrence of somewhat looser connections among these words. Nevertheless, these strings have a familiar ring to them, suggesting that there exist in English some larger, semi-fixed units based on the word *thing*.

Some patterns will fall outside even this extended span of words. For example, examining 28 instances of *it's one thing to*, we discover that, based on the 16 instances in which some kind of “*another*” phrase actually occurs, the number of words occurring before the use of the complementary *another* phrase ranges

from 2 to 60, with the average being 11. These results show that while a narrow context window reveals a number of strong collocations, a wider view may reveal other, possibly more subtle, association patterns.

5. Lexical frameworks

It is illuminating to specify a lexical context of immediately surrounding words, what Renouf and Sinclair (1991) call a framework, and observe the influence of the framework on the intervening collocate. Frameworks have the form *word* ____ *word* and we can use them as a basis for contrasting related frameworks. Thus we can contrast the framework *the* ____ *that*, shown in Table 5, with *a* ____ *that* shown in Table 6.

Table 5. Collocates of *the* ____ *that*

Frequency	Percentage	Collocates of <i>the</i> ____ <i>that</i>
4499	13.1%	the fact that
694	2.0%	the idea that
464	1.3%	the way that
433	1.2%	the view that
409	1.1%	the impression that
380	1.1%	the hope that
369	1.0%	the news that
334	0.9%	the belief that
317	0.9%	the knowledge that
312	0.9%	the ground that

Table 6. Collocates of *a* ____ *that*

Frequency	Percentage	Collocates of <i>a</i> ____ <i>that</i>
439	2.9%	a way that
310	2.0%	a warning that
230	1.5%	a move that
204	1.3%	a company that
177	1.1%	a team that
147	0.9%	a system that
137	0.9%	a sign that
121	0.8%	a car that
113	0.7%	a game that
108	0.7%	a country that

By comparing the two sets of results, we see how each lexical framework, although very similar in form, selects different (but overlapping) sets of collocates. We also observe that the strength of the lexical links is much greater for *the fact that* and *the idea that* than for other collocates. Again, we present this as an illustration of how the nature of the contextual window influences which language patterns are retrieved.

6. More on context

So far we have examined the patterns associated with the word *thing* in arbitrary and rather artificial contexts, namely, within a certain span of words. This perspective has proved to be interesting and revealing. However, such local contexts cannot reveal the relationship between particular words and other natural units such as the sentence or paragraph or turn, in the case of spoken data. In the letter to the editor (Figure 1), we can see that *one thing* occurred sentence-initially and so we might want to know whether there is a general preference for this phrase to occur at the beginning of a sentence. To investigate this question we need to change the viewing context from a span of words to a sentence and examine the placement of *one thing*. Figure 7 shows the first 10 instances of *one thing* in the newspaper corpus, and while this is a small unrepresentative sample, it does suggest that a full investigation might show that the usage in the letter was typical and that *one thing* does usually appear towards the beginning of a sentence.

In contrast to *one thing*, the phrase *kind of thing* tends to occur in the middle or end of sentences, as the ten sentences in Figure 8 suggest.

Choosing the sentence as a contextual unit reveals differences in the distribution of the two phrases. Such differences might be brought into sharper focus if the corpus were annotated to show anaphoric chains and these were used as part of the viewing context.

6.1 Annotation as context

We can imagine performing searches that refer to part-of-speech tags (e.g., *spoke* <*w VVD*>), which are part of the analysis of words in the text, but which for practical reasons occur next to the words they categorise. Thus extending our notion of context, we could say that <*w VVD*> is part of the context of *spoke* and other past tense verbs. Similar comments apply to other annotations such as syntactic categories or semantic information.

1. **One thing** puzzles me.
2. **One thing is clear:** there is no room anywhere for fodder, for cramped TV-style productions or botched little comedy thrillers like *The Steal*, the worst British film to crawl into the light last year.
3. But it is strange, because if there's **one thing** we do know, it is that if calorie consumption exceeds calorie expenditure then we'll put on weight.
4. On both occasions, fiscal policy had to be put into reverse, and the past three budgets have had **one thing** in common, namely, aiming to restore the underlying budget deficit to a more sustainable position.
5. "It is **one thing** being Jackie Stewart, the racing driver," he said, "and quite another to be Jackie Stewart, team owner.
6. There is only **one thing** worse than being named as most expensive provider this year, and that is ending the year in the same slot.
7. **One thing** that can be said about the 7th Earl of Anywhere or the 5th Lord Nobody is that, by succeeding to their titles, there can be no suspicion that they obtained them in anything other than honourable circumstances.
8. He said Mr Blair's speech had used words which appeared to the gullible to say **one thing** but to the hard-nosed analyst gave clear warnings of what Labour in power would be like.
9. It was **one thing** to go to Fleet Street, a real old community, now alas given over to bankers and lawyers – lawyers being far too astute to abandon premises in the true heart of London.
10. **One thing** is certain: the models are mad about it.

Figure 7. Position of *one thing* within sentences

1. People have never heard this **kind of thing** on South African television before.
2. So questions will just be about signs and braking distances, that **kind of thing**?
3. Time was, not so very long ago, when this **kind of thing** in Sydney would have been about as likely as the singing of the Internationale at the inauguration of the US president.
4. Stripped of that connection, it would be no different from **the kind of thing** found in any popular high-street jewellers.
5. This shows 12 different dalmatian dogs dressed up in cowboy, bunny and jungle explorer costumes, and is the **kind of thing** that photo-stylist Linda Groves might, in years to come, wish her name hadn't appeared on.
6. "Oh" drawled Clarke, "I suppose this is the **kind of thing** everyone would expect you to do."
7. Britten in his early twenties, Mendelssohn in his teens: this is the **kind of thing** that a young ensemble should be playing, surely, rather than late Beethoven.
8. A person who is agile enough to take part in various terrorist activities is about the last person to be compensated for this **kind of thing**.
9. "We told them there was no way they could get away with this **kind of thing**."
10. "Just the **kind of thing** Hamlet wants to hear before the court arrives with Ophelia's coffin.

Figure 8. Position of *kind of thing* within sentences

Another kind of annotation that can be loosely described as part of the context is the information relating to the author, date of publication, source text, and so on. In the British National Corpus (BNC), this information is included in a highly elaborated header structure which precedes each piece of text. The header consists solely of tags relating to the source and content of the text, such as date of publication, sex of the writer, etc. Such information can be used in context-sensitive searches to reveal non-obvious distinctions in linguistic patterns associated with different sets of texts.

7. Conclusion

In this paper we have examined some basic issues arising from the digitisation of textual information. A well-known characteristic of digital representations is the potential for duplication of data without loss of information. However, while simple duplication is very useful for the transmission and distribution of corpora, in order to reveal interesting language patterns duplication must be associated with text transformations of different kinds, as we have seen, for example, in the various kinds of display of concordance lines.

An important point illustrated here in different ways is that changing the context window reveals different views of a corpus and therefore different language patterns. This idea is a simple one, but one that is easy to forget since our attention naturally focusses solely on the object of our investigations and not on the tools used. The tendency to perceive only a single object of investigation leads to a situation in which annotations come to be seen as raw data rather than a level of analysis.

In this paper we have deliberately kept shifting attention back to the mode of analysis in order to emphasize the advantages and disadvantages of particular tools and annotation schemes. Being aware of the role of the tools of analysis becomes ever more important as digital transformations of corpus data become more complex and we move from concordances and collocate frequency displays to other, still more abstract, ways of visualising patterns in texts.

References

- Barlow, M. & S. E. Kemmer (Eds.). (2000). *Usage-based Models of Language*. Stanford: CSLI.
Bybee, J. & P. Hopper (Eds.). (2001). *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins.

- Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis* (Special volume of the Philological Society.) Oxford.
- Sinclair, J. McH. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Renouf, A. & J. McH. Sinclair (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics* (pp. 128–143). Harlow: Longman.

Programming

Simple Perl programming for corpus work

Pernilla Danielsson

University of Birmingham

Whether you are compiling your own corpus or simply want to use your corpus with new software, you will find the need for a conversion tool. Despite the fact that there are many conversion tools available, there never seems to be exactly the one you need. This article tries to tackle this problem by teaching the reader how to write simple Perl programs. The programs may not give the answer to every problem, but should be seen as a starting point for further exploration. The toolbox includes the following programs: a tokeniser, a word splitter, a frequency counter and a KWIC concordancer. Full program codes are given which you may type in and test on any computer where a Perl interpreter is installed.

1. Introduction

You do not need to be a Jack-of-all-trades to become a corpus linguist but you may well find yourself, after years in the field, having had to learn a bit of everything. Alongside the theoretical challenges, corpus linguists fight many practical battles with text editors, concordance tools, mark-up format and linguistic annotation. For anyone aiming to compile their own corpus, the endeavours of cleaning up text, tokenising, indexing, and annotating can become an insurmountable task, especially if they are to be done manually. If manual labour is not an option, particularly when working with very large corpora, then another option is often to wait for your organisation's over-worked technical staff to allocate time for investigating the problem. This usually involves some more or less embarrassing moments of mis-communication, where the power is with the one who knows which tasks are impossible to implement, while other tasks are achievable almost instantaneously. For the technical virgin, the difference between an impossible or possible computing task is often hidden in darkness. Why should it be so easy to count frequencies when it is so difficult to mark sentences?

This article will endeavour to offer a third way; the option to do your own programming. Rather than sitting around waiting for someone else to do the job, you could take a little time learning some simple programming in order to perform relevant operations. A few simple commands and some short programs can make all the difference.

The following text will contain a toolbox consisting of four small programs: a tokeniser, a word splitter, a frequency counter and a KWIC concordancer. These programs will not solve all corpus linguistics problems, but should be viewed as a stepping stone to enable further development.

2. When to use Perl

Programming may sound frightening. For those who were around before Windows the old memories of DOS and UNIX commands may come flashing back in horror. However, just as operating systems have become more user-friendly so have the programming languages. One of these more user-friendly and frequently used programming languages in corpus linguistics is Perl.

One reason for Perl's popularity may be found in the ease of use; reading an input file can be stated in one line of Perl code rather than by 20 lines of C code.¹ From its beginning, the intention of Perl was to ensure an easy and efficient way to navigate large text files. As such, the makers themselves state its success: “– the pattern matching and textual matching capabilities of Perl often outperform dedicated C programs” (Wall & Schwartz 1991: xii).

Although Perl programming may be perceived as easy, for anyone who has never programmed before there will be a steep learning curve. You need to be stubborn and persistent at the beginning and trust that you will be successful. All programming takes a bit of getting used to and what may appear as ‘proper English’ is not always used in its correct form. For example, we will use the function ‘*print*’. Contrary to expectations, this function does not send your file to the printer, but instead simply repeats whatever you have written after ‘*print*’ to the screen. A similar example is the use of the UNIX command ‘*more*’, which instead of giving you more of a file, actually displays only a small bit of the text on the screen.

Even if you decide not to persevere with Perl, it might still be worth getting used to a programmer's way of thinking, in order to familiarise yourself with what can or cannot be achieved by simple programming. The increase in the number of linguists and language teachers turning to corpus linguistics has affected the number and type of jobs for which corpora are used. It is

difficult for any programmer of popular corpus retrieval and search tools to incorporate all foreseeable features required by a potential user. Apart from the shortcomings of some of the corpus tools there are also gaps between the interchange formats of available corpora and the input formats accepted by most corpus tools. Today, many of the available corpora are marked up with SGML (Standard Generalised Mark-Up Language) or XML (Extensible Mark-Up Language), however, most corpus tools work on plain text (see Sperberg-McQueen & Burnard 2001; Ide & Veronis 1995 for further details on SGML and XML). The bloated format of SGML requires substantial transposition to return back to plain text; a task every programmer in the field of corpus linguistics has been confronted with at least once.

Further regularly encountered problems concern the line breaks in a text. For example, many text files have additional breaks in the middle of each line, which are there due to the typists thinking that the lines would be too long otherwise. When using a concordance tool such as WordSmith (see Mike Scott's webpage for the tool on <http://www.lexically.net/wordsmith/>) this will alter the display of the text and therefore may require the removal of these additional breaks. Another simple tool may be required to rectify mistakes in the conversion of files from working under MSDOS to working under UNIX, where an uncorrected text has left your document full of mysterious '^M' symbols before each line break. Yet another may help those who work with aligned texts; aligners often produce output to one single output file but in order to use the files with a parallel concordancer, such as ParaConc, the file needs to be split into two halves, one per language, where the line numbers will indicate which sentences are corresponding alignments.

All that is needed to rectify each of these difficulties is a small program. The problem is that the program will be slightly different every time you need it and so it is valuable to be able to adjust the programs yourself. Every tool has its own speciality, as has every corpus.

Any programmer familiar with Perl knows that many tasks can be completed in a number of ways; here, we will focus on programs that are easy to define and read. When you become more familiar with programming, you may find other ways which are more suitable, or alternative ways to reduce the code. Unfortunately, the reduced code is also likely, at least for the novice programmer, to reduce the readability. The four tools described in this article represent the first steps on the long journey to becoming a fluent programmer. The user's specific requirements for each tool may necessitate the slight customisation of the programs to achieve the intended results. It is hoped that the information contained in this article will be sufficient to enable any required alterations.

For this chapter, the operating system is assumed to be UNIX (or Linux), but Perl is also available for Windows and Macintosh. One advantage of using the UNIX operating system is that there is quite an affinity between UNIX and Perl; often, as we shall see, UNIX has a function that makes the Perl program simpler and shorter. If you are using another operating system there may be similar opportunities for such co-operation, but this is not guaranteed.

3. Step 1: Finding the interpreter and changing access rights

In order to try out the following programs, a Perl interpreter must be available on the computer. The programs introduced here will work also in Windows or Macintosh environments, as long as there is a Perl interpreter installed on the computer. If not, you may need help to install it. Perl interpreters for all operating systems can be downloaded from the websites www.perl.org or www.perl.com. Make sure to always download the version referred to as “stable” and not one of the test versions.

If Perl is already available, the initial programming steps involve discovering where the interpreter is located on the computer, which is most easily achieved by using the UNIX command ‘*which*’. This locates where the interpreter is stored and can be used by typing ‘*which perl*’ on a command line, see Figure 1 below. The computer used to test these examples returned the information ‘*/usr/bin/perl*’. Not every computer has the same set-up, however, the placement of the Perl interpreter is not as important as knowing where it is.

It may be worth mentioning that Perl is an interpreted programming language. This makes it different from other programming languages, such as C or Pascal which are compiled. The obvious differences are found in the way that you can read the program file. If you have a program file in Perl you can read the actual code, if you get a compiled program in C then the program code is

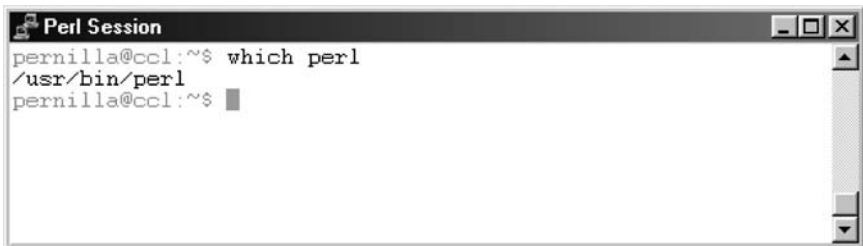


Figure 1. Command line with ‘*which perl*’

hidden from the viewer. For the computer, the differences are that when you run a Perl program, the hard work still needs to be done to convert your program code into machine code. A compiled file, on the other hand, has already been translated into machine code.

Here is a tiny program (not part of the toolbox) for test running Perl. It will only perform one task, writing the word “Typing” to the screen using the function ‘*print*’. *Print* will write to screen (unless another target is specified) whatever is given within “”-marks. Note that at the end of the print line there is a semi-colon (;). The semi-colon is required at the end of each statement line as an indicator that this is the end of the current statement, which we will use in the tools later.

```
#!/usr/bin/perl

print “Typing”;
```

Program code 1. Typing.prl

Type it in and save it as ‘typing.prl’. The programs should be typed in as a text file. Any text editor will do, but all Perl files should be saved with the extension “.prl” or “.pl”. This extension will enable us to recognize them as Perl files.

The first line of a program file should contain information about the interpreter. This was discussed above, using the command ‘*which perl*’. The information should now be put in line 1, preceded by “#!”.

```
#!/usr/bin/perl
```

Normally, the #-mark is used to mark a comment, i.e. something that should be ignored by Perl. However, when found in the very first position on the very first line in a Perl file, it states the location of the interpreter. If this line is left out, the program may still be executed but needs an additional ‘perl’ at the beginning of each command line together with the call to the program file. Compare the two commands below.

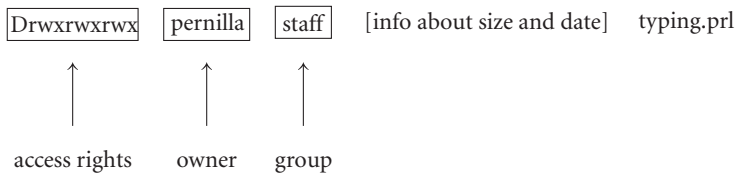
```
perl typing.prl
typing.prl
```

This is a small fact, but it is good programming habit always to include the location statement. Of course, if the programs are run on a different computer, this line may require alteration depending on where the Perl interpreter is located on the new machine.

If the computer responds with something like: ‘typing.prl: command not found’, you may not have the current directory in your path. The computer will only look for the program in the directories stated in the path, and therefore it may not find your new program. You can avoid this problem by adding “./”, which means ‘in this directory’: “./typing.prl”.

Before running the program, the access rights to the program must be executable. If you type ‘ls -l’ on the command line you will be able to view the access rights to each file in your directory on the left hand side. (Make sure you are in the same directory as your saved file. It may be easiest if all your files are in your home directory. You can always return to this directory simply by typing ‘cd’.)

Example 1. Access rights for a file

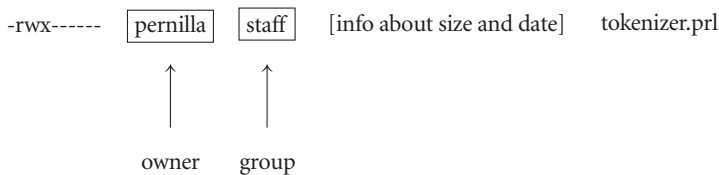


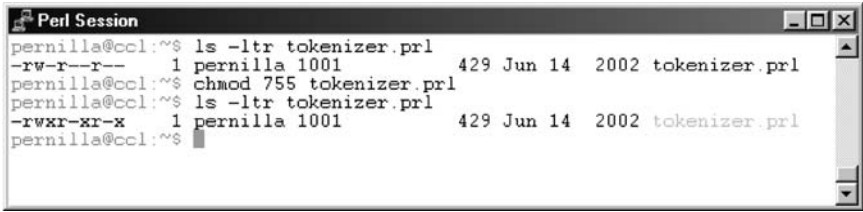
The leftmost information contains the access rights. The very first position indicates if it is a directory (‘D’). The rest of the information is in three parts with three characters in each section. The first three characters contain information regarding the owner’s rights of the file. The letter code above should be read out as:

- r=Read
- w=Write
- x=eXecutable

The example below shows a file where the owner has reading, writing and executing rights to the program. The following two sections hold information about the group rights and anyone else’s rights for the file, which in the example below are left blank, i.e. they cannot read, write or run the program.

