Corpus-Based Approaches to English Language Teaching

Edited by

Mari Carmen Campoy-Cubillo, Begoña Bellés-Fortuño, and Mª Lluïsa Gea-Valor

Research in Corpus and Discourse

Series Editors: Wolfgang Teubert and Michaela Mahlberg

continuum

Corpus-Based Approaches to English Language Teaching

Corpus and Discourse

Series editors: Wolfgang Teubert, University of Birmingham, and Michaela Mahlberg, University of Liverpool.

Editorial Board: Paul Baker (Lancaster), Frantisek Čermák (Prague), Susan Conrad (Portland), Geoffrey Leech (Lancaster), Dominique Maingueneau (Paris XII), Christian Mair (Freiburg), Alan Partington (Bologna), Elena Tognini-Bonelli (Siena and TWC), Ruth Wodak (Lancaster), Feng Zhiwei (Beijing).

Corpus linguistics provides the methodology to extract meaning from texts. Taking as its starting point the fact that language is not a mirror of reality but lets us share what we know, believe and think about reality, it focuses on language as a social phenomenon, and makes visible the attitudes and beliefs expressed by the members of a discourse community.

Consisting of both spoken and written language, discourse always has historical, social, functional, and regional dimensions. Discourse can be monolingual or multilingual, interconnected by translations. Discourse is where language and social studies meet.

The *Corpus and Discourse* series consists of two strands. The first, *Research in Corpus and Discourse*, features innovative contributions to various aspects of corpus linguistics and a wide range of applications, from language technology via the teaching of a second language to a history of mentalities. The second strand, *Studies in Corpus and Discourse*, is comprised of key texts bridging the gap between social studies and linguistics. Although equally academically rigorous, this strand will be aimed at a wider audience of academics and postgraduate students working in both disciplines.

Research in Corpus and Discourse

Conversation in Context
A Corpus-driven Approach
With a preface by Michael McCarthy
Christoph Rühlemann

Corpus-Based Approaches to English Language Teaching
Edited by Mari Carmen Campoy, Begona Bellés-Fortuno and
M^a Lluïsa Gea-Valor

Corpus Linguistics and World Englishes
An Analysis of Xhosa English
Vivian de Klerk

Evaluation and Stance in War News
A Linguistic Analysis of American, British and Italian television news reporting of the 2003 Iraqi war
Edited by Louann Haarman and Linda Lombardo

Evaluation in Media Discourse
Analysis of a Newspaper Corpus
Monika Bednarek

Historical Corpus Stylistics
Media, Technology and Change
Patrick Studer

Idioms and Collocations
Corpus-based Linguistic and Lexicographic Studies
Edited by Christiane Fellbaum

Meaningful Texts
The Extraction of Semantic Information from Monolingual and Multilingual Corpora
Edited by Geoff Barnbrook, Pernilla Danielsson and Michaela Mahlberg

Rethinking Idiomaticity
A Usage-based Approach
Stefanie Wulff

Working with Spanish Corpora
Edited by Giovanni Parodi

Studies in Corpus and Discourse

Corpus Linguistics and the Study of Literature
Stylistics In Jane Austen's Novels
Bettina Starcke

English Collocation Studies
The OSTI Report
John Sinclair, Susan Jones and Robert Daley
Edited by Ramesh Krishnamurthy
With an introduction by Wolfgang Teubert

Text, Discourse, and Corpora. Theory and Analysis
Michael Hoey, Michaela Mahlberg, Michael Stubbs and Wolfgang Teubert
With an introduction by John Sinclair

This page intentionally left blank

Corpus-Based Approaches to English Language Teaching

Edited by

Mari Carmen Campoy-Cubillo, Begoña Bellés-Fortuño and Maria Lluïsa Gea-Valor

continuum

Continuum International Publishing Group
The Tower Building
11 York Road
London SE1 7NX

80 Maiden Lane
Suite 704
New York, NY 10038

© Mari Carmen Campoy, Begona Bellés-Fortuno and Mª Lluïsa Gea-Valor 2010

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage or retrieval system, without prior permission in writing from the publishers.

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

ISBN: 978-1-8470-6537-7 (Paperback)

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress.

Typeset by Newgen Imaging Systems Pvt Ltd, Chennai, India
Printed and bound in Great Britain by MPG Books Group Ltd

Contents

Part One: Corpus Linguistics and ELT: State of the Art

Part Two: Corpora and English for Specific Purposes

Part Three: Learner Corpora and Corpus-Informed Teaching Materials

Part Four: Multimodality: Corpus Tools and Language Processing Technology

Notes on Contributors

Annelie Ädel's (annelie.adel@english.su.se) main research areas are discourse/text analysis, corpus linguistics and EAP. She has been affiliated with Boston University as a visiting scholar and with the University of Michigan's English Language Institute as a post-doctoral fellow and as Director of Applied Corpus Linguistics. She is currently a research fellow in the Department of English at Stockholm University, Sweden.

Mª Ángeles Andreu-Andrés, Ph.D. (maandreu@idm.upv.es) is an Associate Professor at the Universidad Politécnica of Valencia (Spain) with more than 20 years of experience in university education. Her major fields of interest and research involve the teaching methodology and evaluation of engineering students' learning. She is a member of the university teaching innovation groups IEMA and GIMA and a research member of DIAAL.