Example 2. Owner and groups rights





```

Perl Session
pernilla@ccl:~$ ls -ltr tokenizer.prl
-rw-r--r--  1 pernilla 1001    429 Jun 14  2002 tokenizer.prl
pernilla@ccl:~$ chmod 755 tokenizer.prl
pernilla@ccl:~$ ls -ltr tokenizer.prl
-rwxr-xr-x  1 pernilla 1001    429 Jun 14  2002 tokenizer.prl
pernilla@ccl:~$

```

Figure 2. Using ‘*chmod*’ to change the access rights

To change the access right on your files, in order to make them executable, use the UNIX command ‘*chmod*’, and type in the following:

```
chmod 755 typing.prl
```

This will give everyone who has access to your computer the right to use the program, but only you will retain the right to change it. This ‘*chmod*’ operation will have to be done on each new file you create. Any introduction to Unix will explain what the 755 means, and what the other options are.

4. The Toolbox

4.1 Tool no 1: The tokeniser

This first tool is a simple *tokeniser* which is used to separate the tokens apart. The following examples are used to illustrate the point:

```

Hello!
Hello.
“hello”
hello

```

Although they may appear as four occurrences of the word *hello* and thereby four word tokens representing the same word type, this is not necessarily how a computer tool will interpret them. The punctuation and quotation marks, along with the differences in upper and lower case letters, make computer tools interpret these four instances as four independent types. While this may not seem as a problem for the linguist, many freely available part-of-speech (POS) taggers will not be able to handle such ambiguities. A tokeniser is then used as an initial step to solve these problems. Some POS taggers come with their own

tokenisers, such as TreeTagger from Stuttgart (Schmied 1994). However, your text may include problems not dealt with in the provided tokenisers.

Tokenising is often dismissed as a necessary but unimportant step; however, decisions made here may affect your corpus work in the future. If you initially use one type of tokeniser and then later run a part-of-speech tagger, which uses a different tokenisation system, the two may be incompatible. For example, how to handle “*don’t*” in the English language is one problem area. In the Bank of English these are tokenised as two tokens; “*don*” and “*t*”, where the ‘-’ mark was used as a token divider. In the British National Corpus (BNC), it is tokenized as one token “*don’t*”, while many other corpora split it into two items, “*do*” and “*n’t*”. There is no right or wrong in this classification, but it is important to know how the program you will be using expects its input.

4.1.1 *The substitute and translate functions*

Two functions will be used for the tokenising tool; the substitution and translate functions. The substitute function ‘*s*’ has the following syntax:

`s/substitutionpattern/replacementform/g;`

In this tokeniser, the substitution will be used to move all punctuation marks one position away from words in the text; commas, exclamation and question marks included. The first part after the “*s/*” is used to match with something present in the input text, using simple *pattern matching*. A string, or row of characters, will be matched against the same characters in the exact same order in the text. The second part, following the second “*/*”, gives the item for substitution.

Substitutions can become very powerful when *pattern matching* is combined with *regular expressions*. By using *regular expressions* a greater variety of substitution pattern can be covered in a single statement. For example, the word ‘*grammar*’ can be changed into ‘*syntax*’, after matching any of the following regular expressions: ‘*grammar*’, ‘*gram*ar*’, ‘*gr.*ar+*’ to mention a few.

`s/grammar/ syntax/g;`

`s/gram*ar/syntax/g;`

`s/gr.*r/syntax/g;`

The difference between the three is that the latter two will also match a great many more strings. The second example will match any ‘*gra*’ followed by zero or more *m*’s followed by ‘*ar*’ (*graar*, *gramar*, *grammar*, *grammmar* etc.). The third example will match anything beginning with ‘*gr*’ and ending in ‘*r*’. Below you find a short list of some of the most useful regular expressions.

Table 1. Useful regular expressions

text	matches "text"
.	matches any (single) character
.*	0 or more characters
.+	1 or more characters
\.	matches "."
\?	matches "?"
\!	matches "!"
[A-Z]	any (single) capital letter between A to Z
[a-z]	any (single) lowercase letter between a to z
[A-Z]*	0 or more capital letters

Our tokeniser will not make use of the '*' or the '+', which are used to match multiple characters. However, you may recognise them from some available concordancer as a way to truncate search words, where '*place.**' is used to match '*place*', '*places*', '*placed*' etc.

One of the most confusing points with regular expressions refers to the full stop. Of course, "." means a full stop. However, when used in a regular expression it is given a special meaning, namely any one character. If you want to make clear that you are referring to the full stop, and not just any one character, you must put a backslash before it.

This is true not only for the full stop, but also for characters, such as ",", "?" and "!". With the backslash we are stating that in this particular case we mean them literally and not as general regular expressions.

Unfortunately, the confusion gets worse; sometimes a cigar is just a cigar and a full stop is just a full stop... The second part of the substitute function does not make use of regular expressions. Whatever is typed here will be printed out as a substitution to the strings that match the first pattern. We tried to highlight this above by referring to the first part as a substitution *pattern*, where pattern is more general and can match more than one form in the text. The second part was referred to as the substitution *form*, the actual form that will be put in the text.

Below is a brief example of using the substitute function. The additional space before the punctuation mark (in the second field) will move them to the right in the text, creating a space between the word token and the punctuation token, thereby enabling the two to be handled separately. The line below will substitute wherever you get an actual full stop, a word space followed by the full stop; carry on to the end of the line.

```
s/\./ ./g;
```

The final ‘g’ stands for *global* and is used to prompt the function, not to stop after having substituted one occurrence, but to go to the end of the line before it stops.

The next function, the *transliterate* function, will be used to “translate” uppercase letters to lowercase letters. The function ‘tr’ has the following syntax:

```
tr/searchlist/replacementlist/;
```

Here, we will again make use of generalizations in the form of regular expression. Rather than making one line for each letter in the alphabet, we want to make a more general statement.

```
tr/A/a/;  
tr/B/b/;
```

To match all the letters in the alphabet you can refer to the whole set by typing ‘A–Z’. However, if you are working with a language that has special characters (not included in the English alphabet) then you have to specify each and every one of them after the Z. The reason for this is that the letters A to Z refer to a number for the computer and these letters are ordered sequentially (so B is one number up from A).

To change all uppercase letters to lowercase (in English), we will need the following statement:

```
tr/A–Z/a–z/;
```

The Swedish alphabet, for example, has an additional three letters after Z, “Å”, “Ä” and “Ö”. Unfortunately, Å, Ä and Ö do not correspond to the three numbers following after Z, and we cannot therefore say that it is from position A–Ö. Instead we need to specify our pattern as follows:

```
[A–ZÅÄÖ]  matching all Swedish uppercase letters.
```

4.1.2 The ‘while’ loop

Perl is line based, which enables it to read a text line-by-line. This may be stated in different ways in the perl code, but the best way for our purposes here is to make use of a *while* loop.

```
while(<STDIN>){  
    do_something;  
}
```

All programs written in this article will have a certain layout where the code is indented using tabs. This is not introduced to satisfy the Perl interpreter, but to make the code more readable to the human eye. If you are writing a loop, for example, make it a habit to use the tab to indent your text inside the loop. The curly brackets are used to denote the beginning and end of the loop, and you may find it useful to put the end bracket on a separate line. That way, when you go back to read your programs you can easily follow what is going on.

In this *while*-loop, we will read our input from the standard input (<STDIN>). The standard input refers to anything that is given in the command line. The above whileloop continues to ‘do_something’ (which we will specify later) as long as there are further lines to read in from <STDIN>. Once you start programming in Perl, you may find yourself using this *while*-loop a lot. Therefore, Perl has provided us with a short form for reading in from the standard input, simply state <> (as in while(<>)). A text or a corpus can be the input from the command line if we make use of the UNIX command “<” when calling our program.

```
tokeniser.prl < inputtext
```

When a text is being processed, the current text line is always stored in a special variable, \$_. This variable will be called upon whenever we want to print out the current tokenised line. The printing, or rather the writing to file and screen, is performed by the function ‘*print*’. Normally, a variable can be imagined as a small box, in which different values can be stored. When a variable is called upon, it will return the current value, i.e. whatever is in the box at that point. When we use variables in these programs they will be given names that explain their contents, for example, the variable name *\$keyword* if we want it to contain the keywords we are looking for.

4.1.3 Adding it all up

In the program code below you will find the *while* loop combined with the *substitution* and *translate* functions. One further function has been added, ‘*chomp*’. *Chomp* has a very simple mission, it cuts off the newline character at the end of each line. This is a simple precaution to avoid any problems with strange new-line breaks (especially when files are moved between the UNIX and Windows operating systems). The line break is reintroduced when the line is printed out; this is achieved by including the “\n” in the *print* line.

```
#!/usr/bin/perl

while(<>){
    chomp;                                #chopping off end-of-line character

    s/\./ ./g;                            #adding space before punctuation mark
    s/\,/ /g;
    s/\!/ !/g;
    s/\?/ ?/g;

    tr/A-Z/a-z;                            #converts uppercase to lowercase

    print "$_n";                          # print each word (stored in the special
                                          # variable, $_) followed by a new line, \n
}
```

Program code 2. A simple tokeniser: `tokeniser.prl`

Remember to change the file's access rights before running the tool. This is done by typing the following command line:

```
chmod 755 tokeniser.prl
```

The tool can be tested on any text file by typing this command (remember to substitute 'inputtext' for the name of your text file):

```
tokeniser.prl < inputtext
```

4.2 Tool no 2: The word splitter

The basic unit of analysis might vary from one corpus study to another. For example, you may wish to work on sentence level when you are aligning translated texts, but turn to word level when you are attempting POS tagging. Before starting to program we will need to define what our units of analysis are, so the program knows what to look for. The easiest unit is a line, since Perl is a line-based programming language. However, a line has very little linguistic value and might therefore not serve as the best unit. It may be more useful to turn to the units of words.

Before we start to implement the word splitter, we need to agree on a definition of what constitutes a word. The easiest starting point would be to state that a word is a sequence of alphanumerical characters followed by a space. If we use our tokeniser before our word splitter, then this will prove to be a good-

enough definition. The output of this word split will be a file with one word per line; a format that is a useful input to many other tools.

4.2.1 *The split function and the foreach loop*

Perl provides us with a useful function for splitting a line into words: *split*. *Split* can take many different syntaxes, dependant upon whether you wish to state the pattern, expression and upper limit of the number of splits. If you do not state anything, then the default value for pattern will be a split at every space, tab or new line. The default value for the expression will be the special variable `$_` (current line in the while loop), and the default value for the number limit will be endless.

```
split(/PATTERN/,EXPR,LIMIT)
split(/PATTERN/,EXPR)
split(/PATTERN/)
split
```

The result needs to be stored where the separated items can still be kept apart: an *array*. An *array* is an ordered list and can be recognized in Perl code by the initial `@`-sign, for example `@words`. If we examine a line including the following sentence:

“Come and visit the magnificent auditorium”,

then the words will be stored as following after using the split operator:

Table 2. The positions in a sorted array: `@words`

Come	and	visit	the	magnificent	auditorium
<code>@words[0]</code>	<code>@words[1]</code>	<code>@words[2]</code>	<code>@words[3]</code>	<code>@words[4]</code>	<code>@words[5]</code>

Although it is not important for our programs here, it is worth noting that an array starts on position 0, and not 1.

To run through every element in an array, from position 0 to the end, we have another useful loop, the *foreach* loop. This does not ask us how many items we have in our array. Instead, it will perform whatever operation we want, as long as there are elements left in the array in consecutive order. Inside the *foreach* loop, we will print the word followed by a new line, i.e. one word per line.

A one-word-per-line input is used by many programs, for example the Vanilla Aligner (Danielsson & Ridings 1997). This aligner is freely available

```
#!/usr/bin/perl

while(<>){                                # open a while loop which will read one line
                                           # at a time until the end of file

    chomp;                                # chop off the end-of-line marker
    @wordlist=split;                       # split the line into words in an array
    foreach $word (@wordlist){            # for every word in the array, do the following
        print "$word\n";                  # print the word and an end-of-line
    }                                      # end the foreach-loop
}                                          # end the while-loop
```

Program code 3. The word splitter

and can be downloaded from the TRACTOR archive (<http://www.tractor.de>) or from the Swedish Department in Gothenburg (<http://spraakbanken.gu.se>).

4.2.2 *Enhanced tokeniser*

Combining the *tokeniser* with our *word splitter* creates an *enhanced tokeniser*. The example below shows the two merged programs. Note how all the functions are stated within one and the same *while* loop. Furthermore, all substitutions and translations performed on the current line will be part of the input to the *split* function.

```
#!/usr/bin/perl

while(<>){                                # open a while loop which will read one line
                                           # at a time until the end of file

    chomp;                                # chop off end-of-line character
    s/\./ /g;                             # following four lines are doing tokenisation
    s/\,/ /g;
    s/\!/ ! /g;
    s/\?/ ? /g;
    tr/A-Z/a-z;                           # convert uppercase letter to lowercase letter

    @words=split;                          # split the line into words in an array
    foreach $word (@words){               # for every word in the array do
        print "$word\n";                  # print the word and a new line character.
    }                                      # end of foreach loop
}                                          # end of while loop
```

Program code 4. Enhanced tokeniser; tokeniser2.prl

Save the enhanced tokeniser in *tokeniser2.prl* and use '*chmod*' to change access rights ('*chmod 755 tokeniser2.prl*'). To save the output text into a new file, use the UNIX command '>'. This will write the output from the left hand side of the symbol, into the file stated on the right hand side.

```
tokeniser2.prl < text > text.tok
```

To view the new file, use the *UNIX* command ‘*more*’.

```
more text.tok
```

Below you see examples of what the program does:

INPUT TEXT:

How is infection transmitted? Through unprotected sexual intercourse with an infected partner.

OUTPUT TEXT:

how
is
infection
transmitted
?
through
unprotected
sexual
intercourse
with
an
infected
partner
.

4.3 Tool no 3: The frequency counter

Language cannot be “solved” as if it were a mathematical problem. Even so, it is often useful to view language from an arithmetical or statistical perspective. A raw frequency count can give us much valuable information about which words are recurrent around a keyword, i.e. the *collocates*. Corpus linguistics today has seen the growth of statistical methodology. Although it may be said that no single statistical method has the answer to questions about meaning, statistics can be useful for giving a snapshot of the context.

In this section, the chosen tool will provide us with a raw frequency list. With the frequency list as a base, the steps to convert this tool into a statistical calculation program are not far.

The frequency program will introduce a sub function, here named *inorder*. A sub function can be called from the main program several times. We may also argue that it makes the program easier to read if you divide tasks into subtasks.

The program begins as the tokeniser with a while loop to read in each line of the input text. Each line is split into separate words using the function *split*. This provides a good basis for the frequency count. The words are first stored in an ordered array, '@words', but as we start to count frequency the words will have to be stored in a new array. This new array will be an associative array, '\$freqlist{\$word}'; rather than an ordered array. An associative array stores the values in pairs which suits our need to store a word and its frequency. Every time we reach a certain word in the text the frequency count adds one to that word.

Example 3. Counting frequency in an associative array

```
foreach $word (@words){  
    $freqlist{$word}+=1;          # counting the frequency of each $word  
    $corpus_total+=1;             #counting the total number of words in the corpus  
}
```

When the program has worked its way through each line in the text, we need to sort our frequency list. In the subroutine *inorder*, each word is sorted based on the frequency of each word. To retrieve value from an associative array, we will use the *keys* function to produce a list consisting only of the words. To retrieve the frequency, we use *\$freqlist{\$a}*, the first pair of word and frequency in our array, and compare it with the next pair, *\$freqlist{\$b}*. The sort function goes through every pair stored in our associative array until it is all sorted.

Example 4. Sorting an associative array in order high-low

```
sub inorder{  
    $freqlist{$b}<=> $freqlist{$a}  
}  
@sorted_freq=sort inorder keys(%freqlist);
```

The difference between sorting low-high or high-low is defined by the order of the \$a and \$b. As represented above, we will sort from highest first. If you want to start from a frequency of 1, you should swap the two variables around:

Example 5. Sorting an associative array in order low-high

```
sub inorder{  
    $freqlist{$a}<=> $freqlist{$b}  
}  
@sorted_freq=sort inorder keys(%freqlist);
```

Once the frequencies are sorted in the array named `@sorted_freq`, we can print them using the *foreach* loop, introduced under 4.2.1 above. In the ordered array, `@sorted_freq`, only information about the order of the words is available. To retrieve the frequency of each word, we look it up in the associative array, as illustrated in the print line below.

Example 6. Printing information from an ordered array and an associative array

```
foreach $type (@sorted_list){
    print "$type    $freqlist{$type}\n";
}
```

The complete program is illustrated below.

```
#!/usr/bin/perl

while(<>){
    chomp;
    @words=split;

    foreach $word (@words){
        $freqlist{$word}+=1;      # counting the frequency of each $word
        $corpus_total+=1;        # counting the total number of words in the corpus
    }
}

print "Total number of words in the text is $corpus_total\n\n";
print "WORDS      FREQ\n";

sub inorder{
    $freqlist{$b}<=>$freqlist{$a}
}
@sorted_list=sort inorder keys(%freqlist);

foreach $type (@sorted_list){
    print "$type    $freqlist{$type}\n";
}
```

Program code 5. Frequency count program: `Freq.prl`

The output from the program will look like this:

Example 7. Output from `freq.prl`

```
Total number of words in the text is 447
WORDS  FREQ
the    20
.      18
```

to	13
,	12
was	12
a	11
it	11
in	9
that	9

This is a simple way of making a frequency list that has the most frequent word at the top. However, Perl on UNIX has many shortcuts and this whole program can in fact be replicated in a single command line. If we start with the one-word-per-line file which we get as output from our enhanced tokeniser, we can get the frequency list without programming any lines. This is achieved by using two UNIX commands: *sort* and *uniq*. *Sort* will first sort all lines alphabetically and *uniq* will take out all duplicates. By using *uniq* followed with a flag “-c” (which makes ‘*uniq -c*’), it will also count the occurrences. Counting the occurrences will of course mean the frequency of each word type. By using *sort* again, this time with the flags ‘-rn’, *sort* will sort based on the frequency (or sort numerically, ‘-n’) and in reverse order (‘-r’). The retrieved result is a frequency list with the most frequent words at the top in combination with the actual frequency.