Begoña Bellés-Fortuño, Ph.D. (bbelles@ang.uji.es) is an English Lecturer in the Department of English Studies at Universitat Jaume I, in Castelló, where she currently teaches English Philology students as well as in the degree of Industrial Engineering within the frame of the EURUJI Project for European students' mobility. Her research interests are focused on Discourse Analysis, and more concretely, academic discourse both written and spoken, as well as on Contrastive and Corpus Linguistics, as her latest national and international publications show. She has been co-editor of the book *Cognitive and Discourse Approaches to Metaphor and Metonymy.* She has recently published the book *A spoken Academic Discourse Contrastive Study: Discourse Markers in North-American and Spanish lectures* (AESLA, 2008). She also has a co-authorship in the recently published book *Hablar ingles en la universidad: Docencia e Investigación* (Septem ediciones 2008).

Susana Murcia Bielsa (susana.murcia@uam.es) is a lecturer in the Department of English Philology at Universidad Autónoma de Madrid. She was awarded her Ph.D. by the Universidad de Córdoba (Spain) in 1999, although the research was carried out at the University of Stirling (UK). Her thesis and subsequent studies explore instructional texts in Spanish and English through a corpus-based approach. Since 2003 she has worked in several research projects using learner corpora to investigate interlanguage in secondary and university education. She is currently working on error analysis of a learner corpus to inform pedagogical applications. She has held teaching positions in various Spanish and British universities, including The Open University.

José Maria Alcaraz Calero (jmalcaraz@dif.um.es) is an Assistant Lecturer at the Department of Knowledge and Communication Engineering, University of Murcia. He has been involved in several national and international research projects on Computer-Aided Language Learning and Corpus-Based applications. His main research areas are Corpus Linguistics software, E-learning environments, Computee-Aided Language Learning, Knowledge Management and Ontologies. He is currently working towards his Ph.D.

Belinda Crawford Camiciottoli, Ph.D. (Universitat Jaume I, Castellón, Spain, bcrawford@tin.it) is an English language researcher at the University of Florence (Italy). She teaches English for Business Studies at the Faculty of Economics. Her current research focuses on interpersonal features of discourse in business settings. She has published in leading international journals and recently authored the monograph *The Language of Business Studies Lectures* (John Benjamins).

Mari Carmen Campoy-Cubillo (campoy@ang.uji.es) is senior lecturer and Head of the Department of English Studies of Jaume I University, Spain. She has recently co-edited *Spoken Corpora in Applied Linguistics* with M. J. Luzón (University of Zaragoza). She has also co-edited *Oral Skills: Resources and Proposals for the Classroom and Computer-Mediated Lexicography in the Foreign Language Learning Context.* Her main research interests are in the areas of lexicography and the application of corpus linguistics to the teaching of foreign languages. Her forthcoming work includes a contribution to the series *Papers on Lexicography and Dictionaries.*

Winnie Cheng (egwcheng@inet.polyu.edu.hk) is a Professor and Director, at the Research Centre for Professional Communication in English in the Department of English at The Hong Kong Polytechnic University. Her main research interests are corpus linguistics, conversational analysis, critical discourse analysis, pragmatics, discourse intonation, intercultural communication in business and professional contexts, and outcome-based education.

Sylvie De Cock (sylvie.decock@uclouvain.be) is a Lecturer in English language and linguistics at the Université catholique de Louvain and at the Facultés universitaires Saint-Louis (Belgium). She is a member of the Centre for English Corpus Linguistics (CECL, Université catholique de Louvain) and has been involved in the Louvain International Database of Spoken English Interlanguage (LINDSEI) project since 1996. Her main research interests include learner corpus research, phraseology, pedagogical lexicography and the study of English for Specific Purposes (Business English).

Izaskun Elorza (iea@usal.es) is Assistant Professor of English Language and Linguistics at the University of Salamanca (Spain). Her research focuses on written discourse, especially on newspaper discourse in English, from the perspectives of discourse analysis, systemic functional linguistics and corpus linguistics, as well as on their applications to reading and writing in ESL/EFL. Recent work has been published in *Language & Intercultural Communication.*

Begoña Montero Fleta, Ph.D. (bmontero @ idm.upv.es) is an associate professor of English for Computer Science at the Universidad Politécnica de Valencia (Spain). Her areas of interests are the application of new technologies to language teaching and the linguistic characteristics of scientific discourse and its assessment. As well as textbooks and material design for academic purposes, her publications and international conference presentations have addressed classroom innovations and new technologies applied to language teaching and comparative discourse analysis.

Inmaculada Fortanet-Gómez, Ph.D. (fortanet@ang.uji.es) is a Senior Lecturer and researcher at Universitat Jaume I (Castellón, Spain), where she coordinates the Group for Research on Academic and Professional English (GRAPE; www.grape.uji.es). Her research interests are related to academic and professional English. She has co-edited several books, such as *ESP in European Higher Education. Integrating Language and Content* (John Benjamins 2008). She has published chapters in well-known national and international books, as well as articles in journals such as *English for Specific Purposes, Discourse Studies* and *Journal of English for Academic Purposes.*

Blanca García-Riaza, M.A. (bgr@usal.es) in English Studies and holder of a research fellowship at the University of Salamanca (Spain). Her main research interests lie in the field of Systemic Functional Linguistics, Discourse Analysis, especially newspaper discourse, and Corpus Linguistics, on which she has published several papers and is developing her doctoral dissertation.