There is no need not to save the file between each individual command here. Instead, use the pipe ‘|’ to combine them together. The pipe will take the output file of the previous command as the input file to the next, which is exactly what you want. The full command line is illustrated below, and the result is the same as Program Code 4.

Example 8. Result from using UNIX command ‘*sort*’ and ‘*uniq*’

```
sort text.tok | uniq -c | sort -rn | more
20 the
18 .
13 to
12 was
12 ,
11 it
11 a
9 third
9 that
9 in
```

4.4 Tool no 4: A plain concordancer

The final tool in the toolbox is a simple Key Word In Context (KWIC) concordancer. A KWIC concordancer displays a keyword with a specified number of context words on either side, called the “span”. Very often the span is set to be 4 words on either side, creating a window of 9 words (4 + keyword + 4). To obtain a correctly sized window, we will use the UNIX command, *grep*. This command you can type in on your command line and immediately receive response. *grep* matches a given pattern, the keyword in our case, in a specified file. There are versions of *grep* available for Windows as well, however, you may need to install them first.

The command below is enough to let us read all lines of “eye” in a text file entitled *text.tok*. The search pattern includes regular expressions; the ^-symbol states that nothing should come before “eye” on the line and the \$-symbol states nothing should come after.

Since it is a UNIX command, you may first try it out on the command line (not in a program file) to see that it works.

```
grep -A4 -B4 “^eye$” text.tok
```

When using UNIX system commands, as opposed to regular programming functions, in Perl scripts, the command has to be given within single backward quotation marks (‘*grep “eye” my_file*’). In order to set the size of the context window, we will use the two options *-A*, after-context, and *-B*, before context. These two options enable us to state how many words before and after the keyword are required. Using the data file from our program above, with one word per line, makes the number of lines above and below the same as the number of words in the span.

In the program below, we have made use of the operators and functions already introduced in this article: *split*, *foreach*, *print*. However, apart from *grep*, only one further new part is introduced. This is a special array entitled ‘@ARGV’. @ARGV contains the list of filenames or, in our case, list of keyword, span number and text file, from the command line. The @ARGV enables us to state our variables when calling the program, for example:

```
conc.prl eye 4 text.tok
```

In the above command, the computer is instructed to make a concordance using the program “conc.prl”, of the word “eye” in the file “text.tok”, with a span of four words on either side of the node word “eye”. Note that I ask for the span number only once, and this number will be used with the *grep* command, both

for –A4 and –B4. Between every retrieved dataset, the *grep* command will produce a ‘—’. This indicates end of the span. We will make use of this to split the input into separate concordance lines stored in an array named ‘@lines’. This is illustrated below in the line ‘@lines=split(/–/, \$input);’.

```
#!/usr/bin/perl

($keyword, $number, $file)=@ARGV;

$input=`grep -A$number -B$number "^${keyword}" $file`;
@lines=split(/–/, $input);

foreach $line (@lines){
    @words=split(/\n/, $line);
    foreach $word (@words){
        print "$word ";
    }
    print "\n";
}
```

Program code 6. A KWIC concordancer

The following example illustrates a KWIC concordance made on the word ‘*corpus*’. The text file used in the example was first run through ‘*tokeniser2.prl*’, saved into a new file, and then recalled on the command line for the ‘*conc.prl*’ programme.

conc.prl corpus 4 text.tok

Note that in calculating the span, punctuation marks and dashes are counted as if they were words.

Example 9. The output from our concordancer: *conc.prl*

with the topic “multilingual	corpus research” bansko, bulgaria
years title, “multilingual	corpus research”, we would
especially the plato multilingual	corpus, available from the
we invite papers on corpus-based/corpus-driven research including, but	
translation equivalence – contrastive	corpus linguistics – the semantics
of corpus linguistics –	corpus software a detailed abstract

Now that you have your four programs, you may combine them in any order you like. For example, a concordance file like the one above can be used as input to your frequency count program. The result will be a frequency list of

the collocations around the specified keyword within the span you selected for the concordance. If all you want is the frequency list, you do not need to save the output file from the concordance program. Instead you can *pipe* the files through to the next program, i.e. let the output of the concordance program be the direct input to the frequency program. *Piping* is performed by the use of the “|” (vertical bar), see example below.

```
conc.prl corpus 4 text.tok | freq.prl
```

5. Conclusion

The four tools in this toolbox are to be seen as a starting point. They can easily be improved and altered to cover also other similar tasks. To continue the pursuit of Perl programming, there are several good books.

Learning Perl (Schwartz & Phoenix 1992) is a good starting point, which takes you through the basics of programming. If you prefer Windows operating system to UNIX, you may be pleased to hear that there is a book focusing on programming Perl on Windows: *Learning Perl on Win32 Systems* (Schwartz et al. 1997).

Once you have learnt the basics, the next step is found in the books that list possibilities rather than explain basics, for example, *Programming Perl* (Wall & Schwartz 1992), which is a good reference book listing most of the functions and operators available in Perl. If you want a collection of tiny coding snippets to be inspired from, then the book to buy is *Perl Cookbook* (Christiansen & Torkington 1998). Once you start programming, you will find that simple programming is not so hard to achieve.

Note

1. The programming language C is probably the most widely used programming language to date. It has the benefits of giving the programmer maximum control and efficiency; however, the code can be rather cryptic at times.

References

- Christiansen, T. & N. Torkington (1998). *Perl Cookbook*. Sebastopol, CA: O'Reilly & Associates.

- Danielsson, P. & D. Ridings (1997). Practical presentation of a 'Vanilla' aligner. In U. Reyle & C. Rohrer (Eds.), *The TELRI Workshop on Alignment and Exploitation of Texts*. Ljubljana: Institute Jozef Stefan.
- Ide, N. & J. Veronis (Eds.). (1995). *The Text Encoding Initiative: Background and Context*. Dordrecht: Kluwer Academic.
- Sperberg-McQueen, C. M. & L. Burnard (2001). TEI P4 guidelines for electronic text encoding and interchange, XML compatible edition. TEI Consortium (<http://www.tei-c.org/P4X>).
- Schmied, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing* (pp. 44–49). Manchester.
- Schwartz, R. L., E. Olson, & T. Christiansen (1997). *Learning Perl on Win32 Systems*. Sebastopol, CA: O'Reilly & Associates.
- Schwartz, R. L. & T. Phoenix (1992). *Learning Perl*. Sebastopol, CA: O'Reilly & Associates.
- Wall, L. & R. L. Schwartz (1992). *Programming Perl*. Sebastopol, CA: O'Reilly & Associates.

Network

Learner oral corpora and network-based language teaching

Scope and foundations

Pascual Pérez-Paredes

Supposing is good, but finding out is better

Mark Twain in *Eruption*; Mark Twain's autobiography

Although learner corpora are said to be less suitable for input-orientated Data-Driven Learning (DDL) applications, it is debatable whether purely language input and *learning* input should not differ. Networked learner oral corpora present ample opportunities for DDL that is based on awareness-raising as they stand out as powerful tools for the creation of situations where learning is organized around students' analysis of language data which leads to language-interlanguage knowledge restructuring. This paper examines the relations between Learner Oral Corpora (LOC) and Network-Based Language Teaching (NBLT), and stresses the need for institutions to develop their own LOC functional taxonomy. An analysis of the scope and foundations of a computer-mediated learner oral corpora integrative implementation cycle is presented in terms of a contribution to DDL methodological applications.

Introduction

In a recent work (Pérez-Paredes 2003) I outlined the scope of a paradigm which integrates the functions of Learner Oral Corpora (LOC) within a Local Area Network (LAN) and enables them to be implemented.¹ I called this paradigm a Computer-mediated Learner Oral Corpora Integrative Implementation Cycle. Following a tradition which can be traced back to the times of the audio-lingual language laboratory heyday (Locke 1965), it is my belief that the *integration* of software and hardware tools and Learner Oral Corpora should be a priority in certain foreign language learning contexts, particularly where

a cycle which includes data gathering, research, teaching, assessing and needs-analysis is implemented. At the University of Murcia, Spain, a cycle of such characteristics has been put into operation in the language laboratory facility at the Faculty of Arts.² Although much of the work carried out in this facility is presented in Pérez-Paredes (2003), in this paper I particularly want to examine the relations between Learner Oral Corpora and Network-Based Language Teaching (NBLT).

A word of caution is required at this point. The following discussion is based on the use of infrastructures, both human and material, which are typically found in western and Asian universities. Its application is therefore very restricted as is most of the literature on technology and learning.

1. The scope of Network-Based Language Teaching (NBLT)

Network-Based Language Teaching is seen by many as a natural extension of Computer Assisted Language Learning (CALL). Hughes (1997) points out that a networked computer classroom eventually results in better management and argues that in order to exploit machines efficiently and cost effectively it is important to link them in a network. Hoffman (1996) goes a little bit further when he defines computer networks as linkages of two or more computer workstations for the sharing of software, data, peripheral devices and the provision of communication. Essentially, this approach is particularly aimed at laying stress on aspects which affect data shareability and network organization. In a way, computer networks are perceived in Hoffman's CALL-oriented discussion as management solutions as their organizational aspect is significantly underpinned.

Other approaches (Kelm 1992, 1996; Gu & Xu 1999), in contrast, are farther-reaching as they go beyond in terms of methodological implications. Kern and Warschauer (2000) define Network-Based Language Teaching as the use of computers connected to one another in either local or global networks for the teaching of languages. The emphasis, in this case, is laid on the communication facilities that networks present. In fact, the use of computer networks has extended the role of CALL over the last three decades (Chapelle 2000) into a whole new array of computer-based learning events which were just undreamed a few years ago. In this new century, one expects that linguistically enriched environments such as networked mediated multimedia, that is, language learning environments with high levels of information and data retrieval, will definitely help rethink the pre-network conceptualisation of

CALL.³ This definition process is still underway as the presence and impact of computer-mediated literature is very recent, especially when compared to more traditional learning approaches.

In essence, CALL practice and research have focused their attention on a great variety of topics such as Second Language Acquisition (SLA), conversation analysis, implementation procedures, CALL as a socio-cultural model for language learning and teaching, language learning immersion theory, learner autonomy and critical ethnography (Levy 2000). However, the relevance of Network-Based Language Teaching in CALL research is far from being significant at all. When Levy (2000) outlines the keywords used in CALL research throughout 1999, it is noteworthy that NBLT or related constructs are not present in his corpus consisting of all the major CALL journals available (*CALICO*, *CALL*, *LL&T*, *ReCALL* and *System*) and relevant CALL reference books (Cameron 1999a; Cameron 1999b; Debski & Levy 1999; Egbert & Hanson Smith 1999).

(CALL) applies research from the fields of second language acquisition, sociology, linguistics, psychology, cognitive science, cultural studies, and natural language processing to second language pedagogy, and it melts these disciplines with technology-related fields such as computer science, artificial intelligence, and media/communication studies.⁴

Despite the assumption that CALL is multidisciplinary and integrative, NBLT is not on the research agenda. The fact that Network-Based Language Teaching is absent from this picture may appear to clash with the notion that CALL, as an emergent discipline, favours ground-breaking initiatives in teaching and learning development and research. All things considered, I tend to believe that the role of Network-Based Language Teaching should be better understood as an umbrella notion which extends the scope of pre-digital networked CALL, that is, CALL practices which range from computer assisted instruction such as the well-known Programmed Logic for Automatic Teaching Operations (PLATO) to speech applications (Salaberry 2001).

Kern and Warschauer (2000:11) suggest that NBLT takes advantage of both computer-mediated communication and the newest delivery of data and communication for language learning via globally linked hypertext. The combination of these two powerful features may, eventually, shape new language education paradigms:

CALL researchers may stress one or more of these areas, as they engage in systematic inquiry seeking to discover new information, create or revise theories, and develop learning tools. Depending on the needs and goals of the project

or institution, pedagogical, budgetary, or student needs may drive the search for new technology-based materials and improved instructional approaches. Development of these CALL solutions leads to new practical applications and to additional research.⁵

This final statement on the development of solutions is evident in the case of the use of corpora in Foreign Language Teaching (FLT). As an example, we can see how users of Learner Oral Corpora, researchers and compilers, are given the opportunity to record truly communicative non-monitored students' performances. This *learner* foreign language can be the result of a wide range of different communicative interactions such as individual monologue-like texts, face-to-face pairwork, telephone pairwork, face-to-face groupwork, telephone groupwork and whole classwork interaction. Furthermore, Chun (1994) has shown that computer-assisted class discussion provides students with the opportunity to acquire and practice more varied communicative proficiency. On top of these advantages, students' data can be stored in different digital formats.

Before dealing with the relationships of Network-Based Language Teaching and Learner Oral Corpora more extensively, it is important to distinguish here between two possible major approximations to using computer networks for the teaching of foreign languages: (1) the technology-enhanced language e-learning approach and (2) the technology-delivered language e-learning approach.⁶ The first approach is characterized by the presence of teachers in the computer rooms in real time and instructor-led sessions, while in the second, also known as Distributed Education or Distance Learning, the learners are never in physical proximity to the instructor. The figure of the tutor is substituted by non-real and real time virtual interactions delivered through a blend of asynchronous and synchronous technologies (Jackson 2001). While technology-enhanced learning has been under the influence of computer assisted instruction in the past, the second approach is at the heart of purely distance e-learning postulates based on the use of internet-related technologies. The type of work proposed in this paper falls within the technology-enhanced learning context, as the type of learner corpora I advocate result from students' communicative interaction within the formal classroom learning situation.

Despite the overlapping terminology which is found in the literature, *e-learning* is mostly associated with learning activities which involve computers and interactive networks simultaneously. Although the computer then does not need to be the central element of the activity or provide learning content, the computer and the network must hold a significant involvement in the learning activity. On the contrary, web-based learning is associated with learning mate-

rials delivered in a web browser, including when the materials are packaged on CD-ROM. As NBLT may encompass different notions of networking practice, for the purposes of corpus usability described in this paper, I will adapt from Rosenberg (2001) a broad, moderate view on e-learning which integrates the following principles:

1. E-learning is networked: put simply, more resources are used more efficiently. The key word that facilitates empowering in this new environment is *digital*. However, every learning organization should ideally define its own policy on e-learning and how knowledge management, electronic performance support and learning directions are characterized.⁷ The level of expertise in CALL instruction and senior management support is also to be considered within this definition.
2. Internet and/or LAN-based technology facilitate end-user access to contents: better accessibility implies facilitated input.
3. Learning is understood in a very broad sense, going beyond the traditional paradigms of training: e-learning based on NBLT practices is an integrative paradigm where communication events based on different concepts of language learning -structural, cognitive and sociocognitive postulates- easily coexist. Kern and Warschauer (2000) have pointed out that the true nature of NBLT is best exploited in sociocognitive language learning where the learners' focus is brought towards the use of language in social interaction. The structural and the cognitive approaches to language learning stressed the transmission of contents from competent users and the operation of innate cognitive heuristics on language input, respectively.

When NBLT is founded on technology-enhanced language learning approaches, which has been the case in most institutions where old language labs have been converted into computer-based language labs, the following further features should ideally characterize the learning environment:

- Information and data can be updated instantly.
- Information and data can be stored and retrieved instantly.
- Information and data can be distributed and shared.
- TCP/IP network protocols and web browsers create a universal delivery platform of digital information.⁸
- Information is delivered and tools provided that are aimed at improving students', teachers' and researchers' performance.

These features intend to facilitate the design, compilation, analysis and delivery of corpora contributed by the students. Otherwise, I imagine the task of so doing would be overwhelming or simply unapproachable.

2. The scope of Learner Oral Corpora (LOC)

So far, language corpora have been available to individual users mostly in academic institutions for research purposes. Bearing in mind the characteristics of computer networks in language teaching outlined above, I suggest that learner corpora take up a more active role and be made available to students themselves to open up the learning possibilities that posit their own examination of *interlanguage*. This is precisely where NBLT comes in.

Learner Oral Corpora differ substantially from other language corpora in, at least, two important aspects. On the one hand, they feed on foreign language students' output as this is the primary target of learner corpora, that is, the language used by students. On the other hand, the use of this output may possibly have a considerable impact on teaching and general methodological decisions.⁹ In this sense Granger (1998b) highlights three areas which, potentially, benefit from learner corpora-based research: curriculum design, materials design and methodology adjustments. This being so, it is sensible to think that, among other initiatives, teachers can build learner corpora that can contribute to changes and adjustments in their methodology, particularly changes concerning those aspects that are more directly connected with developing students' oral skills. However, I would add a fourth area which NBLT favours: networked LOC can be conveniently used to promote students' language awareness on different aspects of learners' FL production. This potential for change has previously been underpinned by Carter (1993). Tognini Bonelli (2001:42–43) has also pointed out how the introduction of corpora in the language classroom will bring about changes in the way the different participants in the learning experience will interact:

In order to classify the new type of evidence (...) teacher and students will have to become co-workers, research collaborators, and the difference in status endemically present in the activities of teaching and learning levels out.

LOC, either compiled or delivered through local networks, appear to be feasible in diverse fields: first, as assessing tools; second as synchronic feedback information from an individual, a group or groups of students, and third as diachronic feedback information from a group or groups of students. In this way

the enriched data information that LOC provide can reinforce teachers and students' perceptions of learner oral language, increasing what Skehan (2001) and Schmidt (1990) call *noticing*.¹⁰ This notion is central in foreign language instruction as it precedes language change and performance development and improvement.

One of the most interesting features of all three domains presented above is that they can interact and offer valuable and extensive information for FL teachers. Simultaneously, LOC used by learners themselves can be enormously practical as awareness-raising tools. This awareness comes basically from (1) contrasting one's own oral production with others'; (2) detaching oneself from one's oral output; (3) establishing a new type of insight into lexical and performance oriented facts, such as the commonest word forms in a student's interlanguage, the central patterns of usage or the combinations learners typically form (Tognini Bonelli 2001:40). Finally, (4) (inter)-language awareness may come from reviewing bookmarked points in the flow of discourse where, for example, strong forms are overused, segmental pronunciation is wrong or discourse organization is faulty (Pérez-Paredes 2003). Bookmarking an audio file is similar to bookmarking a website while you visit it: you can easily return to it later with a simple mouse click, rather than having to remember or type a very long or sometimes cryptic URL. In a way, bookmarks in multimedia learner corpora play the role of tags in standard corpora. Digital players like DIVACE DUO offer this facility (see Pérez-Paredes 2003 for further details).

To fully appreciate the complexities associated to LOC use, we believe it is essential for any language learning organization to approach their own taxonomy of LOC which accounts for the participants and motivations in using this type of corpora. As an example, I offer a chart with an LOC taxonomy which is instrumental in my practice as a language teacher. All types of corpora are collections of texts contributed by students of English as a Foreign Language (EFL).

In this particular taxonomy, there are five types of corpora (LOC 1 to LOC 5), which, whether teacher-controlled or student-controlled, serve different purposes. In our case, these are linked to assessment, research and awareness-raising but, obviously, purposes can be refined and adjusted to teachers' needs in different teaching institutions. Different types of corpora yield different compilation and networking delivery options. Although all the students' oral productions are recorded in the computer facility, they present important differences. While LOC 1 are usually the product of an oral interview with the teacher where students are given a closed set of topics to talk about, and accordingly it is obvious that learners are influenced by a potentially stressful situation,

Table 1. Functional LOC taxonomy

Type of corpus	Operated by teachers	Operated by students	General purpose of the corpus
LOC 1	X		Assessing students' oral production
LOC 2	X		Investigating students' oral production from a synchronic perspective
LOC 3	X		Investigating students' oral production from a diachronic perspective
LOC 4		X	Raising L2 awareness through introspection into one's own oral production
LOC 5		X	Raising L2 awareness through introspection into a whole group's oral production

LOC 3 are based on oral presentations on the same topic which are recorded at different points in time. Sample presentation topics include immigration, the environment or the description of a picture or film. LOC 1 through LOC 3 are used by the teacher for assessment or research purposes. On the other hand, LOC 4 and LOC 5 are typically stored in the lab local area network and can be accessed by individual students who want to examine their own spoken output (LOC 4) or by a group of students examining the productions of the whole class (LOC 5). In a recent work (Pérez-Paredes & Cántos-Gómez 2002) we have outlined the type of work these students carry out with these corpora and the guidelines and materials they are presented with.

3. Strategic foundations of LOC in networked environments

In modern societies we see that new approaches to e-learning have appeared that are based on new strategies, both instructional and informational. The new online training and learning opportunities and the new notions on knowledge management are just a token of this new interest.¹¹ Learning architectures (Rosenberg 2001) have also started to change as new infrastructures have made it possible for the FL community to start up new directions in computer assisted language teaching, learning and research.

Organizations now have technological capabilities for the delivery and management of learning that were unimaginable a decade ago. This has had a great impact on the learning culture of students and the “teaching” culture of teachers. Research, of course, is no exception to this and has seen how its praxis has been altered. For example, in the world of business and econ-

omy, the concept of management has witnessed a radical redefinition of its role. These days, organizations understand that success is linked to learning support strategies which embrace Information and Communication Technologies (ICTs) and computer mediated communication. Rosenberg (2001), in this sense, thinks that old students' days of tuition revenue are gone forever. New adapted organizations, and the type of learning they promote foster models which support, rather than limit, the growth of e-learning (whatever the view on e-learning which is adopted). Why should language teaching and learning remain impassive?

New opportunities in teaching and learning based on ICTs are becoming more apparent as western societies opt for a new model of organizing the way knowledge is acquired and delivered (Terceiro & Matías 2001). This paradigm presents important advantages to the FL community. Esch and Zähler (2000) have pointed out the presence of three relevant features in ICTs, and consequently in computer networks, which make them perfect vehicles for language learning. Put briefly, individual learner characteristics (Skehan 1989), especially aptitude and motivation, adapt well to the learning interactions which are deployed in computer networks suitable for the learning and teaching of languages, ranging from monologues to group-discussion with students who are not sitting close to each other. In this sense, research shows that students who would be inhibited in face-to-face communication do get more involved in communication in networked environments (Warshauer & Kern 2000). Knowledge construction of the type which is derived from the work with language corpora is similarly facilitated through ICTs as rich information, that is, information which integrates text and audio, can potentially be adapted to students' learning styles and preferences. Esch and Zähler (2000: 10) use the case of concordances to make this point:

use of concordances to analyse a corpus makes learner control available as it is possible to approach the linguistic information from different angles or levels which reflect more the investigator's approach to the corpus than the software structure.

Finally, learning languages through or with ICTs has an added value: "socially constructed reality is the conceptual glue that holds societies together and much of our knowledge is extrinsic, collective shared knowledge" (*ibid*).

Given the networked society we are living in, networked contents and language learning delivery via LANs, Wide Area Networks (WANs) or external networks such as those which run on HTTPs naturally fit in an increasingly digitally-oriented culture.¹² This new phenomenon is well rooted in both

technology and economy and permeates most social human activities. As a manifestation of that general paradigm, this computer-mediated learner oral corpora integrative implementation cycle presents a major breakthrough in the Data Driven Learning (DDL) methodological applications. The reasons are varied and among others I have outlined the following (Pérez-Paredes 2003):

Shareability: users of Learner Oral Corpora (students, teachers and researchers) share the same environment and tools.

The perfect integration of a multi-purpose environment and tools favours a natural introduction of corpora in foreign language instruction. This way, corpora can be incorporated into everyday lessons and into the learning context itself. Lock and Tsui's (2000) experience is worth considering as the use of corpora as an integral part of their teachers' network in Hong Kong has opened up new directions in the application of general functional grammar in the teaching of English. Again, the key issue here is integration.

Methodology-independence: technologies should always support the curriculum, not determine it. Different learning methodologies fit well in this paradigm as long as communication is a crucial ingredient of the learning experience since the data gathering process is built in the very learning and teaching environment.¹³

Learner autonomy: one of the applications of networked LOC is the possibility for students to access the data simultaneously and to do this in different autonomous environments.

These four foundations are based on a SWOT (Strength – Weaknesses – Opportunities – Threads) analysis performed on students' oral language productions and how these relate to the methodology and classroom management.¹⁴ The acronym above is easily understood when we examine the features of the organization where students learn to speak and master the L2. In the case of the University of Murcia, Spain, those characteristics were analysed in the following terms:

- Strengths (SWOT): formal L2 instruction which accommodates an official, academic curriculum.
- Weaknesses (SWOT): students' groups are too crowded and situations for meaningful interaction are not abundant.

- Opportunities (SWOT): Computer Mediated Communication and the gathering of LOC could help the FL community gain insight into the nature of the interlanguage used by students.
- Threats (SWOT): the learning situation either continues to be the same or further deterioration can be expected.

The opportunities chapter is crucial as it highlights those areas where headway could be made if the strengths were played up and weaknesses played down (Waters & Vilches 2001). In this sense, the use of corpora in FL instruction is highly adaptative in terms of reinventing both the learning organizations we work in and the learning experiences we create as language teachers and linguists. These four foundations currently guide my use of learner corpora with the advanced students of English as Foreign Language at the University of Murcia. These learners see in corpora a natural way to explore their own oral productions in exactly the same way as they do with their writing after being marked. Transcriptions of selected output and the corresponding audio files are stored in the same classroom which is used for their everyday lessons, which enormously facilitates the students' process of familiarization with the environment.

The old language teaching paradigm is heavily dependent on textbooks and teachers as learning mediators. In this traditional banking model teachers deposit knowledge into students who, in turn, are expected to duplicate behaviour out of contact with the teacher and observation of his behaviour.¹⁵ Kennedy (1991) describes the traditional model of teaching and learning languages as the grammar paradigm. This paradigm is, in effect, lexis bound and neglects oral discourse for three reasons: it is more orientated towards paradigmatic choices, it presents a focus on written varieties of the L2 and usually presents, if not a normative approach, certainly a bias towards the language and linguistic information provided.

As a way of contrast, a methodology that integrates Learner Oral Corpora will surely find itself at ease within the new knowledge paradigm as independent learning and form-focused, meaningful instruction clearly adapts successfully to the new learning approach. Johns (2000) finds that new learning and teaching trends in the 21st century will place "emphasis on fitting the corpus to the learner: on alternative approaches such as Reciprocal Learning; and on the potential of the Web in defining and supporting a 'worldwide DDL community'" (quoted from conference abstract). All three trends, and this is fascinating enough, converge in networking LOC.

4. The digital bridge

As already outlined, Kern and Warschauer (2000) believe that technologies which support NBLT not only serve the new and emerging teaching and learning paradigm but also help build it. One question remains unanswered, though: how are corpora, in particular LOC, being shaped by such technologies? And equally interesting, how are LOC going to help define the new paradigm? If one attempts to understand the issue, first the areas of application of such a paradigm must be framed.

Computer networks are formed by workstations which are interconnected through different hardware and software. Similarly, communication protocols vary from network to network. In essence, more than the quantity of computers being connected or the technical peculiarities which condition their configuration as an entity, we should examine carefully the use these computers may be put to in terms of language instruction and, in our case, LOC manipulation procedures. Considering the existing technology and its possible applications, I find that there exist three types of language learning computer network of interest for the work with learner oral corpora.

First, we find networks where computer assisted instruction has served a structural notion of CALL. In these networks most software is courseware, that is, guided instruction with little room for language students' roles as researchers. In these networks, teachers would have little control over the delivery of the corpus material. Furthermore, communication among users and teachers is not even contemplated. Also, these pre-digital networks allow very little room for corpus compilation and implementation other than single workstation monitoring and reading. It is highly unlikely that in this network environment LOC could be transformed into any sort of interactive customized courseware. An informative purpose could, however, contribute to contrastive interlanguage analysis, either instructor or student led, of the type of work proposed by Granger (1998a).

Second, technology-enhanced networks will exemplify the four foundations which were advanced previously. Communication is the key issue here for users. A truly enriched language environment is presented on their computer screens as LOC use integrates a very extensive typology of students' interactions along with audio and, if feasible, video, transcription facilities and strategic information such as bookmarks on recorded files. The digital format facilitates this integration and ease of manipulation. In this environment, LOC may have both informative and formative roles. Networked delivery of multimedia

and interaction with rich data will require a somewhat steep learning curve for teachers and students. Once this is fixed, the benefits are obvious.

Third, the technology-delivered network is based on both web-based Instruction (WBI) and computer-based training (CBT).¹⁶ Of course, postulates about technology-enhanced networks also apply here although the emphasis is laid on distance learning and autonomous work. For understandable reasons, distance learners are unlikely to contribute to communicative activities in real time and hence data gathering has to be reconsidered until broad bandwidth delivery is made available to most potential learners. Notwithstanding, worldwide access to LOC-based materials is of great interest and relevance to students, teachers and researchers in general. Online learning events and Internet-based courseware has begun to be developed to access corpora and new opportunities are being envisaged.¹⁷ The MICASE corpus at the University of Michigan is an excellent example of an L1 oral corpus which can be accessed through the web in text format.¹⁸ The second and third types of language learning computer networks partake in McEnery and Wilson's potentially beneficial uses of general corpora as they very clearly favour the integration of purposes, access, an enriched-data context and, most outstandingly, the collection of naturalistic learner language.¹⁹

The need for a paradigm like the computer-mediated Learner Oral Corpora integrative implementation cycle proposed here becomes even more evident when one reads Hoffman's (1996) remarks on the future of networks in FLT. He believes that it is necessary that some requirements are met before *net* benefits are exposed to a big audience of professionals. Those requirements include the implementation of reliable networking standards, the improvement of the quality of interfaces and, last but not least, the appearance and consolidation of creative ways for teachers and learners to use and exploit the potential of networks. I firmly believe that the time has come for language-learning networks to be an essential ingredient of FL instruction and, coincidentally, for the role of corpora to expand in the language classroom. Let us find out how this can be done.

These days, networking standards have been incorporated into major operating systems that allow even non-experts the interconnection of stations. Ethernet networking is, no doubt, the most popular networking technology available; data transmission is fast, inexpensive if the computers are in the same room, very easy to maintain and set up, expandable in terms of the unlimited number of extra devices that can be connected and, primarily, very reliable. Other possibilities are not so readily available or cost-effective, for example phone-lines, which are very slow, power-lines and wireless ethernet.

However, online audio delivery is, disappointingly, best accomplished through hardware, rather than software which, too often, turns unstable. Although the situation is likely to change shortly, an audio network is the most reliable and cheapest solution to guarantee stable capture, delivery, storage and retrieval of audio files in the framework of networked LOC.

User-friendly interfaces in e-language (lab) networks are within the means of even low-budgeted academic institutions. These professional customized solutions are the most convenient way to start up technology-enhanced e-language learning network-based instruction and, accordingly, the easiest way to implement a paradigm like the one presented here. Ideally, these solutions should be instrumental and transparent as regards technology so that the focus of the students and teachers praxis rests upon L2 communication and learning. I believe that it will not be too long before this is accomplished and stable software and hardware solutions allow jointly for drastic improvements in data shareability and multimedia management.

The e-language network of the future should fade away as a technological entity, and give way to an infrastructure easy to customize that should be in charge of sustaining environments where better, richer learning conditions are provided. Hoffman (1996: 60) stresses precisely this feature of computer networks in language learning:

The important point is that computer networking enriches and expands the opportunity to learn and use the target language naturally and with a communicative purpose.

Along with Hoffman, I see in communication and in the impact on language learning the two most outstanding features of networked LOC; the first, communication, because through networks students are offered more freedom to communicate,²⁰ more collaboration of the type which is found in everyday situations where no direct monitoring or supervision is imposed on speakers,²¹ more diversified interaction and, as a corollary, more opportunities for the negotiation of meaning. This communication framework will favour more realistic accounts of learner language and a more comprehensive picture of students' interlanguage, especially of contexts other than interviews, and, finally, cheaper, easier ways to integrate rich multimedia digital files.

In turn, the impact of networked LOC on language learners can be exemplified through general assumptions on the advantages and benefits of corpus linguistics.²² LOC data do certainly (1) open up doors to the objective verification of L2 production, (2) bring about a better insight into the students'

interlanguage competence and (3) offer quantitative information to linguists and SLA researchers.

At the University of Murcia, a COLOCINC paradigm has been set up that is fundamentally oriented at raising students' awareness of their own language oral output through the examination of naturalistic data recorded during interactive communication activities in the language laboratory. These data can be further transformed into a wide range of study patterns ranging from word lists to digital audio files with meta-information such as bookmarking, including concordance lines or different word/lemma contexts. Besides, this exploration is suitable for both divergent and convergent learning activities (Leech 1997). Each student can save his or her own database of use and compare his or her portfolio with other fellow classmates'. This type of work, suggested by Tognini-Bonelli (2001) or Skehan (2001), builds upon the notion of *noticing* mentioned above and the idea that discourse manipulation is beneficial to students. Besides, research suggests that the transcription of students' output favours form-focused learning (Lynch 2001).

Although learner corpora are said to be less suitable for input-orientated DDL applications (Kaszubski 1998), it is debatable whether purely language input and *learning* input should not differ. Networked LOC present ample opportunities for DDL that is based on awareness-raising as, given the technology available, they stand out as powerful tools for the creation of situations where learning is organized around students' analysis of language data which leads to language-interlanguage knowledge restructuring (Barlow 1998). This general approach coincides closely with Pennington's (1996: 1) view on the general role of CALL:

The power of CALL in language learning and language teaching is to introduce new types of input, from both a quantitative and a qualitative perspective. The added quantity of input leads to a richer language learning environment, while the unique quality of CALL input means different possibilities for accessing and developing information.

Bringing computer networks, methodological approaches and LOC together, we can bridge the gap between technology fascination, computer-mediated communication and effective language learning (Chapelle 2001b), providing the FL community with the type of insights and understanding which derive from corpus-based studies and the functionalities of computer networks in language learning environments.

Notes

1. For the purposes of this paper, *Learner Oral Corpora* will be understood as collections of texts contributed by students of a foreign language engaged in communication activities. These corpora consist of both transcriptions and their corresponding digitised audio files. For a discussion on learner corpora see Granger (Ed.) (1998a). A LAN is a group of computers and associated devices that share a common communications line. These computers may share a single processor or server within a small area such as a classroom, a language laboratory or a building. In a typical LAN, the server stores applications and data which is shared by multiple users.
2. See <http://www.um.es/engphil/lab>, consulted: 27.11.03.
3. From a purely technical perspective, in these contexts the concern is managing resource usage so that the quality of service requirements of the retrieved continuous-media data can be met. This is of relevance in multimedia language learning which integrates different types of text, audio and real-time communication in the classroom.
4. Scholarly activities in computer-assisted language learning comprise development, pedagogical innovations, and research. Joint policy statements of CALICO, EUROCALL, and IALLT, arising from a Research Seminar at the University of Essen, Germany, 30 April-1 May 1999 can be found at http://calico.org/CALL_document.html (as of November 12, 2001).
5. Ibid.
6. See Spector & Davidsen (2000) and Rosenberg (2001).
7. "Knowledge management supports the creation, archiving and sharing of valued information, expertise and insight within and across communities of people and organizations with similar interests and needs. Many knowledge management systems are facilitated by Internet technologies". Rosenberg (2001:60).
8. TCP/IP are the set of protocols that make FTP, e-mail, and other services possible among computers that don't belong to the same network. However, TCP/IP protocols can be used in LANs to develop an intranet or to exchange information.
9. Basturkmen (2001) has pointed out that oral production and analysis is neglected in the foreign language classroom.
10. Kramsch and Andersen (1999) maintain that in multimedia environments a discursive gap is created by the powerful way the computer has of both imitating and representing life. For them multimedia re-enacts the original, lived context in which language was used and transforms it into readable "discourse". The gap between these two processes creates the communicative challenge "par excellence".
11. See note 7.
12. A WAN is a computer network which spans great distances, connecting different LANs together. HTTP stands for hypertext transfer protocol.
13. See Pérez-Paredes (2003) for details.
14. See Gołębiowska (1990), McGinity (1992) and Pérez-Paredes and Mena (2001) for in-depth discussions on oral communication management in FL classrooms.

15. See Tom Heany's webpage on Freirean pedagogy. URL at <http://nlu.nl.edu/ace/Resources/Documents/FreireIssues.html> (as of November 1, 2001).
16. Web-based instruction is a hypermedia-based instructional program which uses the attributes and resources of the World Wide Web to create a meaningful learning environment where learning is fostered and supported. Computer-based training uses a computer as the focal point for instructional delivery. Note that the stress of CBT is laid on *delivery* and not on communication exchange.
17. See for instance Lock and Tsui (2000).
18. See <http://www.hti.umich.edu/m/micase/>, consulted: 27.11.03.
19. McEnery and Wilson in ICT for Language Teachers. See <http://www.ict4lt.org>, Module 3, Chapter 4.
20. See Kelm (1992).
21. See Kelm (1996).
22. McEnery and Wilson in ICT for Language Teachers. See <http://www.ict4lt.org>, Module 3, Chapter 4.