Maria Lluïsa Gea-Valor (gea@ang.uji.es) is a Lecturer in English Language, Linguistics and ESP at Universitat Jaume I (Castelló, Spain). Her research interests lie in the field of genre analysis, especially evaluative and promotional genres. She is the author of *A Pragmatic Approach to Politeness and Modality in the Book Review* (2000), and has published articles in indexed journals such as *Ibérica, RESLA, RLyLA, Pragmalingüística*, etc. She has recently co-edited the volumes *Internet in Language for Specific Purposes and Foreign Language Teaching* (2003), *Language @ Work: Language Learning, Discourse and Translation Studies in Internet* (2005) and *The Texture of Internet: Netlinguistics in Progress* (2007).

Maria Georgieva, Ph.D. (mageorg@nlcv.net) is currently Associate Professor in the Department of British and American Studies, 'St. Kliment Ohridski' University of Sofia. Her research interests and publications are in the field of Applied Linguistics, Sociolinguistics, Intercultural Pragmatics, Communication Strategies and Canadian Studies. Author of monographs, articles and EFL textbooks for primary and secondary school.

Rafael Alejo González (ralejo@unex.es) is a Lecturer at the Universidad de Extremadura, Spain. He teaches at the Faculty of Education and his research focuses on how to apply linguistic and discursive analyses to the teaching of English to foreign language learners. His recent publications touch on the use of metaphor in Economics and the problems posed to learners by phrasal verbs.

Lilyana Alexandrova Grozdanova, Dr. Litt. (lilian.gr@gmail.com) Professor of Linguistics. Department of English and American Studies, University of Sofia "St Kliment Ohridski". Her Research areas are Theoretical and Applied Linguistics, Teaching Methods, and Materials Design. She is the author of monographs, articles, course books, reviews, and annotations. She is also a Consultant to projects and Member of the Academic Fund for English and American Studies and the Bulgarian Society for British Studies.

Aurora Astor Guardiola (aastor@idm.upv.es) is an English teacher at the Polytechnic University of Valencia where she has taught English for Specific Purposes for many years, mainly in the field of architecture. This has oriented her research interests towards the meaning of architecture and landscape in literature, not only for research purposes but also for its didactic value as reading material within specific contexts.

Jorge Arús Hita (jarus@filol.ucm.es) teaches English language and linguistics at the Facultad de Filología, Universidad Complutense de Madrid, where he earned his Ph.D. in English Linguistics in 2003. His publications include articles on corpus and contrastive linguistics as well as on second-language teaching in various national and international journals and edited volumes. He is copy-editor of *Atlantis* and blended-learning coordinator at the Facultad de Filología, UCM.

Julia Lavid, Ph.D. (lavid@filol.ucm.es) is Full Professor in English Linguistics at the Department of English Philology I, Universidad Complutense of Madrid (Spain), where she teaches several courses on English Linguistics, Computational and Corpus Linguistics, and the contrastive analysis and translation of English and Spanish. Her research expertise focuses on functional and corpus-based approaches to the study of English in contrast with other languages, as well as their application to educational and computational contexts. As the team leader of several international projects, she has published extensively in these research fields. She is the author of the book *Lenguaje y Nuevas tecnologías: Nuevas Perspectivas, Métodos y Herramientas para el lingüista del siglo XXI* (Madrid, Cátedra, 2005), and co-author of the research monograph *Systemic-Functional Grammar of Spanish: A Contrastive Account with English* (London: Continuum, in press).

Penny MacDonald (penny@idm.upv.es) is a member of the Department of Applied Linguistics and lecturer in English for academic and professional purposes at the Universidad Politécnica de Valencia, Spain. Her main research interests lie in the field of Corpus Linguistics, Critical Discourse Analysis and the analysis of interlanguage errors in foreign language learning.

María José Luzón Marco, Ph.D., (mjluzon@unizar.es) is Senior Lecturer in English for Specific Purposes at the University of Saragossa, Spain. She has a Ph.D. in English Philology and has published papers on academic and professional discourse and on language teaching and learning in the field of English for Specific Purposes in national and international journals. Her current research interests include corpus-based research of academic and professional discourse and the use of new technologies in English language teaching and learning.

María Boquera Matarredona (mboquera@idm.upv.es) holds a degree in English and German Philology from the University of Valencia (UV) (Spain). In the following two academic years she attended the Universidad Complutense of Madrid where she obtained a Post-Graduate Certificate in Translation. She has also worked as an auxiliary translator for the Spanish Division of Translation of the European Parliament in Luxemburg.

Amaya Mendikoetxea, Ph.D. (York, amanya.mendikoetxea@uam.es) is a Lecturer in English Syntax in the Department of English Philology at the Universidad Autónoma de Madrid. Her research interests include both theoretical linguistics (syntax and lexicon) and applied linguistics (second language acquisition and corpus linguistics). She has published widely in those areas and has directed several externally funded research projects.