References

- Abbey, B. (Ed.). (2000). *Instructional and Cognitive Impacts of Web-Based Education*. Hershey: Idea Group Publishing.
- Aijmer, K. & B. Altenberg (Eds.). (1991). *English Corpus Linguistics. Studies in Honour of Jan Svartik*. London: Longman.
- Barlow, M. (1998). A new paradigm for teaching and language concordancing. Paper presented at the Teaching and Language Corpora Conference. Keble College, Oxford, 24–27 July.
- Basturkmen, H. (2001). Descriptions of spoken language for higher-level learners: The example of questioning. *ELT Journal*, 55(1), 1–10.
- Berglund, Y. (1999). Exploiting a large spoken corpus: An end-user's way to the BNC. *International Journal of Corpus Linguistics*, 4(1), 29–52.
- Cameron, K. (Ed.). (1999a). *CALL: Media, Design and Applications*. Lisse: Swets and Zeitlinger.
- Cameron, K. (Ed.). (1999b). *CALL and the Learning Community*. Exeter: Elm Bank Publications.
- Carter, R. (1993). Language awareness and language learning. In M. Hoey (Ed.), *Data, Description, Discourse* (pp. 139–149). London: Harper Collins.
- Chapelle, C. (1997). CALL in the year 2000: Still in search of research paradigms? *Language Learning & Technology*, 1(1), 19–43.
- Chapelle, C. (2000). Is network-based learning CALL? In M. Warschauer & R. Kern (Eds.), *Network-based Language Teaching: Concepts and Practice*. Cambridge: Cambridge University Press.
- Chapelle, C. (2001a). *Computer Applications in Second Language Acquisition*. Cambridge: Cambridge University Press.

- Chapelle, C. (2001b). Innovative language learning: Achieving the vision. *ReCALL*, 13(1), 3–14.
- Chun, D. M. (1994). Using computer networking to facilitate the acquisition of interactive competence. *System*, 22(1), 17–31.
- Debski, R. & M. Levy (Eds.). (1999). *WORLD CALL: Global Perspectives on Computer-Assisted Language Learning*. Lisse: Swets and Zeitlinger.
- Egbert, J. & E. Hanson-Smith (Eds.). (1999). *CALL Environments: Research, Practice and Critical Issues*. Alexandria, VA: TESOL.
- Esch, E. & C. Zähler (2000). The contribution of Information Communication Technology (ICT) to language learning environments or the mystery of the secret agent. *ReCALL*, 12(1), 5–18.
- Golebiowska, A. (1990). *Getting Students to Talk*. Hertfordshire: Prentice Hall.
- Granger, S. (Ed.). (1998a). *Learner English on Computer*. Harlow: Longman.
- Granger, S. (1998b). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer* (pp. 3–18). Harlow: Longman.
- Granger, S. & C. Tribble (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.) *Learner English on Computer* (pp. 199–209). Harlow: Longman.
- Granger, S. & S. Petch-Tyson (Eds.). *Extending the Scope of Corpus-based Research. New Applications, New Challenges*. Amsterdam: Rodopi.
- Gu, P. & Z. Xu (1999). Improving EFL learning environment through networking. In R. Debski & M. Levy (Eds.), *WORLD CALL: Global Perspectives on Computer-Assisted Language Learning* Lisse: Swets and Zeitlinger.
- Hoffman, R. (1996). Computer networks: Webs of communication for language teaching. In M. Pennington (Ed.), *The Power of CALL* (pp. 55–78). Houston, TX: Athelstan.
- Hughes, G. (1997). Developing a computing infrastructure for corpus-based teaching. In A. Wichmann et al. (Eds.), *Teaching and Language Corpora* (pp. 292–308). Harlow: Longman.
- Jackson, R. (2001). *Web Based Learning Resources Library*. See <http://www.knowledgeability.biz/weblearning/>, consulted: 27.11.03.
- Johns, T. (2000). Data-driven learning: The perpetual challenge. Paper presented at The Fourth International Conference on Teaching and Language Corpora, TALC 2000, Graz, Austria, July 19–23.
- Jones, R. (1997). Creating and using a corpus of spoken German. In A. Wichmann et al. (Eds.), *Teaching and Language Corpora* (pp. 146–156). Harlow: Longman.
- Kaszubski, P. (1998). Learner corpora: The cross-roads of linguistic norm. Poster presented at *Teaching and Language Corpora Conference*, Keble College, Oxford, 24–27 July.
- Kelm, O. R. (1992). The use of synchronous computer networks in second language instruction: A preliminary report. *Foreign Language Annals*, 25(5), 441–454.
- Kelm, O. R. (1996). The application of computer networking in foreign language education: Focusing on principles of second language acquisition. In M. Warschauer (Ed.), *Telecollaboration in Foreign Language Learning* (pp. 19–28). Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.

- Kennedy, G. (1991). Between and through: The company they keep and the function they serve. In K. Aijmer & B. Altenberg (Eds.) *English Corpus Linguistics. Studies in Honour of Jan Svartik* (pp. 95–110). London: Longman.
- Kern, R. & Warschauer, M. (2000). Theory and practice of network-based language teaching. In M. Warschauer & R. Kern *Network-based Language Teaching: Concepts and Practice* (pp. 1–19). Cambridge: Cambridge University Press.
- Kramsch, C. & R. Andersen (1999). Teaching text and context through multimedia. *Learning & Technology*, 2(2), 31–42.
- Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann et al. (Eds.) *Teaching and Language Corpora* (pp. 1–24). Harlow: Longman.
- Levy, M. (2000). Scope, goals and methods in CALL research: Questions of coherence and autonomy. *ReCALL*, 12(2), 170–195.
- Lock, G. & A. Tsui (2000). Customising linguistics: Developing an electronic Grammar Database for teachers. *Language Awareness*, 9(1), 17–33.
- Locke, W. (1965). The future of language laboratories. *Modern Language Journal*, 49, 294–304.
- Lynch, T. (2001). Seeing what they meant: Transcribing as a route to noticing. *ELT Journal*, 55(2), 124–132.
- McGinity, M. (1992). Come together: The search for practical answers to specific problems in ESP. *Teachers Develop Teachers Research Conference* held at Aston University from 3–5 September 1992.
- Muñoz, C. et al. (Eds.). (2001). *Trabajos en lingüística aplicada*. Barcelona: Univerbook SL.
- Pennington, M. (Ed.). (1996). *The Power of CALL*. Houston, TX: Athelstan.
- Pérez-Paredes, P. (2003). Integrating networked learner oral corpora into foreign language instruction. In S. Granger & S. Petch-Tyson (Eds.), *Extending the Scope of Corpus-based Research. New Applications, New Challenges* (pp. 249–261). Amsterdam: Rodopi.
- Pérez-Paredes, P. & F. Mena (2001). La implicación del estudiante en el aprendizaje de la lengua oral. Las propuestas de W. Littlewood y su conceptualización en un nivel universitario. In J. R. De Mendoza Ibáñez (Ed.), *Panorama actual de la lingüística aplicada. Conocimiento, procesamiento y uso del lenguaje*, Vol. 1 (pp. 1783–1792). Logroño: Universidad de La Rioja.
- Pérez-Paredes, P. & P. Cántos-Gómez (2002). Some lessons students learn: self-discovery and corpora. Paper delivered at The Fifth International Conference on Teaching and Language Corpora, TALC 2002, Bertinoro, Italy, July 27–31.
- Rosenberg, M. (2001). *E-learning*. New York: McGraw-Hill.
- Salaberry, M. R. (2001). The use of technology for second language learning and teaching: A retrospective. *Modern Language Journal*, 85(1), 39–56.
- Sánchez, A. (2000). Language teaching before and after “digitalized corpora”. Three main issues. *Cuadernos de Filología Inglesa*, 9(1), 5–37. Murcia: Corpus-based Research in English Language and Linguistics. Universidad de Murcia.
- Sánchez, A. et al. (1995). *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*. Madrid: SGEL.
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 17–46.
- Skehan, P. (1989). *Individual Differences in Second Language Learning*. London: Arnold.

- Skehan, P. (2001). The role of a focus on form during task-based instruction. In C. Muñoz et al. (Eds.) *Trabajos en lingüística aplicada* (pp. 11–24). Barcelona: Univerbook SL.
- Spector, J. & P. Davidsen (2000). Designing technology enhanced learning environments. In B. Abbey (Ed.), *Instructional and Cognitive Impacts of Web-Based Education* (pp. 241–261). Hershey: Idea Group Publishing.
- Terceiro, J. & G. Matías (2001). *Digitalismo: El nuevo horizonte sociocultural*. Madrid: Grupo Santillana.
- Tognini Bonelli, E. (2001). *Corpus Linguistics at Work* [Studies in Corpus Linguistics 63]. Amsterdam: John Benjamins.
- Warschauer, M. & Kern, R. (2000). *Network-based Language Teaching: Concepts and Practice*. Cambridge: Cambridge University Press.
- Waters, A. & M. Vilches (2001). Implementing ELT innovations: A needs analysis framework. *ELT Journal*, 55(2), 133–141.
- Wichmann, A. et al. (Eds.). (1997). *Teaching and Language Corpora*. Harlow: Longman.

Prospects

New evidence, new priorities, new attitudes

John Sinclair

As well as supporting many of the daily routines of language teaching, the use of corpora raises some more radical questions, and this paper tries to anticipate some of them. It picks four familiar features of the way language is understood and presented in language teaching, features which can be problematic in the classroom, and asks some questions about them, calling on a corpus for evidence. From the discussion emerge several points for guidance, and some implications for methodology.

Introduction

This paper picks out four aspects of the way we perceive and handle language that can be a nuisance in teaching and learning. They are briefly characterised, and then examined one by one to see if they are inalienable features of language and not merely constructs of the way we perceive and describe language. The next section derives from this analysis a number of hypotheses about language which, if they could be incorporated into a theory, would improve the relationship between language description and meaning, and the final section draws out some implications for materials and methods that arise from the arguments.

The impetus for this review comes from observations made using large electronic corpora. In addition to the critique that corpus evidence is providing of the details of structure, pattern and meaning, this paper opens a discussion at another level, by reviewing some largely unchallenged characteristics of language and its description. In general from a classroom perspective the emergence of corpora may not seem to be good news – a large amount of new information to absorb, and an unsettling failure to confirm the consensus view of language that has been considered adequate for most classrooms for many years. The present paper suggests that much of the apparent difficulty arises not from corpora but from a poor fit between the models we use and the data that corpora uniquely provide; many of the problems just dissolve when the

theoretical adjustments are made. In the long run it is unlikely that corpus evidence can just be ignored, and this paper complements the various specific reports on language patterning by examining how the incorporation of corpus evidence into language pedagogy may proceed, and what benefits may accrue as a result of the incorporation.

1. Four features of language and language description

I suppose that every language teacher has his or her personal list of things about language that make it difficult to teach and learn – I certainly have mine. The complexity, the abstraction, the adaptability and the remarkable resources of a natural language make the task of teaching it quite formidable. One motivation for this paper is to check that it really is as difficult as it looks – and, just perhaps, to conclude that it looks more difficult than it is. Checking of this kind is immensely aided by the availability of evidence from a corpus, and I will use the large resource of The Bank of English¹ to support my argument.

Let me first pick out some items from my blacklist – features of language that appear to be unavoidable and which complicate the teaching and learning of languages.

Ambiguity. Words have far too many different meanings, and there is no simple correlation between form and meaning. Learning would be so much simpler if there was just a constant relationship of “one word – one meaning”.

Variation. This is almost the opposite of ambiguity; there are far too many ways of saying much the same thing, and it is often pedantic to suggest that they differ significantly in meaning. Learners need to keep their feet on the ground and feel that there are clear and simple answers to their problems of expression and understanding. If they are given ifs and buts and perhapses, they will turn off and rapidly conclude either that the language or the teacher is lacking in clarity.

Terminology. Although most of the familiar grammatical categories – like *first person* – have names that suggest meanings, they are not related to everyday life in any simple way. After a while they get learned and used without thinking, and as labels for obvious items or processes in a language they are as good as any others. The problem is that they let you down just when you need them to mean something. There are two kinds – those that seem to mean something, like *past*

or *negative*, except that they don't always have their obvious meaning; and those that are part of a private language, like *perfective* or *subjunctive*. Just what does *finite* mean for current English? How does it differ from *non-finite*? Can you tell just by looking at the form of the verb? Is the imperative a finite verb?

And another point about terminology. While there are plenty of terms to describe grammatical objects and events, lexical terms – like *idiom* – are in short supply and are defined only vaguely. The recent interest in lexis in language teaching has exposed an embarrassingly broad range of categories which, while incontrovertibly linguistic entities, have no names. For example, what kind of things can be *rife*, and where are they likely to occur (see Partington 1998:67)? Do you *get in contact with* the same people or things as you *get in touch with*? Current methods of describing languages offer no apparatus for classifying the answers to these questions. See further Sinclair (2003).

Incompleteness. The point about the lack of a terminology for lexis is part of a larger picture. A substantial proportion of the meaning of a text is not set out systematically in existing descriptions, and so is difficult to teach, and difficult for the learner to become aware of and gain control over. There are no established terms, even in the grammar, for these aspects of organised meaning. For example if you read “It’s about as useful as a...”, you know that the rest of the noun phrase following *a* is something that is both useless and ridiculous. No grammar or dictionary will explain how this meaning arises, and there is no semantic set of “useless and ridiculous things”.

2. Are they inherent in the language or do they arise in the description?

This is quite a formidable list of awkward features, and it is worth considering at the outset whether all of them are inalienable features of language, or whether some of them may be problems of the way we observe and describe languages. If some are of the latter kind, then improved descriptions could alleviate the problems or even eliminate them.

Ambiguity II

I have argued for some time that ambiguity is not an inherent feature of language, but one that arises largely from the way we describe it. (Sinclair 1998:9–13). Essentially our viewpoint is too limited, and most of what appears to be ambiguity resolves itself when a wider range of text is examined. Of course, in such a complex organisation as a natural language there will always be the pos-

sibility of occasional accidental overlapping of forms, but this should only be a tiny and insignificant percentage, no barrier to learning. If a more accurate description eliminates most of the apparent ambiguities, the language should be easier to learn because the relationship between form and meaning will be more transparent.

There is a general property of language, called the *arbitrariness* of the linguistic sign, which has been one of the cornerstones of modern linguistics since Saussure pointed it out. It points up the fact that, leaving aside a very small area of vocabulary such as how we describe animal noises, there is no way in which the physical form of a linguistic sign can be predicted from its meaning or vice versa. Given that there are many thousands of signs, each in an arbitrary relationship with its meaning, it does not need much exercise of the imagination to appreciate that there is a strong chance that the same sign may – purely by accident – come to stand for more than one meaning.

In practice this rarely happens, though it is highly probable in theory. One explanation, for which there is evidence, is that there are diachronic principles at work in language works to avoid such accidents. One, developed by Orr (1939) suggests that one of the confusables simply drops out of usage; so when the words *quean* (prostitute) and *queen* (monarch) fell together in sound, one of them had to go. In the case of multi-word signs, distinctive phrasings are often developed for each meaning, so that, say, an informal word for clothing is *gear* and a container is a *box*, yet it seems unlikely that a speaker would use *gear box* for a clothes container, since its use as a component of a motor car is well established. Indeed, there are no instances of the former use in The Bank of English.

A more theoretically-loaded explanation for the rarity of ambiguity in practice (which is not in conflict with the others) is that signs with different meanings *must* affect their environments in characteristically different ways, and that therefore, returning to our first formulation, we are not talking about ambiguity as much as *differentiation*. By reconceptualising the problem in this way we may be led to a solution.

Variation II

On the other hand, variation is certainly an important feature of language itself, rather than of our way of observing language. For several reasons it is an essential feature – for flexibility in fitting phrases together, and for nuances of expression, among others. But from a teaching point of view variation is a nuisance. Instead of having a simple object to be taught – a word or phrase that can be directly and exclusively associated with a meaning, teachers are often

faced with a bewildering range of alternatives. Conrad (this volume), dealing with another aspect of variation, points out how teachers find this aspect of language one of the most frustrating.²

A good pedagogical description of a language will organise the variation and prioritise the variants for language teaching purposes. This process of classification moves from the initial listings of variants to an understandable pattern of relationships that can be arranged according to different criteria such as complexity and familiarity. Here the most superficial findings of corpus analysis can be used – the frequency of occurrence of items. It happens that in most cases of varying realisations of a phrase, one of the alternatives is far more frequent than any of the others, and an obvious candidate for a canonical form,³ easy to teach and with the authority of a corpus behind it. In this way the problems of variation can be deferred until the principal sign-meaning relationships have been acquired.

The distinction between reception and production can be used to advantage when teaching in an area of great variation; when dealing with production there need be no variation allowed for at all in the first instance, and as variants turn up in the learners' receptive experience of language they can be associated with the canonical form and minor adjustments of meaning can be pointed out.

As an example of a typical pattern of variation, consider the collocation *save* with *skin*. The Bank of English offers 201 examples of a form of the verb *SAVE*⁴ followed within three words by *skin*; in most cases this collocation indicates a lexical item in which the meaning of *skin* is no longer just a reference to the outermost body covering, but combines with *SAVE* to make a subtle meaning. It refers to someone's unmerited escape from the unfortunate consequences of their incompetence or reprehensible behaviour, with more than a hint of selfishness and lack of concern for others.

Where there is one word intervening between *SAVE* and *skin* it is almost always a possessive of one kind or another; the most common are the possessive adjectives *her*, *his* etc., with *her* and *your* especially prominent. Other single words are either nouns in the possessive form, ending with "s" or the definite article *the* introducing an *of*-phrase that has a similar meaning to the possessive.

Where there is more than one word intervening, the likeliest combination is a possessive adjective followed by the word *own*. *His* is still very common as the possessive, but the others are less so. Just over half the instances show the addition of *own*, which stresses the self-centredness of the meaning.

Skin can occasionally be replaced by another word, e.g. *hide*, or, strangely, *bacon* (see Cobuild 2002 for some background on this word). In American English *ass* or *butt* can replace *skin*; it is regarded as a vulgar usage, pointing up the disapproval that a speaker or writer expresses by using this idiom.

Terminology II

The problems associated with terminology are also largely of our own making. When we come to describe the association of form and meaning, our first step is to separate these two into an area devoted to the description of structures, without reference to the lexical choices that realise the elements of structure, and an area devoted to classifying the meanings of the words, without reference to the structures in which they are organised. The former may be called *grammar*, and the other *lexis*. If there was genuine independence between these two aspects of language patterning, then the split would be justified, and the terminology of grammar could be as arbitrary as the terminology of any exact science. But it is now clear that semantic considerations are intricately associated with grammatical choices. Unfortunately, the two halves of language form cannot be reunited once split, because they subdivide according to quite different criteria, and so the terminology of grammar is left in an uncomfortable position, unable to reflect the usage and yet not precise enough to avoid some reliance on the implied meaning of the terms. (Sinclair 2001)

Some grammars have begun to recognise this problem in recent years, and the “pattern grammars” (e.g. Francis et al. 1996) are an excellent move towards realigning structure and meaning; a comprehensive set of structural patterns around the major word classes was compiled and The Bank of English was searched to discover which realisations of the major classes were found in each pattern. So for example a verb pattern “V *so*” is associated with a class of just nine verbs – ASSUME, BELIEVE, FEAR, HOPE, IMAGINE, PRESUME, SAY, SUSPECT, THINK (Francis et al. 1996: 120).⁵ The word *so* refers to an earlier report, and there is no obvious reason why only these verbs regularly take the pattern. The following features emerge on further investigation:

- (a) Several of the verbs have more than one meaning attributed to them, and this pattern is not available in all the meanings – thus the Cobuild Dictionary (2nd edn. 1995) recognises five senses of FEAR as a verb, of which only one allows this pattern.
- (b) There is a semantic harmony about the members of this class, to do with attitudes to the likelihood of future events, and this kind of organisation characterises the classes that are uncovered in the pattern grammars.

- (c) The verbs SAY and THINK are fairly ubiquitous reporting verbs, and thus have less strong attachments to the shared meaning of the class as a whole, so the association between meaning and pattern is not absolute or ring-fenced; no doubt other reporting verbs occasionally use it. But a learner can take from this evidence a clear, safe and useful phrasing.

More generally speaking, “lexicogrammar” has become fashionable, but lexicogrammar does not question either (a) the advisability of the original separation, or (b) the priority of grammar as the basic descriptive frame, with lexical material woven in from time to time. Lexicogrammar is still firmly a kind of grammar, laced, or perhaps spiked with some lexis.

The reason that “singular” does not reliably mean “one of” is because each of the thousands of nouns which select from the number system has a unique relationship with countability, discreteness and other factors that affect the meaning of the choice of “singular”; for example the word form *eye* is not the singular for which *eyes* is the plural. It happens in life that eyes tend to come in pairs, and so the singular, referring to just one eye, is only required in special cases like anatomy, injury or surgery. The form *eye* is thus free to form part of a rich variety of idiomatic phrases, such as *catch his eye*, *keep an eye on*, *in my mind’s eye*, *the naked eye*, *turn a blind eye to*. Similarly, “past” does not reliably mean “earlier” because each verb meaning not only has a unique relationship with the time dimension but also relates to other verbs in the same sentence in ways which can affect the timing of what meaning the verb realises.

The early separation of lexis and grammar fragments and obscures another important level of organisation of language – the functional. Even the grammars that call themselves functional grammars cannot make the leap from text pattern to action with any reliable predictability, because to specify a verbal action requires control over both the lexis and the grammar, simultaneously. Recently a new kind of grammar called a *local grammar* has emerged (Barnbrook & Sinclair 2001) which is specially constructed for a small descriptive job, and which can call on any aspect of the written or spoken language for inclusion among its categories.

But let us return to the familiar current situation. Deprived of lexical content, structures cannot quite align with meaningful categories; deprived of structure, the vocabulary of a language also cannot organise its meanings, which is one reason why the terminology of lexis is so inadequate – indeed impoverished. The removal of the essential connection with structure reduces lexis essentially to a list, which is the way in which it is generally handled in grammars. Complementary to the grammar there is a lexicon, which is also a

list of words; with each word goes a set of criteria that restrict its use, and on any occasion when these criteria are met the word can be slotted into the text and will fit – if the grammar is good enough. Lexis is reduced to vocabulary.

This point goes well beyond mere terminology, and exposes another area where problems are caused for teaching and learning by the inadequacies of our models of language rather than inherent features of the language like variation. Because of the isolation of the lexis, semantic theories have been obliged to search for and import a suitably profiled external system for organising meaning to map onto the vocabulary. As with the terminology of grammar, the match is not very good, and there are large and systematic gaps; furthermore, the main concerns of ontologies and the like seems to be far removed from the needs of describing everyday language. Logical mappings suffer the same fate – their categories are not very closely related to those of natural language and their concerns do not seem very important from the point of view of linguistic communication.

There is a touch of irony in this situation, since both ontologies and logical systems derive pretty straightforwardly from natural language and are simplifications and regularisations of the way words and phrases are normally used. The subtlety and flexibility of meaning that is so characteristic of its everyday use is regularised and sanitised to make the words stable nodes in a network that is far removed from their textual origins. In these circumstances it would indeed be strange if they could substitute for the contextually sensitive relationships that are contracted in actual text.

As an example, consider some of the principal verbs used to express perception. SEE is not primarily used for observation with the eyes, but for understanding, the phrase *I see* being one of the commonest responses in the spoken language, and *you see* being one of the commonest discourse particles (Aijmer 2002). TOUCH is more often used to express communication in general rather than physical contact – strangely, *contact* is quite often the verb of choice for that meaning. HEAR is principally used of coming to know, not necessarily via the spoken medium. SMELL is indeed common for olfactory perception, but is by no means a neutral technical term; there are prominent figurative and ironic uses of it (like *the sweet smell of success*), and plenty of bad smells around. Much the same can be said of TASTE, where *good taste* and *bad taste* refer well beyond the eating sensation. So although these verbs represent the five senses, and would have that place in an ontology or thesaurus, such a classification is clearly marginal to the way in which they are used to make meaning in speech and writing. It would in fact be confusing to a student to be introduced to

these verbs as expressing the five senses when most of their lively patterning is elsewhere.

Many people have a notion of “core” vocabulary (Carter 1987), and would formulate a clever argument, based on hindsight, that related the common meanings of these verbs to the presumed core, which is their role in the five senses. The relation between seeing and understanding, hearing and getting to know, smelling and smelling bad smells, touching and communicating with – these are understandable semantic movements over time. And so they are, but since they cannot be predicted they explain nothing; if the historical development had actually moved in the opposite direction, and SEE, originally a verb of understanding, took on the special role of ocular vision, our present situation would be just the same as it is.

We need to enter a caveat here. Whenever we find, as here, a strong tendency among language users to explain meaning in a certain way, regardless of the hard evidence from a corpus, we must respect it and seek a way of reconciling the positions. Intuitive reactions are important even though they are not necessarily supported by evidence. While we are perfectly at liberty to over-rule our intuition, we should be prepared to reinstate it at a later date, with a deeper understanding of the way language works (Sinclair 2004a).

Provisionally, however, we have to note that the imported schemata for the organisation of the lexis of the language are all unsatisfactory, and cannot replace the organisation that was lost when the structure was removed, and cannot bring the two together again.⁶ When we study a text directly, we interpret it holistically; analysis by its nature disintegrates, so there must be a means of reconciling the atomistic results of analysis with the experience of the communicative event that is analysed. Firth (in Palmer, ed. 1968:177) called this *renewal of connection* with the data, and it is this area where many attempts at automatic analysis have bitten the dust, in part because, having separated grammar and lexis at the beginning, they are unable to put them together again.

Incompleteness II

What was referred to as “incompleteness” above is an inherent feature of descriptions rather than languages, but in order to understand it we have to consider the underlying concepts and find a better name for this feature. First, we can take it for granted that no description of meaning will ever be quite complete – it is too complex and too fuzzy a phenomenon for that. Variation ensures that there is no limit to the number and type of meaningful patterns, which in turn means that any codification of the patterns with respect to mean-

ing will have to leave a proportion of the patterns to be interpreted by the users.⁷

Accepting that there are no complete descriptions, we must, secondly, ask ourselves why the particular kind of incompleteness that we find in conventional language descriptions is unfortunate for language teaching and learning, and search for a preferable incompleteness.

The most important single step is to revise the model of description in order to make it fit the data better. A corpus is a powerful investigative tool for use in this revision, and – in a limited way – it has been put to use for some years now in improving parsers, taggers and other tools of analysis. It has opened up the prospect of elaborating lexical structures, which when integrated into the model as a whole should reduce the overall complexity. When the model is made as efficient as possible, then the residual, unavoidable gaps will be seen as exploitations of the regularities on which the description is based.

3. New hypotheses

Let us now look at each of the problem areas, and consider how we could adapt our conventional theories of language in order to improve the situation. We will try to formulate a hypothesis for each, or perhaps more than one, and then rebuild the theory around them.

Ambiguity III

We have seen that the reason for so much wrongly-perceived ambiguity is that the pieces of language under scrutiny are too small. For a number of reasons, quite sensible in earlier times, our approach to vocabulary and meaning is to associate meaning with the shortest possible segment of language, and the one that shows the least variation. Most commonly it is the word.

This can be called a minimal approach; the alternative, a maximal approach, would be to extend the dimensions of a unit of meaning until all the relevant patterning was included – all the patterning that was instigated by the presence of the central word. At least, to stay on the practical side, we should extend the unit until the ambiguity disappears (as it does in almost every case). An appropriate hypothesis would be:

§1 The lexical item is best described maximally, not minimally.

A lexical item consists of one or more words that together make up a unit of meaning. Inevitably this alteration in the perception of vocabulary will highlight variation, because the extended dimension of the unit of meaning will take it well beyond the present way of seeing a lexical item as not much more or less than a word. But variation is also part of the agenda, and one of the major jobs of any new theory is to account for and allow for the large amount of variation that occurs.

Variation III

To deal with variation we need a very strong hypothesis, one that has been shaping itself over many years:

§2 Each distinct meaning in a language can be associated with a word pattern that is unique to it.

It will be some years still before we know if this is true in all cases, but it is an important guiding principle. Using a combination of the maximal lexical item and the frequency of the individual words in it, a canonical form of the lexical item can be proposed, which will both be one of the commonest forms of the lexical item but also it will be different from the canonical forms of all other items.

This will probably not be the only form of the unit, but it is the ideal one for introducing to the student. Each meaning will be associated with a phrase that reliably creates that meaning and no other; questions like “What preposition goes with this verb?” will disappear because the most appropriate preposition will be part of the model expression.

Most lexical items will include a substantial range of variation in their make-up, which keeps the management of variation in the hands of the teacher. In some circumstances the variation can be explored as a teaching point in itself, showing the nuances of meaning that can be created, the limits of the alternatives, and the possibilities of exploitation of the structure to create ironies and figures of speech.

We can formulate another hypothesis to clarify this area:

§3 Lexical meaning is created at two levels – the general meaning of the lexical item, and the modulation of this meaning by selections of individual words within the item.

Much of the variation of language is contained within the lexical items, and this is an important point to note as we adjust to the new evidence that comes from a corpus. One of the common reactions of newcomers to corpus study is alarm at the huge amount of detailed and sensitive patterning that the corpus reveals. There is no doubt that fluency in the language entails mastery of all this patterning, and it is no consolation that languages have been taught for centuries without the benefit of this explicitness. The whole job of teaching and learning suddenly assumes new, enlarged dimensions, bewilderingly complex.

This first impression arises largely because the descriptions we are familiar with do not organise this data for us because they are not constructed to handle such patterning; furthermore, the theories that inform the descriptions have not envisaged this situation and need some adjustment. When the theories catch up, and their dependent descriptions are appropriately revised, then the data will appear organised once again, in a more appropriate framework.

One way of handling great complexity is to recognise different types of complex pattern and put those of the same type together, and insulate each type from the others, as far as possible. This of course was the origin of the division into grammar and lexis, which is now being questioned. Instead, it is proposed to contain much of the low-level complexity within the lexical item. The two levels of meaning in lexis gives promise of great clarity.

The number of lexical items is thus likely to be a lot fewer than the initial impression of complexity might suggest; also the relations between them will not be very complicated. This is an encouraging conclusion, because in recognising variation within the basic unit of meaning, we run the risk of a combinatorial explosion, leading to an unmanageable number of lexical items.

In the early stages of a study that recognises frequency of occurrence as an important element of the description, it is usually necessary to impose arbitrary cut-off points to limit the amount and diversity of the data. It is however only a provisional tactic, and part of the job of the linguist is to replace the arbitrary decisions with linguistically motivated ones. For example it is common practice to ignore single occurrences of a pattern when organising the more common patterns, and that usually puts aside something like half of the evidence available, which is very convenient. It can be justified on linguistic grounds with the argument that unless a pattern is recorded from at least two apparently independent sources, it cannot be evaluated; it could be a mistake, a

personal quirk, an oddity of transmission, or one of many other factors, which are unimportant when the job is to find the large-scale repeated patterns.⁸

There is a happy ending to this tactic, giving the lie to the saying that you cannot have your cake and eat it. Once the main lines of analysis are laid down, many of the single occurrences will be found to fit well into one or other of the categories that have been set up to accommodate the repeated events. A single occurrence may be classified as a minor variant of an established pattern, or a slight extension of a semantic grouping, thus increasing the comprehensiveness of the description. At present, however, the computer cannot be relied on to provide this finesse, and the allocations have to be carefully monitored until more of the semantic structure of the language is incorporated in the description.

The handling of single occurrences is a point that is taken up again in Incompleteness III; returning to the two levels of lexical meaning, it is clear that the internal level of selection within the item is a paradigmatic choice at a place in structure. There are several different kinds of paradigmatic choices, ranging from the familiar grammatical ones, called *colligations*, to the choices called *semantic preferences*, where the list of options can be open-ended. Unlike the usual descriptions of paradigmatic choices, these occur within the structure of a lexical item and are thus lexical choices; the paradigms are prioritised on frequency grounds, so even the same colligational set may have different priorities within different lexical items.

Many of the features of this view of subsentence structure will be familiar to anyone versed in English grammar. A syntagmatic structure is established with a number of elements, some optional, and at each place in structure there are paradigmatic choices. But there are three notable differences between this model and the conventional one.

- (a) the syntagmatic framework – the lexical item – is lexical, and therefore does not respect the boundaries of grammatical units
- (b) because it is lexical, the item is sensitive to the overall meaning that is being created, so that if a choice that could be made at one place in structure would alter the meaning of the whole, it must indicate the presence of another lexical item which shares some of the same elements; this quality is not available to grammars
- (c) the paradigmatic choices are not always realised at a particular place in the apparent grammatical structure, because they are interpreted semantically; so a “negative” is not only realised by a grammatical negative structure, but

by any expression that can be interpreted as containing the implications of negation – occasionally it is even inferred (Sinclair 1998).

So in (a) we see that the lexical and the grammatical item are not necessarily coterminous, at any rank of the grammar; that is to say, the syntagmatic patterns of grammar and lexis bear no systematic relationship to each other. In (c) we see that the same point can be made for the sets of paradigmatic choices – those that are made at a place in the structure of a lexical item can be more restricted than the relevant grammatical class and at the same time much wider.

Terminology, etc. III

The problem of terminology breaks down into an easy part and a difficult one. The easy part is to overhaul our definitions of items and processes which are not directly related to meaning. The hard part is to devise a way of talking about the ever-provisional nature of meaning itself.

First, however, we should consider the rather fundamental problem exposed by the discussion of terminology. Here is a corrective hypothesis.

§4 Meaning does not divide into lexical meaning and grammatical meaning, but remains a holistic creation of a stretch of language.

A conclusion from this is that the difficulties incurred by learners in talking about or even perceiving the way meaning is created can be laid at the door of the descriptions and not of the language itself. This is not to say that there are no problems to come, no complexity, no indeterminateness. There will be plenty, but at least if we get the descriptive parameters correct the problems will be linguistic ones.

(a) As an example of the easy part, consider the verb forms in English. Most verbs have four distinct forms, one without inflection, one with an added “s”, one with an added “d” and one with an added “ing”. In each case there may be some minor adjustments to accommodate the ending, so that the “s” ending may be realised by “-es” as well. The forms have various syntactic functions, both alone and in combination. The uninflected or *base* form realises the present tense except for the third person singular, which is the only role of the *s* form; it is also the imperative, and the form used in combination with modals; it is the form that combines with *to* to make the infinitive. After certain other verbs, like *HELP*, the base form is also used.

Similar statements can be made about the *d* and *ing* forms. Then there are a number of common verbs which distinguish two forms covering the range of the *d* form, like *ate* and *eaten* where the regular verbs have, for example, *consumed*. We can call these forms the *t* form and the *n* form respectively, though there are several different realisations of the endings. Any reference grammar will give a full listing of the forms. The difference between them is that the *t* form is finite and does not combine with other forms, while the *n* form is non-finite and is much used in complex verb forms such as the passive voice, the perfect and the pluperfect tenses.

For a small number of verbs there is no *d* form, and the base form is used, e.g. *put*, which has only three forms. Then there are oddities like *read*, where the base and the *d* forms are only distinguished in speech. And finally there is the verb *be*, which has more forms than any other verb.

It is easy to talk about the forms in this way, and the curious terminology of the combinations can be avoided – so that the “passive voice” can be identified as a form of BE followed by the *d* form which is defined as including the *n* form. The meaning does not come into the reckoning at this point in the description.⁹

From this discussion can be framed another hypothesis:

§5 Categories of conventional syntactic structure do not relate reliably to meaning.

From this we can derive guidance about the kind of terms that are best to use.

(b) We need a new way of talking about lexical choices, rather than a terminology. Terms entail definitions, and definitions entail a fixed relationship to a meaning; the signs are that meanings never get fixed enough to allow them to fit in with this orderly scheme, because meaning is only ever provisional, and it has enough of a personal element in it to defy ultimate categorisation.

§6 Terminology which is sensitive to meaning is inherently unstable because meaning is always provisional and negotiable.

As an example, consider the observation, now made about many lexical choices, that they largely pick out unpleasant things. Reviewing some of my own work, I find that the things that HAPPEN are usually not very nice, nor are those that SET IN, and if they are RIFE they are uniformly awful. Something you are on the BRINK of is likely to be very nasty indeed. And so on. But if we list the words and phrases that realise the nastiness, we can make two further

observations – (a) it is normal for a few items to be prominent in each list, and for there to be a long tail of unique realisations, and (b) while the lists overlap, they are not identical. That is to say, there is not a coherent class of “nasties” in English, but rather a large, untidy pool from which each syntagmatic frame prioritises certain collocates. For descriptive purposes those items that recur are the important ones, and the singletons either show clear semantic similarity to a recurrent one, or can be overlooked until their text is being interpreted as a communication.

A useful model for this state of affairs is the inclusion relationship of set theory – the accumulation of all the sets. The main criteria for membership of this “superset” are a perceived similarity of meaning and a defined similarity of the syntagmatic environment, the *cotext*.

§7 Lexical preferences can be represented in “supersets”, which are made up of sets each of which contains all the items that occur at a place in lexical structure.