Magali Paquot (magali.paquot@uclouvain.be) is a postdoctoral researcher at the Centre for English Corpus Linguistics (UCL, Belgium) and a fellow of the Belgian National Fund for Scientific Research (FNRS). Her research interests include EAP vocabulary, phraseology and corpus-based analysis of L1 transfer in Second Language Acquisition. In 2008, she launched the Varieties of English for Specific Purposes database (VESPA) project which aims at collecting a large corpus of learner texts in a wide range of disciplines, genres and degrees of writer expertise in academic settings.

Pascual Pérez-Paredes (pascualf@um.es) started his collaboration with the English Department in the University of Murcia, Spain in 1996. After a research stay in the University of Texas at Austin, he completed his Ph.D. in Applied Linguistics in 1999. He currently teaches CALL, Legal English and Applied Linguistics. He is also an Official Translator. His main interests are quantitative research of register variation, the compilation and use of language corpora and the implementation of Information and Communication Technologies in Foreign Language Teaching/Learning. He is a member of the Research Group Lingüística Aplicada Computacional, Enseñanza de Lenguas y Lexicografía (LACELL). Pascual Pérez-Paredes has been project coordinator of a MINERVA initiative funded by the European Commission: SACODEYL (http://www.um.es/sacodeyl) and at the moment coordinates Corpora for Content & Language Integrated Learning [BACKBONE], a LLP K2 Transversal programme.

Carmen Pérez-Sabater, Ph.D. (cperezs@idm.upv.es) Associate Professor, has been lecturing in English for Computer Science at the Universidad Politécnica de Valencia (Spain), Department of Applied Linguistics, since 1990. She is currently working in the field of Comparative Discourse Analysis, Computer-Mediated Communication and Language Teaching in a university environment.

Josep Roderic Guzmán Pitarch (guzman@trad.uji.es) is Senior Lecturer at the University Jaume I, Spain. His research ranges from discourse analysis to translation studies. He has translated several books, movies and TV series. His publications focus on language learning, pragmatics and translation for language teaching. He has coordinated several research projects on the use of corpora in translation.

Mercedes Querol-Julián (querolm@ang.uji.es) is a predoctoral research fellow at the Department of English Studies of the Universitat Jaume I. Her research interests are in the areas of corpus linguistics and discourse analysis, with particular focus on the analysis of academic spoken discourse and the development of multimodal corpora. Currently, she is writing her Ph.D. thesis, *Discussion sessions in specialised conference presentations: a multimodal approach to analyse evaluation.*

Tom Rankin (tom.rankin@wu-wien.ac.at) is a teaching and research assistant at the Vienna University of Economics and Business. He is completing a Ph.D. in second language acquisition at the University of Vienna.

Paul Rollinson, Ph.D. (Universidad Autónoma de Madrid, paul.rollinson@uam.es) is a lecturer in English Language (Writing) in the Department of English Philology at the Universidad Autónoma de Madrid. His research interests are within the area of applied linguistics and foreign language teaching, especially in relation to the teaching of Writing. He has participated in several research projects and has coordinated the compilation of WriCLE (Written Corpus of Learner English).

Ute Römer (uroemer@umich.edu) is currently Director of the Applied Corpus Linguistics Unit at the University of Michigan English Language Institute where she manages the MICASE (Michigan Corpus of Academic Spoken English) and MICUSP (Michigan Corpus of Upper-level Student Papers) projects. Ute's primary research interests include corpus linguistics, phraseology and the application of corpora in language learning and teaching. Her current research focus is on how corpus tools and methods can be used to identify meaningful units in specialized discourses.

Eva Alcón Soler (alcon@ang.uji.es) is Full Professor at the University Jaume I, Spain. She has been working on discourse and language learning since 1993. Her research focuses on interlanguage pragmatics, interaction and second language acquisition. Her publications include *Intercultural language use and language learning, Investigating pragmatic learning, teaching and testing in foreign language contexts.* She has been guest editor for different journals.

Encarnación Tornero Valero (nanitornero@yahoo.es) has been a Spanish as a Foreign Language teacher for a decade now. She has been involved in teaching immigrants in slum areas, incorporating her Psychology background to the field. She holds a Diploma in Español como Lengua Extranjera, and has taken part in the transcription and annotation process of the Spanish component of SACODEYL (http://www.um.es/sacodeyl).

Juan Rafael Zamorano-Mansilla (juanrafaelzm@yahoo.es) has been Lecturer in English Studies at the Universidad Complutense de Madrid since 1998. He received his Ph.D in 2006 for his work on the automatic generation of tense and aspect in English and Spanish (*La generación de tiempo y aspecto en inglés y español: un estudio funcional contrastivo,* Editorial Complutense). His research interests lie primarily in tense, aspect and modality, natural language generation and attributive constructions.