Incompleteness III

The previous analysis of incompleteness concluded that it was an unavoidable part of the relation between a language and its description that there would be gaps in the description. This is the general condition of descriptions, neatly put by Borges in considering how accurate a map could be:

the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless

(J. Borges, “On Exactitude in Science”, in *The Maker*, 1960¹⁰)

There are, however, several reasons for the gaps, and we can identify three at this point:

- (a) some gaps may be accidental; there is no particular reason that they are not filled, and their discovery may lead to an improvement in the description
- (b) some gaps may be the result of deficiencies in the description – a description can obscure a pattern just as easily as it can expose one; the objective here is to try to use the most relevant description available
- (c) some gaps, however, are the result of sequences of unique text choices, so unlikely to recur that they cannot be included in a description on principle.

The starting point for all three reasons can be summarised in the following hypotheses:

§8 A description of the way in which language creates meaning is considered adequate if it accounts for all the patterns that recur.

This is a hypothesis worth elaborating further, because it can lead us to widen our notion of meaning and give much greater weight to the actual patterns of text. It has to be conceded that as formulated above it begs a lot of questions – what exactly is a pattern, what is a recurrence, what happens to the once-only patterns and how are they distinguished from text where the choices are unrestricted in their lexical patterning? This is not a suitable occasion for teasing out all the detail; we have clear enough ideas about a pattern and an occurrence to be going on with, and some experience in the continuum of which one end is the totally free occurrence of any item, and the other end is the entirely predictable occurrence of one particular item (the *open-choice* and the *idiom* principle – see Sinclair 1991: 109 seq). For this paper we will consider further the once-only patterns which are not open-choice – that is where the cotext sets up a strong likelihood of one of a small set of collocates occurring, and the word that is chosen is not one of them, but is an unusual choice. For example, a newspaper headline a few years ago announced that two countries were “on the brink of peace”. The phrase is easy to understand, but the usual choice of words in the position occupied by “peace” is quite the opposite – war, famine and other dreadful things. So strong is the expectation of the reader that a sense of irony is created, as if the countries were so habitually at war with each other that peace would be a dreadful outcome, that neither of them really wanted peace, etc.

The model of meaning that describes this state of affairs has two main components, a set of expectations built up from experience of using the language and an interpretative faculty for deriving meaning from unexpected co-occurrences. The set of expectations is roughly represented by our current works of reference, though they are in need of drastic revision. It is the interpretive faculty that is our focus of interest, because it is clearly of central importance. Language text is hardly ever entirely predicable, and while ironies and other unusual phrasings, gnomic utterances, puns and poetic niceties can be left for more leisurely study, the exercise and training of the interpretive faculty must start from day one and feature prominently throughout a language course.

§9 We understand text by relating the phrasings to our stored experience of recurrent meaningful patterns, and interpreting those that vary from what is stored.

4. Implications for teaching and learning

Following on the analysis and interpretation of the four “awkward” features presented to the language teacher and learner in current practice, this final section considers how the new energies and concepts produced by corpus investigation might improve the situation. One very positive point is that both teacher and student can make use of a corpus right away, with only a modest few hours of orientation; there is no need to wait for the new textbooks and reference books. Only fairly simple queries can be handled at this stage, but the results can be illuminating and very helpful. Several of the papers earlier in this book give examples of the information that can be gleaned at the first stage.

For this, you will need a computer of normal performance, a corpus and some query software. Will the corpus be 100% reliable, comprehensive and representative? Of course not, but do your present textbooks match these targets? Or your reference grammars and dictionaries? Or any native speaker models? Or any combination of these? Of course not. Any source of information about language has to be evaluated carefully, but at least you will know what is in your corpus and where it came from; what is more, if any pattern or usage occurs more than once from apparently independent sources then there is a very strong possibility that it is a regular pattern in the language.

Ambiguity IV

It was pointed out above that most ambiguity arises from not looking at a large enough “window” of text, and widening the window is a job that is easily done with a corpus. Hypothesis §1 suggests that units of meaning should be described exhaustively, so we need a *concordance* in order to pick out the typical *cotext* of the ambiguous forms. See Michael Barlow’s chapter in this book for details. You choose a word or phrase that you think is common to all the instances of the meaning that you are examining, and ask for a concordance of it. You can then organise the cotext so that the surrounding patterns show clearly – sometimes you will find it necessary to re-sort the cotext several times, and most software allows you to do this.

For example, it may be useful to differentiate among various uses of the word *fire*, particularly between shooting and burning. Using a concordance and reference to collocational profiles, it can be established fairly quickly that the

first usage – the more common in The Bank of English – is frequently combined with *cease*, *under* and *open*. The characteristic use of *cease fire* (which is often hyphenated) is a verb such as *arrange*, *negotiate*, followed by *a cease-fire*; of *under* it is *came/come under fire from* and of *open* it is *open fire on*. While these are by no means the only possibilities, they show the natural usage and can be safely followed.

These patterns are so strong that they emerge from the most superficial observation of the word form. There are several minor usages which are well-represented but which do not emerge with such numerical weight; however, if a student encounters one and wishes to check that it is regular, the corpus will confirm this.

If we remove these collocations, the next most significant combination is with *set*, and there are two phrases of almost equal frequency – *set on fire* and *set fire to*, and we are now dealing with burning rather than shooting. One of the things people set on fire is *the world*, but of course this is an idiomatic usage – they don't set fire to it; closer examination of its cotext shows that the phrase is used negatively, so that if something didn't exactly set the world on fire we know that it was a failure and a disappointment.

These usages with *set* are the typical ways in which we talk about fires started by someone; the next pattern is used where the arsonist is unknown or the fire starts spontaneously or accidentally – the phrase *catch/caught fire*.

At this point we have identified the two main uses of the word *fire* and shown that with sufficient cotext they are not ambiguous because they are *coselected* with other words round about that serve to make the meaning clear (§2).

If we now take out these instances, the next strong combinations are *the Fire Brigade* and *the Fire Department*, UK and US usages respectively. Beyond this point the less prominent and more subtle usages need special techniques to bring them out, and in the classroom at the present time this is about as far as it is wise to penetrate – apart from checking on particular usages that have been noticed. For example there is a phrase with *caught* and *fire* which is actually an instance of shooting rather than burning – *caught in the cross fire*. It is often used to describe a difficult position in an argument, as well as on the battlefield. The phrasing, and the collocation with *cross* is quite distinctive, so there is no risk of confusion despite its appearance among the instances of *caught...fire*.

As well as clarifying the meaningful units of the language, the corpus also allows the prioritisation of usages on the basis of frequency. No-one would argue that frequency is other than a rough indication of the importance of a sense or a phrasing, but it is an indication; for example, with the passage of

time usages that no longer occur in numbers can be summarily discarded from teaching materials, if only for the reason that there is plenty to learn among the more widely used patterns. Without corpus evidence, it is difficult for a dictionary to omit an obsolescent sense or usage.

Using the information in a corpus the teacher can present a lexical item confidently, knowing that the whole expression of which it forms a part is typical of the target language, and knowing that the form presented is quite distinctive. For production, this eliminates the risk of accidental ambiguity and offers reliable phraseology beyond the boundaries of the lexical item. For reception the distinctive form gives a sound basis from which to interpret variations.

Variation IV

We move naturally to the topic of variation, and we can return to some instances of the long lexical item above, about setting the world on fire. This item became obvious during an examination of the form *fire*, so we must first check that *fire* is an obligatory component of it. At present this is a little laborious because the usual software does not anticipate the need, so we have to enter something like “any form of the verb SET, followed by *the world*”. Sure enough, *alight* occurs about as often as *on fire*, and *aflame* only a few times; but these are possible variants.

Notice at this point that we are dealing with a single lexical choice whose realisation is six or seven words long, and within which there is some variation. It is an excellent exercise for learners to track and enumerate the variations, and see what effect they have on the meaning. Let us start with a fairly straightforward instance – the most typical instance of the variant with *fire*, and summarise the main patterns found.¹¹

This galloper hasn't exactly set the world on fire but has favourable conditions here.

Among the 55 instances the inflected forms of SET occur, *sets* and *setting*, so that the phrase can be fitted into the prevalent time and aspect of the utterance. Then we can look in more detail at the strong negative orientation of the phrase. The usual grammatical negatives occur, *not*, *n't* and *never*, and also *none of*, and *without* in front of *setting*. A lexical realisation of negativity occurs with the verb *failed* and the adjective *unlikely*. About one in ten instances appear to be positive, and it is worthwhile following those up to see where they come from and if they are genuine or perhaps ironic. Most software allows you to choose a wider context than just the one line, and for many purposes that is sufficient. In our example it does seem that most of the positive in-

stances are genuine; they all come from media publications given frequently to melodrama, making the bulk of the positive instances look like cynical realism.

This is an important result, because some lexical items have such a strong negative meaning that even if there is no word carrying it we still assume it. It means that a positive example from the more strident newspapers and magazines is not wrong nor ironic, but just the hyperbole that we half-expect anyway.

One counter-example is where the sentence starts with *Apparently* . . . , allowing the writer to disassociate himself or herself from the otherwise positive statement:

Apparently she is going to set the world on fire. That's good. We'll see what happens.

Finally we can look at the “exactly” word, which occurs in about a quarter of the instances. There are quite a number of adverbs that occur here, seminegatives like *scarcely* and *hardly* (which do not have another negative with them), and *quite* and *really* with a negative. These bring out the possibility of sarcasm in the phrase, and some students might be keen to follow up the ability of these words to create this semantic slant in other cotexts.

We can now examine the variant with *alight* to see if it shares the various shades of meaning and structural options with the one that we have been looking at. Yes, it does; the addition of *scarcely* and *hardly* is more common, *failed* recurs, and there are fewer clearly positive instances; one shows the ironic possibilities of this phrase rather well:

At Health, Frank Dobson sets the world alight just as dimly as Stephen Dorrell did before him

Frank Dobson was a rather dull Minister of Health in the UK government in the closing years of the last century. The contrast between *alight* and *dimly* shows the bathos, and we know from other research that the structure *as X as* is often used to make absurd comparisons, especially when preceded by *about* or *just*.

Here are several other leads which can be pursued by interested students.

The third possible main pattern was with *aflame*. There are only three instances, but if they show the same semantic features we can treat them as a minor variant. Here they are:

will shoot forth such a spark as will set the world aflame. When Holden probed, ‘for this craft to set the world aflame, then? clash against Zimbabwe was about to set the World Cup aflame.

The first two instances come from science or mythological fiction and are clearly ways of indicating archaism or remoteness – the second one, in a wider context, seems to be intended literally. The third is a typical journalistic development from *world* to *World Cup*; the previous clause is “If the pre-match hype was to be believed...”, thus sabotaging the force of the expression.

Anyone setting priorities for language learning can quickly tell that this variant is unimportant, and potentially confusing; best ignored. Luckily teachers can ignore inconvenient evidence, and if it is like this, minor and unclear, then the decision is easy. If a student comes across such a rare event the meaning will be obvious by association with *alight* and *on fire*, and if a student challenges the teacher with this evidence, the explanations are easy. It is helpful if students are encouraged to challenge, using corpus evidence, any statements about language because either the exercise will bring out a hitherto unknown pattern (of which there are thousands) or they will fail to evaluate their evidence accurately, and open an opportunity for guidance.

In this example I have outlined in the greatest brevity some simple ways in which variation can be controlled and harnessed, rather than being perceived as a nuisance. It shows the teacher firmly in control, able to evaluate different kinds of variation and able to deal with all the information that a large corpus will throw up. The student has the security of seeing familiar patterns repeated frequently, while indulging his or her curiosity by exploring uncharted textual territory, developing techniques of self-learning and assessment of evidence which will be valuable skills for life outside the classroom.

The two-level structuring of a lexical item – first a *coselection* of (usually) several components in a number of relationships of collocation etc., then individual choices of words for each component – is just the first step in what may become a descriptive apparatus as powerful as a grammar (§3). The two levels are distinguished as follows: the upper level creates the overall meaning, especially the *semantic prosody*, a sort of attitudinal or pragmatic meaning that gives the reason for the choice and the deployment of the item. Coselection takes place on what is usually called the *syntagmatic* dimension, the unfolding of text. The lower level is the *paradigmatic* level, where the fine tuning is done, fitting the item into its context, and creating ironies and special effects. Separating these two types of meaning-creation is clarifying and makes the vocabulary easier to master, and light-years away from lists of words and “their” meanings.

Terminology etc. IV

A reconsideration of the division into grammar and lexis will take time and then more time for textbooks and reference material to be produced to make a holistic approach teachable (§4). In the meantime teachers can point out the intimate relations between meaning and choice, using the direct experience of interrogating a corpus. It is easy to point out that in the expression *set the world on fire* the word *on* cannot be exchanged for any other. In the long term this kind of observation, multiplied a thousandfold, threatens the definition of a preposition because a conventional grammar declares that all the prepositions may occur at any place in structure where one of them can. This argument can be left till later – the individual observations of the constituency of lexical items can be quietly accumulated. It is most important, in coping with variation, that words that cannot be varied are identified, because they are part of the *core* of the lexical item, the means by which it is signalled in running text.

It is also easy to use neutral terminology for most of the categories of structural grammar, and thus to avoid the implications of meaning that can be so misleading (§5). The attitudes of educational authorities and examining bodies to terminology should be taken into consideration and the new terms can be introduced gradually and as alternatives to the old. Once the terms are clearly decoupled from meanings, they will be a lot easier to use and will not attract the awkward questions that they sometimes get nowadays.

Where there are grammatical realisations of a particular meaning as well as lexical ones, then there is no objection to a semantically-loaded term; thus, as we saw above, “negative” was by no means only a grammatical choice in the verb, but could be realised by a variety of adverbs, prepositions, adjectives and lexical verbs (like *failed*). That is to say, it is a semantic choice, and one of the ways in which it can be realised is through the verb.¹²

There is a need, as we have seen above, for a new approach to terminology in following the creation of meaning (§6). While we can talk in rather vague terms about *semantic preferences* having a homogeneity among them, it may be almost impossible to give it a name – and what’s more, we will soon find another one, very close in meaning but not deploying exactly the same realisations. And we will find more and more, and as we search larger and larger corpora the class membership will shift a little bit and will discourage circumscription. For example the class of verbs that are regularly followed by *so* where the *so* refers to a report is difficult to give a name to, and it is difficult to limit it because any of hundreds of reporting verbs might occasionally “borrow” the construction. We have no examples in the corpus of “I surmise so” but it does not sound completely unacceptable.

In the teaching/learning process this does not need to be a big problem. Using a corpus will for some years to come be a voyage of discovery at every level of education – the student, the teacher, the class, the institution, the educational authority, the curriculum planners, the publishers. As new patterns and relationships emerge, they can be referred to quite informally and provisionally. In the longer term the organisation of meaning may not become much more fixed, because of the opportunity that a speaker has in every utterance of developing the meaning of an item by manipulating the context. We may have to talk in rather general categories, which is not attractive to teachers who want a firm foundation for their statements.

Once again, the much-maligned feature of frequency of occurrence can be used to define levels of granularity in description. Most of the apparent problems in the classification of meaning are in the less common events, and for patterns which are often repeated the computer can find hard evidence, and that evidence, as the pattern grammars begin to show, is often semantically homogeneous.

Lexicographers distinguish between *intensive* and *extensive* definition. The former is the usual kind, assigning a word to a superordinate and adding a feature that distinguishes it from its co-hyponyms. Extensive definition, on the other hand, simply enumerates all the items that may be named by the word. So “A **liquid** is a substance which flows...” (intensive), but “A **substance** is a solid, powder, liquid or gas...” (extensive) (Cobuild 1987).

The enumeration of the most frequent members of a class is a simple and effective way of characterising the class, and it is an accurate definition at a specified level of granularity; once we get into fine detail, of course, the lists get too long to be manageable and something more like intensive definition is needed. But that is some way ahead, and in our present discovery mode the extensive definition is quite adequate.

There are many ways of working this kind of exploration into teaching material. Many English word forms can occur both as nouns and as verbs, sometimes with almost the same meanings, and sometimes with subtle differences, such as *profit* and *combat*. You may come across several lexical items all of which suggest doom and gloom, but perhaps of different kinds. Many words have a range of meanings often described as ranging from “literal” to “figurative”. Such factors suggest that useful groupings might be made on an informal basis, and “supersets” (§7) formed. This can be the basis for discussion of similarity and difference in meaning, comparisons of overlapping lists of realisations, and insights into the nature of the vocabulary.

The importance of becoming aware of these semantic groupings and characterising them cannot be overestimated, because these are the operational parameters that eventually determine the appropriacy of a phrasing; a keen understanding of them leads to sensitive interpretation when reading and listening, and guides effective communication in speech and writing. Without access to a corpus and some simple tools for exploring it, this appreciation of the semantic preferences and prosodies can only be acquired through inductive learning over a long period; current reference books give guidance only on the extremely striking and obvious semantic parameters.

Remember that this is the practical side of meaning, and not the abstract classification of thesauruses (including visual ones), lexicons, ontologies, semantic webs, wordnets etc. These deal with another aspect of meaning, one which has little to do with the deployment of words in texts, and consequently is of little use in applications to textual analysis.

Incompleteness IV

Finally we come to the question of the gaps that must exist in any description, and we must try to make a description whose gaps are in places that are not very important to the intended application – in this case language teaching and learning. As §8 points out, an adequate description for all purposes should account for all observed patterns except the one-offs. This target in turn depends on the size and design of the corpus, because if a corpus doubles in size then we can reasonably expect that some repetitions of the original single occurrences will be found, and lots of new one-offs as well.

The “granularity” model based ultimately on frequency of occurrence is a good basis for dealing with this issue. Something that occurs a thousand times is likely to be more use to a learner than something that just occurs a few times. This is an extension of the principle of §8, but a justifiable one.

One way of looking at a corpus is as a repository of “used language” (Brazil 1995). The main repetitive lines of its holdings probably reflect fairly closely what is readily available to an average user – his or her stored knowledge of the language. A learner, presumably, has a much smaller, more patchy and less reliable store, and by becoming familiar with the corpus may be able to assimilate quite a lot. §9 claims that there are two ways by which we understand text; one is by referring to our stored knowledge and the other by interpreting those portions of a text which are not explained by the stored knowledge. It follows that a learner needs to develop strategies of interpretation (or adapt them from first language proficiency), and use this experience to feed and expand the store.

Concordances are ideal material for developing interpretive strategies, as Bernardini (this volume) points out. Returning to our example of FIRE for the last time, let us pick out what might appear to be a strange collocation, requiring some interpretation. This is the phrase *friendly fire*. If it was from the “burning” set of meanings rather than the “shooting” ones, then it might mean the warm look of a bright fire on a cold night; if it was from the shooting set, then it could mean supportive fire from colleagues in the area. Unfortunately it does not, and we can see from a tiny but unbiased selection:

the casualties came from misdirected ‘friendly fire’ from a Russian helicopter gunship.
 giving details of what are known as friendly-fire incidents in which the Americans
 determine whether those men died from friendly fire, a phenomenon which he said
 The men were killed by US friendly fire. They died when an American
 that has been causing so many ‘friendly fire’ casualties.