Acknowledgements

This book has been funded by a grant from the Ministerio de Educación y Ciencia (Spain) HUM2007/30057-E

MINISTERIO
DE EDUCACIÓN
Y CIENCIA

Part One

Corpus Linguistics and ELT: State of the Art

This page intentionally left blank

Chapter 1

Introduction to Corpus Linguistics and ELT

Mari Carmen Campoy-Cubillo, Begoña Bellés-Fortuño
and Maria-Lluïsa Gea-Valor
Universitat Jaume I, Castelló

From its origins, Corpus Linguistics has had a strong link with language teaching. John Sinclair's impact on dictionary making and his pioneering work on corpus research (Sinclair 1987, 1991, 2004) have been the starting point for many corpus-based approaches to language teaching (Wichmann et al. 1997; Burnard and McEnery 2000; Granger et al. 2002; Kettemann and Marko 2002; Aston et al. 2004; O'Keefe et al. 2007; Aijmer 2009, to name but a few). The common ground for all these approaches is that they are based on empirical evidence, thus leading to the elaboration of better quality learner input and providing teachers and researchers with a wider, finer perspective into language in use, that is, into the understanding of how language works in specific contexts.

Corpus-Based Approaches to ELT presents work by leading linguists exploring different ways of applying corpus-based and corpus-informed research to language teaching environments. More specifically, the volume tackles three main areas of special interest today: the use of corpora for teaching English for Specific Purposes, pedagogically motivated uses of corpora, and the potential of corpora-mediated multimodal tools for the language learning context.

The compilation, description and analysis of domain-specific corpora is one of the widest areas of research in corpus linguistics, especially as regards academic and professional settings. This book provides an in-depth analysis of academic and professional texts by means of corpus-based methodologies in order to enhance English for Specific Purposes (ESP) teaching. A wide perspective into ESP corpora is offered, as the chapters include written and spoken academic discourse, the use of English language in professional contexts, and the use of both native English speaker corpora and ESP learner corpora, that is, corpora in which learners attempt at producing professional texts.

The second issue examined in this volume has to do with how English language teaching may benefit from corpus data to improve language learner input (the so-called corpus-based and corpus-informed approaches) and the different ways in which corpora may aid in understanding learner and teacher discourse. In this sense, the volume illustrates the way corpora may be used directly in the classroom and how corpus research may be applied to inform syllabi and classroom materials.

Finally, the third dimension reflects on the role of corpus tools and multimodal devices, where corpora-based research plays a central role to inform teaching materials. Multimodal corpora are still in their infancy when compared to corpora where only one discourse mode is used. Challenges in this area lie not only in the design of such corpora, a difficult task per se, but also in the reflection on how information is organized and connected among the different text modes. Far from being just an inclusion of one or more corpora within a learning package and allowing users access to concordance and collocational information, this entails having a clear idea of the pedagogical goals of both tool and tool applications and how corpora are integrated in the tasks a learner is intended to carry out. It also implies a lot of research into feasible text mode combinations and consensus on issues such as possible tagging categories and terminology in order to be able to contrast studies carried out by different researchers.

The volume opens with Ute Römer's chapter, in which she presents and discusses the state of the art in the field of corpus linguistics and language teaching. The author provides an overview of the past, present and future developments in corpus linguistics, reviewing the applications of general and specialized corpora. Römer insightfully points at the need to foster the use of pedagogical corpora and draws a work agenda around three main topics: focus on learner and teacher needs, indirect uses of corpora in language teaching and direct uses of corpora in language teaching.

From this introductory chapter, the volume goes on to study the close relationship between corpus linguistics and language teaching, and is divided into three more Parts, namely Corpora and English for Specific Purposes; Learner Corpora and Corpus-Informed Teaching Materials; and Multimodality: Corpus Tools and Language Processing Technology.

1.1 Corpora and English for Specific Purposes

Part I of this book contains six chapters describing various scenarios related to the field of *English for Specific Purposes* (ESP), including academic and professional settings. Although both learner and expert corpora are discussed in this part, it should be pointed out that (mostly small) learner corpora are frequently used in the ESP field, since teachers are concerned with the production of their students in contexts of specialization. While general corpora have proven to be most effective for the study of the structure and use of language, specialized corpora which focus on specific genres are required when exploring language in specific academic and professional settings (Connor and Upton 2004a). According to Flowerdew (2004), specialized smaller corpora offer more advantages than general corpora from a methodological perspective because they provide more contextual information (i.e. the communicative situation) than larger corpora. When complete texts are included, the implementation of top-down analyses of the textual and generic features present in the texts is made feasible.

Similarly, genre analysis clearly benefits from the use of specialized corpora, which help to grasp more accurately the function and use of language in genre. In this sense, corpus linguistics reveals itself as an essential and indispensable framework which, combined with genre analysis (Swales 1990; Bhatia 1993), may provide new insights and ultimately help ‘to improve the training of novice writers and to encourage the development of better and more effective [texts]’. (Connor and Upton 2004b: 254).

The first two chapters in the part Corpora and English for Specific Purposes study the use of written and spoken academic English corpora. Annelie Ädel (Chapter 3) provides a thorough review of the challenges that lie ahead in the use of corpus for the teaching of academic writing. She discusses the scarce attention paid to the potential of corpora in the context of writing instruction. As she rightly states (Ädel, this volume: 41): at this point in time, it takes a corpus linguist to offer a corpus-based writing class.’ To alleviate this situation, she presents seven different challenges involved in using corpus-based approaches in teaching writing in English for Academic Purposes (EAP) settings. Among these challenges we find the lack of corpus availability; the difficulty of finding what users are looking for, where and how, without getting lost in large amounts of data; how to evaluate and present corpus patterns to language learners; how to manage decontextualized data; and how to connect surface forms to meaning.