With slightly wider cotexts, each of these instances makes it clear what the phrase means e.g. (the fourth line above)

The men were killed by US friendly fire. They died when an American aircraft fired on two British armored personnel carriers by mistake.

The collocational profile of this phrase confirms the interpretation:

by	killed	casualties	incident	so-called
deaths	were	victims	Gulf	US
during	from	American	marine	tragedy
caused	died	British	hit	been

The whole story is there; the presence of *so-called* as a prominent collocate suggests that the latent irony in the phrase is still vivid for many people, though it is used as a technical term in military reports and discussions.

An examination of this kind is very likely to add *friendly fire* to the stored knowledge of the learners, along with its tragic prosody and its provenance of warfare.

Conclusion

I should reiterate that even the biggest corpora available are still small, especially for retrieving information about multi-word lexical items, which more and more are becoming the focus of research. The software available to someone who is not a computer specialist is simple and not very flexible, and sometimes it maddeningly refuses to do the obvious. We can certainly look forward

to having much more powerful tools and almost limitless text to try them out on – the Web itself is now being examined by search engines looking for concordances and collocations, and that, while still rather a jungle compared with a reasonably tidy corpus, is another huge source of language that is available in the classroom or the study at home.

To summarise the results of this investigation, we argued against ambiguity being an important property of language, but merely the result of bad theories; most of it can be dispelled by widening the context, and that gives rise to a whole panoply of activities that can derive from a corpus, large or small. On the other hand, variation is a major and essential property of language, but it can be controlled easily using the information provided by the corpus, and need not confuse the learner nor suggest that language is more complicated than it is.

The problem of terminology will not be solved in a short article, but the case is made for introducing a less misleading set of terms, which are easily coined, and to move gradually from one to the other. New terms, which will be needed in large numbers once the new information begins to be codified, can be extensively defined, leaving them open-ended and flexible. Corpus information can be used to control the inevitable gaps in the description, by introducing the notion of granularity into description and starting from the most clearly outlined and most often used patterns.

All the work suggested here can be merged with traditional models of language and language-teaching; the emphasis is different, and gradually we can expect the interests of the students to shift into new areas, but there is nothing revolutionary in my proposals as offered here. As time goes on there will be alternative theoretical positions formulated, and new descriptions begun, but for the present any teacher or student can readily enter the world of the corpus and make the language useful in learning.

Notes

1. At the time of finalising this paper, The Bank of English consisted of around 450 million words of contemporary English, last updated in 2002, and covering most major types of text, drawn principally from British English but with a substantial representation of other native-speaker varieties, particularly transatlantic and antipodean English. The corpus contains large amounts of transcribed spoken material.

The Bank of English is jointly owned by The University of Birmingham and HarperCollins plc., and can be accessed from the website <http://www.cobuild.collins.co.uk>. I am grateful for continued access to this unique resource.

2. Susan Conrad's paper in this volume deals with variations in usage that are associated primarily with register, or language suited to the occasion of its utterance or dissemination, so I will leave that area to her, and concentrate on the variability that does not entail contrasting varieties.
3. I am using *canonical form* to mean the most explicit, full and unambiguous presentation of a lexical item that can be achieved.
4. SMALL CAPS indicates that the word form cited stands for the whole *lemma*, that is all the usual inflected forms of the word. So SAVE is short for *save, saves, saving* and *saved*.
5. To this I would add EXPECT, and there may be one or two other contenders.
6. The consequences of this faulty view of language structure go well beyond a few teaching problems; it is now standard opinion in Information Technology that language is quite inadequately structured, especially semantically, and requires the erection of a massive external edifice to "code" the meanings and usages. These are called "lexicons" or "wordnets", and have consumed huge resources over the last fifteen years. None so far seems to work very well.
7. In a fully formal grammar there are devices that allow a finite system to match any number of patterns – for example Chomsky's well-known recursive rules (1965:6). Although indefinitely large, such a list of sentences will never be complete.
8. On other occasions it could of course be the single occurrences which are the object of enquiry.
9. Chomsky's early phrase structure grammar used such a device, where the passive was formed by adding (be + en) to the verb phrase structure, and a later transformational rule transferred the "en" to the end of the following word.
10. I had forgotten where this reference came from, but such is the wonder of e-mail that in a few hours I was reminded of it, courtesy of Anthony Chenells and Bill Louw of Zimbabwe.
11. The instance is arrived at by choosing all the lines containing the next most numerous collocate and repeating that process until there is only one left. This routine is automated in the program C-lect, and is very useful whenever you want to isolate a very typical example of a pattern.
12. Halliday (2002) makes the same point with reference to modality.

References

- Aijmer, K. (2002). *English Discourse Particles* [Studies in Corpus Linguistics 10]. Amsterdam: John Benjamins.
- Barnbrook, G. & J. Sinclair (2001). Specialised corpus, local and functional grammars. In M. Ghadessy, A. Henry, & R. Roseberry (Eds.), *Small Corpus Studies and ELT* [Studies in Corpus Linguistics 5] (pp. 237–276). Amsterdam: John Benjamins.
- Brazil, D. (1995). *A Grammar of Speech*. Oxford: Oxford University Press.
- Carter, R. (1987). *Vocabulary*. London: Allen and Unwin.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: The MIT Press.

- Cobuild (1987). *Collins Cobuild English Language Dictionary* ed. J. Sinclair, P. Hanks et al. 2nd edition 1995. London: HarperCollins.
- Cobuild (2002). *A Dictionary of Idioms*. Glasgow: HarperCollins.
- Francis, G., S. Hunston, & E. Manning (1996). *Grammar Patterns 1: Verbs*. London: HarperCollins.
- Halliday MAK (2002). "Judge takes no cap in mid-sentence": On the complementarity of grammar and lexis. University of Birmingham, Department of English.
- Orr, J. (1939). On Homonymics. *Studies in French Language and Mediaeval Literature presented to Professor Mildred K. Pope*, 253–297. Manchester.
- Palmer, F. R. (Ed.). (1968). *Selected Papers of J. R. Firth 1952–1959*. London: Longman.
- Partington, A. (1998). *Patterns and Meanings. Using Corpora for English Language Research and Teaching* [Studies in Corpus Linguistics 2]. Amsterdam: John Benjamins.
- de Saussure, F. (1917). *Cours de Linguistique Générale*. Paris: Payot.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair, J. (1998). The lexical item. In E. Weigand (Ed.), *Contrastive Lexical Semantics* [Current Issues in Linguistic Theory 17] (pp. 1–24). Amsterdam: John Benjamins. Reprinted in Sinclair (2004b).
- Sinclair, J. (2001). Lexical grammar. In M. Gellerstam, K. Jóhannesson, B. Ralph, & L. Rogström (Eds.), *Nordiska Studier I Lexicografi 5*, Proceedings of the Fifth Conference on Nordic Lexicography, Göteborg 27th–29th May 1999 Göteborg, Skrifter utgivna av Nordiska föreningen för lexikografi 6. Meijerbergs Arkiv för Svensk Ordforskning 27 Göteborg, Meijerbergs Institut för Svensk Etymologisk Forskning (pp. 323–343). Göteborg: Göteborgs Universitet. Reprinted in Sinclair (2004b).
- Sinclair, J. (2003). *Reading Concordances*. London: Longman.
- Sinclair, J. (2004a). Intuition and annotation – the discussion continues. In B. Altenberg & K. Aijmer (Eds.), *Proceedings of the 23rd ICAME Conference*. Amsterdam: Rodopi.
- Sinclair, J. (2004b). *Trust the Text*, ed. R. Carter. London: Routledge.

Notes on contributors

Silvia Bernardini currently has a research contract with the Department of Intercultural Studies in Translation, Languages and Culture of the University of Bologna at Forlì, Italy, where she is involved in the construction of the CEXI corpus, a parallel bi-directional corpus of English and Italian. Her research interests include corpus-based translation studies and contrastive linguistics and the didactic applications of corpora in English language teaching and translation. She is co-editor of the journal *Languages in Contrast*.

Amy B. M. Tsui is Chair Professor in the Faculty of Education of The University of Hong Kong. She obtained her PhD in linguistics in 1986 at The University of Birmingham. She has published widely in the areas of discourse analysis, language policy, teacher education and ICT in teacher education. Her most recent publications include three books, *Understanding Expertise in Teaching* (2003), New York: Cambridge University Press; *Medium of Instruction Policies: Which Agenda? Whose Agenda?*, Mahwah, NJ: Erlbaum, co-edited with J. Tollefson (2004), and *Classroom Discourse and the Space of Learning*, Mahwah, NJ: Erlbaum, co-authored with F. Marton et al. (in press).

Susan Conrad is an associate professor in the Department of Applied Linguistics at Portland State University. Her work in corpus linguistics includes collaborations on *Corpus Linguistics: Investigating Language Structure and Use* (Cambridge University Press), the *Longman Grammar of Spoken and Written English*, and the edited collection *Variation in English: Multi-dimensional Studies* (Longman), as well as articles in journals such as *TESOL Quarterly*, *System*, and *Linguistics and Education*.

Anna Mauranen is professor of English at the University of Tampere, Finland. Her major publications are in contrastive rhetoric, discourse analysis and corpus linguistics, including *Cultural Differences in Academic Rhetoric*. Her current research and publications focus on corpus linguistics, speech corpora, English as lingua franca and translation studies. She is compiling a corpus on

spoken English as lingua franca (the ELFA corpus) and running a large research project “Translated Finnish and Translation Universals”.

Luísa Alice Santos Pereira is a High School teacher of Portuguese as mother language, with experience on teaching Portuguese as foreign language. She works in the Linguistic Research Center of Lisbon University (Centro de Linguística da Universidade de Lisboa – CLUL) as collaborator of research in the group of Corpus Linguistics. She collaborated with the redactorial team of the *Dictionary of Contemporary Portuguese Language* (2001) (Dicionário da Língua Portuguesa Contemporânea).

Nadja Nesselhauf is an Assistant at the Department of English at the University of Basel, Switzerland, where she teaches courses in corpus linguistics, second language acquisition, and phonology. She has just finished her PhD, which investigates the use of collocations by advanced learners of English on the basis of a learner corpus. Her main research interests are foreign language teaching (in particular the use of corpora for teaching), second language acquisition, and phraseology.

Gyula Tankó is an Assistant Lecturer at the Department of English Applied Linguistics of Eötvös Loránd University in Budapest. He has taught EFL courses, academic writing courses, discourse and corpus analysis courses for teaching purposes at undergraduate level, and methodology courses on the teaching and assessment of writing both at undergraduate level and as part of in-service teacher training courses. His research areas are discourse analysis, corpus linguistics, research methodology and testing. He is currently working on his PhD dissertation in which he investigates the Rhetorical Move Structure of argumentative essays.

Ute Römer studied English linguistics and literature, Chemistry and Pedagogy at the University of Cologne and now works as a researcher and lecturer in English linguistics at Hanover University. She is currently writing up her doctoral thesis on the functions, contexts and didactics of English progressive verb forms, taking a corpus-driven approach to the topic. Main research and teaching interests include corpus linguistics and discourse intonation. She has recently co-edited *Language: Context and Cognition. Papers in Honour of Wolf-Dietrich Bald's 60th Birthday* (2002) and has published articles on corpus linguistics and language teaching.

Michael Barlow completed his PhD in Linguistics at Stanford in 1988. Since that time he has compiled the *Corpus of Spoken Professional American English* and has created a variety of text analysis tools including *MonoConc*, *ParaConc* and *Collocate*. The main strands of his other research interests are: usage-based models of language and the use of corpora in language teaching.

Pernilla Danielsson (PhD from Gothenburg, 2001) is the Academic Director for the Centre for Corpus Research at Birmingham University. She moved to Birmingham in 2000, to take up the role of project manager of the EU concerted action TELRI II. Since the end of the project in 2002, she has been involved in setting up the new centre as well as lecturing on the Birmingham's master degrees in corpus linguistics. She is also involved in running short courses in corpus linguistics, both in Birmingham and at the Tuscan Word Centre. Her own research covers areas of identifying units of meaning in corpora. She is the co-editor of *Meaningful Texts* (together with Geoff Barnbrook and Michaela Mahlberg) and is working on publishing her monograph 'Retrieving Meaningful Units from Corpora'.

Pascual Pérez-Paredes has worked as an EFL teacher in Spain since 1989, first in Secondary Schools and later, for eight years, in Escuelas Oficiales de Idiomas (State-run Language Schools). Since 1996 he has been working at the English Department in the University of Murcia, Spain. He completed his doctorate in English Philology in 1999 and currently teaches English Language and Translation. He is also a Sworn Translator.

John McH. Sinclair was Professor of Modern English Language at the University of Birmingham for most of his career, and Editor-in-Chief of *Cobuild* for much of that time. His education and early work was at the University of Edinburgh, where he began his interest in corpus linguistics, stylistics, grammar and discourse analysis. He now lives in Italy, where he is President of The Tuscan Word Centre. He holds an Honorary Doctorate in Philosophy from the University of Gothenburg, and an Honorary Professorship in the University of Jiao Tong, Shanghai. He is an Honorary Life Member of the Linguistics Association of Great Britain and a member of the Academia Europæa.

Index

A

adverbial connectors – distribution
168
adverbial connectors – position 175
adverbial connectors – semantic
relationship 171
adverbial connectors 70, 157seq.
ambiguity 272seq.
arbitrariness of the linguistic sign
274
arrays 240
arreigar-se 116
authenticity 19, 91

B

Bank of English 272, 297
bi-directional corpora 20
BNC 23, 185

C

célebre 116–118
classroom 15, 99
Cobuild 126
cohesion 160
collocates 26, 212seq.
collocation 212seq., 289
communicative utility 94
comparable corpora 20
concession 71
concordance 209, 288
concordancer 243
contrast/concession 71
corpus access and analysis 205seq.
coselection 292
CRPC (Portuguese) 109

D

data-driven learning 16, 126
day after day 48
day by day 49
deduzir 113–115, 121
definite article 53
digital bridge 260
dimensions of variation 75
discovery learning 22

E

elaborated reference 76, 84

F

famoso 116–118
formulaic expressions 95
frequency counter 239
frequency of occurrence 40
friendly fire 296

G

graduar 116

H

hailed 24–25
hands-on 102
high 45–48

I

idiom principle 18
impersonal style 76, 85
imply 57seq.

incompleteness of description
272seq.
infer 57seq.
informational production 76, 83
involved 76, 82

L

language awareness 41
language pedagogy 16
learner autonomy 28, 258
learner corpora – lists 129, 150–152
learner corpora – potential and
limitations 131
learner corpora and pedagogic
material 137seq.
learner corpora availability 133
learner corpora 125, 127seq.
learner corpus studies 134seq.
learner oral corpora – taxonomy
255
learner oral corpora 129, 152,
249seq.
lexical frameworks 217
lexical item 18, 281–283
lexicogrammar 277
Lexicon of Portuguese 110
linking adverbials 70, 157

M

mark-up 206
MICASE corpus 100
modal auxiliaries – co-occurrence
189
modal auxiliaries – frequency 186
modal auxiliaries – meanings 187
modal auxiliaries 185seq.
modals in EFL teaching 190seq.
modals in textbooks vs. modals in
corpus 193seq.
multi-dimensional comparisons
74seq.

N

narrative 76, 83
native speaker 30
naturalistic learning 94
network-based language teaching
152seq.
none of 53
notável 116–118

P

parallel corpora 20
pattern grammar 276
Perl – changing access rights 228
Perl interpreter 228
Perl programming 225seq.
persuasion 76, 84
Português Falado 110
prefabricated units 90
prefabs 96
programs – concordance 209
programs – concordancer 243
programs – frequency counter 239
programs – tokeniser 231
programs – word splitter 236
programs – wordlist 207

R

recurrence 287
regular expressions 233
resources for teaching 113
rules and evidence 50

S

schema theory 17
semantic preference 293
semantic prosody 292
serendipity 23
Sir 29
situation-dependent reference 76,
84
sofisticar 116
spoken corpus 89
staunch 24–25
sub-corpus 27

subject-verb agreement 50
suggest(s/ed/ing) 141, 142
supersets 286
synonymous items 44

T

tall 44–46
Telecorpora 43
TeleNex 42
terminology 272seq.
the case 96
there's 50
thing 210seq.
though 70–73
tokeniser – enhanced 238

tokeniser 231
translation corpora 19

U

used language 91, 295

V

variation 67seq., 272seq.
verbs of perception 278

W

well-experienced 55
word splitter 236
wordlist 207

In the series STUDIES IN CORPUS LINGUISTICS (SCL) the following titles have been published thus far:

1. PEARSON, Jennifer: *Terms in Context*. 1998.
2. PARTINGTON, Alan: *Patterns and Meanings. Using corpora for English language research and teaching*. 1998.
3. BOTLEY, Simon and Anthony Mark McENERY (eds.): *Corpus-based and Computational Approaches to Discourse Anaphora*. 2000.
4. HUNSTON, Susan and Gill FRANCIS: *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. 2000.
5. GHADESSY, Mohsen, Alex HENRY and Robert L. ROSEBERRY (eds.): *Small Corpus Studies and ELT. Theory and practice*. 2001.
6. TOGNINI-BONELLI, Elena: *Corpus Linguistics at Work*. 2001.
7. ALTENBERG, Bengt and Sylviane GRANGER (eds.): *Lexis in Contrast. Corpus-based approaches*. 2002.
8. STENSTRÖM, Anna-Brita, Gisle ANDERSEN and Ingrid Kristine HASUND: *Trends in Teenage Talk. Corpus compilation, analysis and findings*. 2002.
9. REPPEN, Randi, Susan M. FITZMAURICE and Douglas BIBER (eds.): *TUsing Corpora to Explore Linguistic Variation*. 2002.
10. AIJMER, Karin: *English Discourse Particles. Evidence from a corpus*. 2002.
11. BARNBROOK, Geoff: *Defining Language. A local grammar of definition sentences*. 2002.
12. SINCLAIR, John McH. (ed.): *How to Use Corpora in Language Teaching*. 2004.
13. LINDQUIST, Hans and Christian MAIR (eds.): *Corpus Approaches to Grammaticalization in English*. n.y.p.
14. NESSELHAUF, Nadja: *Collocations in a Learner Corpus*. n.y.p.
15. CRESTI, Emmanuela and Massimo MONEGLIA (eds.): *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. n.y.p.