English subject curricula should take into account language aspects that go beyond linguistic features to introduce real language into the classroom. Thus, in Chapter 4, Begoña Bellés and Mari Carmen Campoy explore the uses of the phrase ‘I feel’ and its variants in contiguous and non-contiguous collocational patterns by analysing MICASE data, indicating lexico-grammatical and textual features that should be taken into account in the teaching of communicative functions in spoken academic discourse at a wide range of linguistic levels. They suggest that these findings should be included in class so as to contribute to raise the student's awareness of the connection between grammatical, sociolinguistic and pragmatic uses of phraseological items of the English language. The analysis of 'I feel' is a good example of how the MICASE search engine may aid in teaching the use of a stance verb by highlighting it as marked in terms of uneven distribution among genres, speech event interactivity rate and in its use among different genders. This is a complex teaching approach to the analysis and understanding of modality devices that teachers may only carry out thanks to the annotation of speaker and speech-event categories that the corpus search interface makes possible.

The following four chapters in this section (by Winnie Cheng, María José Luzón, Belinda Crawford, and Maria Georgieva and Lilyana Grozdanova) deal with corpora and English for Professional Purposes (EPP). An interesting feature of learner corpora in this context is that text or speech production on the part of learners does not usually coincide with the text types and genres collected in native speaker corpora. In the area of EPP, however, this situation is changing. EPP teachers are now gradually becoming more engaged in trying to get their students to produce texts based on language use situations in which they might find themselves in their future as professionals in a specific area of work.

Cheng (Chapter 5) shows how *ConcGram©* (Greaves 2009) may be used to elicit data from a corpus representing the English language of the engineering sector in Hong Kong, and discusses how the results may be used to deal for instance with the aboutness of the text and to help EPP students to learn the language used in their profession. Regarding the use of NS and NNS corpora for ESP teaching (see Gavioli 2005 for ESP corpora designed with teaching purposes rather than for language description), characteristic discourse moves may be studied by learners so that they become aware of those common expressions that are typical of the genre under analysis within the wider perspective of move sequencing.

Luzón (Chapter 6) studies the misuse or atypical use of organizational items in a small learner corpus in contrast to the information gathered from the BNC corpus. The problematic areas found in Luzón's study include errors regarding the word class, meaning, or function of an item and its position in a sentence, as well as atypical or incorrect use of (sometimes inexistent) lexical bundles and genre phraseology. Errors discussed in Luzón include those involving signalling nouns and their use to create cohesive relations across-clause level. Likewise, problems regarding the use of informal or oral discourse in a formal context are brought to light. In this chapter it is made clear that in order to design effective teaching materials it is essential that both native speaker and learner corpora should be brought together to better understand learner's needs and problematic areas in order to identify language patterns used by learners which clearly differ from those used by experts.

In Chapter 7, Crawford introduces a spoken business corpus and derived classroom activities that may improve ESP materials through corpus-based pedagogical applications. Drawing on a small specialized corpus, the author explores key business English lexis and demonstrates that corpus-based activities can help students better understand content lectures in English. This is vital for the learners' success not only in their academic studies but also in their future careers. In this way, Crawford (this volume, 104) contributes to 'bridging the gap between ESP research and ESP pedagogy'.

The last chapter in this part, by Georgieva and Grozdanova, pays special attention to English as an International Language (EIL) or English as a Lingua Franca (ELF) corpora (Seidlhofer 2005; Mauranen 2007). EIL/ELF corpora are particularly focused on the production of native-like speakers in academic and professional contexts. For the majority of ESP learners, competent professional communication is one of the highest motivations to learn a language. Georgieva and Grozdanova intend to answer a different set of questions, such as which strategies participants in intercultural communicative encounters use to overcome differences in the process of communicating with other speakers; or which are the most widely used patterns that come up in order to communicate successfully.

1.2 Learner Corpora and Corpus-Informed Teaching Materials

The corpora explored in this part may be termed *pedagogic corpora* (Hunston 2002) or *(E)LT discourse corpora*,[1] in a similar fashion to EIL corpora. They include the language used in classroom or in formal teaching and learning contexts and situations (exams, office tutorials, etc.) and may take into account teacher-learner relationship patterns. A comprehensive example of this kind of corpora is the T2K-SWAL corpus designed to test to which extent the language of ESL/EFL materials and assessment instruments represents 'real' English language (Drescher 2007; García 2007). This group would also include English Language Teaching (ELT) materials corpora, in the sense that textbooks, for instance, are meant to represent NS production as a model for the language learner (Römer 2005; Amador-Moreno et al. 2006; Cheng 2007). Authenticity of the written/spoken texts is questioned here in terms of the language used, the text types provided and the authenticity of tasks. Corpus Linguistics has a lot to say in the assessment/improvement of the aforementioned levels of authenticity. Corpora based on the interaction between teachers and learners which should be considered EPP corpora would fall into this category. Examples of analysis of this interaction may be seen in the MICASE corpus (Csomay 2007), or the POTTI corpus (Farr 2007) and also in O'Keefe et al. (2007: 220–243).

In this volume, the part devoted to ELT corpora focuses on three main dimensions: the first one deals with the compilation and exploitation of learner corpora; the second explores error analysis using learner corpora and comparable native speaker corpora; and the third has to do with the use of corpora to create teaching materials.

The compilation and use of corpora as a means to enhance language learning practices takes us to the issue of criteria in corpus compilation which determine the end product and how and by whom it may be used afterwards (Luzón et al. 2007: 4–6). Among these, there are at least three essential criteria that affect corpus-based language learning and teaching: (1) the purpose and principles behind the compilation of the corpus, (2) its availability, not only for the researcher but also for materials writers, teachers and learners and (3) the use of various resources in multimodal corpora. In this sense, as may be seen in the articles collected in Ghadessy et al. (2001), it is a well-known fact that a good number of teachers prefer the use of small *ad hoc* corpora that have been designed with a very specific aim in mind and addressed to a particular group of learners. There are two obvious reasons for this: one is that, given the opportunity, teachers would not avoid the possibility of tailor-made resources; the other is that, in most cases, small *ad hoc* corpora are easier to handle in the classroom.

If we consider the issue of corpus compilation purposes, another interesting feature stands out: how the texts are obtained, i.e. the compilation methodology. Thus, we think that an important point when dealing with corpus-based methodologies is that learner corpora follow a task-based instead of a text-type based approach in their compilation and database organization. This takes us to the subject of how learner corpora differ from corpora with other speaker profiles. In learner corpus compilation, an important debate revolves around the kind of task selected to elicit learner language production, and the extent to which the elicited language may be seen as authentic. In this sense, it is important to bear in mind that any chosen task for the learners is not going to be considered as natural as those performed by native speakers since the former are produced in a more or less imposing learning situation where fully spontaneous speech may not be attained, though it may be argued, as in Sylvie De Cock's chapter (Chapter 9), that the learning situation is in fact a real situation for learners.

In her chapter, De Cock extensively reviews the use of spoken learner corpora in ELT. She discusses the two fundamental aspects in learner corpora: learner variables and task variables. Learner variables pose a number of questions regarding the complexity of the description of speaker profiles and of the compilation of speaker production corpora where speakers follow the same procedures and belong to a similar learning profile. Task variables largely influence not only what may be done with the corpus in question but also the possibility of research replications in subsequent investigation. For the creation and analysis of oral tasks, communication problems arising from inability to convey a message are one of the main concerns when querying corpora and they constitute a central issue when designing pedagogically relevant materials.

Moreover, De Cock complains about the scarce availability of materials derived from spoken corpora, which are also still in its infancy regarding classroom exploitation. Direct and indirect use of spoken learner corpora requires participation on the teachers' part that could at this stage perhaps only be carried out if the teacher is a corpus linguist or is trained specifically to deal with such corpora, since spoken corpora are difficult to handle at least in depth or to obtain as many benefits as possible on the part of both learners and teachers.

In the second chapter of this section, Julia Lavid, Jorge Arús and Juan Rafael Zamorano explain details about the compilation and exploitation of a small bidirectional corpus of written texts. The texts in their online corpus include originals and their translations in English and Spanish, and allow for the analysis of individual texts as well as for 'whole-corpus reading'. In an effort to guide teachers and learners, the authors also include other tasks which would fit into what is called direct use of corpora, designing possible hands-on tasks as part of their corpus-based materials.

Regarding the use of corpora to analyse learner output, Chapters 11, 12 and 13 (by Rafael Alejo, Mª Ángeles Andreu et al. and Amaya Mendikoetxea et al., respectively) explicitly deal with corpus-based error analysis and learners' non-prototypical use of English. Many studies analysing learner corpora focus to a large extent on language proficiency and on possible errors in a set-up task. In Chapter 11, Alejo explores the Spanish and Swedish components of the ICLE corpus and the Written School and University Essays from the BNC to compare the use of the particle 'out' in both corpora in terms of over- and underuse, prototypicality, avoidance and erroneous use of this particle. Similarly, Andreu et al. (Chapter 12) analyse written production of EFL students in an error-annotated multilingual corpus of students learning English, Spanish, French and German as a foreign language, and also Catalan, as a first, second or foreign language. Comparable and parallel multilingual corpora incorporate the production of speakers (NS or NNS) whose mother tongue may represent two or more languages. They are most common in corpus-based translation studies. The possibilities are varied: researchers, teachers and students may be using comparable and/or parallel corpora in two or more languages to analyse possible translations and/or to check on a specific language issue. Other multilingual corpora discussed in this volume may be found in other parts (see Lavid et al.; Alcaraz et al.; Guzmán and Alcón).

Mendikoetxea et al. (Chapter 13) aim at the development of teaching materials drawing on a database of learner errors extracted from a corpus of essays written by Spanish learners of English at university level in order to identify problematic areas and to develop relevant pedagogical materials, thus improving curriculum design. Their project (INTELeNG) combines contrastive analysis (CA) and error analysis (EA). Despite advocating for the use of learner corpora, the authors highlight the benefits of the combination of learner and native corpora for the elaboration of teaching materials and curriculum design as part of classroom methodology aimed at fostering students' language awareness and, ultimately, their language proficiency.

The last three chapters of this part deal with the creation of corpus-informed language teaching materials taking into account lexicography, grammar and representativeness in language learning. Leaner corpora may be used to obtain feedback for the improvement of existing pedagogical materials. In this area, corpus-based updating and improvement of pedagogical dictionaries is one of the most widely exploited fields of research. Grammar and textbook design are now also receiving more attention in the field of indirect corpus applications. Cheng (2007) and Römer (2005) are examples of how differences between actual language use and textbook language may be tackled by means of corpus analysis.

In Chapter 14, Sylvie De Cock and Magali Paquot discuss the design of corpus-based information in dictionaries that are meant to aid learner language production. They focus on the work carried out in the *Macmillan English Dictionary for Advanced Learners* to describe how the International Corpus of Learner English (with writings of learners from 16 different countries) is used, together with information drawn from a 15 million-word corpus of academic English in order to provide improved information on those areas where difficulty was detected in the learner corpus. Thus, corpus information is an added value in the form of 'Get it right' boxes, grammar sections and academic writing sections, increasing the dictionary's productive use potential.

Chapter 15 also deals specifically with corpus-informed teaching and learning materials. Here, Tom Rankin analyses adverb placement in an advanced learner corpus suggesting ways to improve grammar teaching materials. Adverb syntax is a particularly problematic area for EFL learners but, paradoxically, it has been neglected in most grammar textbooks. Rankin contends that specific discourse and pragmatic contexts must be taken into account when teaching adverb placement and suggests that corpus data can inform the selection and sequencing of materials and 'provide practical help in choosing which type of semantic and syntactic features prove most problematic for the learners and should therefore be included in teaching examples and exercises' (Rankin, this volume: 305).

Finally, the issue of representativeness in corpora use and compilation is discussed by Izaskun Elorza and Blanca García-Riaza in Chapter 16. These authors tackle the question of how the compilation of a successful pedagogical corpus of written academic texts should be done in terms of size, topic, authenticity and representativeness. The focus remains on the texts chosen for the corpus, since learners will take them as a model of the language used for 'real' communication. The authors suggest that the use of a specific (pedagogic) corpus can influence the definition of the model of language to be used in the classroom. As the authors indicate (Elorza and Riaza, this volume: 221),

> when dealing with the corpus compilation of the written input we cannot ignore the great variety of the texts used in higher education courses. The need for using texts from different types seems to impede the very possibility of compiling a representative corpus in terms of typological representativeness.

Thus, they study wordlist statistics, rank and frequency of word types in relation to text length, completeness and representativeness and compare and contrast data to the first hundred most frequent words in the BNC corpus.

1.3 Multimodality: Corpus Tools and Language Processing Technology

The development of corpus tools and the integration of different modes of communication in corpora are key issues in the use of corpora for learning purposes. Also, CD and online availability allow both learner and teacher to use corpus resources at ease. Together with this availability is the issue of user-friendliness in the design of both corpus and corpus tools. The fact that most educational institutions have access to the internet has promoted the use of the web as corpus (Kilgarriff and Grefenstette 2003; Sharoff 2006) in the making of self-compiled *ad hoc* corpora, since educators worldwide find it easy to download the exact text types they need to use in the classroom and make them part of a corpus in a do-it-yourself fashion (e.g. *CorpusBuilder* in *SketchEngine*). In this sense, web as corpus research facilitates the study of multimodal features through the use of corpora. Moreover, the development of customized corpora such as ACORN (the Aston Corpus Network) and its focus on, and open access to, corpus and corpus output materials show how corpora are increasingly present in today's educational institutions.

Some CD and online language learning packages also include corpora as part of their components (see for instance the Virtual Language Centre, Hong Kong Polytechnic at http://www.edict.com.hk/vlc/ and its *WebConcordancer*), so learners may play around with several search routes which allow for teacher work on various language proficiency levels. Furthermore, with the combination of different discourse modes in multimodal corpora, learners may develop all four competences.

We would also like to point out the advances that have been made since Tim Jones' pioneering work in DDL (Data Driven Learning), when most research was based on concordance and collocation data. The future that lies ahead regarding corpus tools that may be used by learners and teachers alike is more complex and exciting than ever. First, the availability of a wide range of corpora, which may be operated through diverse corpus tools, enables teachers to design a wide range of materials and tasks for the classroom. The creation of corpora such as MICASE including speech events, speaker status and academic position, speaker level or interactivity rating of the event, makes it possible to go beyond the word and its lexico-grammatical patterns into other discourse levels. Secondly, a surge for pedagogic annotation and annotation tools (Braun 2006; Alcaraz et al., this volume) reflects the interest of teachers and researchers alike to use annotated corpora in the classroom and in the creation of language teaching materials. New corpus tools such as *SketchEngine* and its *Word Sketch* automatically provide the user with a complete collocational and grammatical pattern of searched words and phrases; others, like the *Word Sketch differences*, show lexical contrasts between two selected words in terms of their collocates.

The possibility to study word association and the combination of genre and keyword analysis (Scott and Tribble 2006) by means of tools such as *WordSmith Tools* gave corpus studies a wider dimension. The development of new tools in this direction may be seen in *ConcGram@* (Greaves 2009), a programme which determines the phraseological profile of the language contained in a specific corpus. As described in Cheng et al. (2006), many word associations do not occur in one fixed grammatical pattern so, taking this into account, *ConcGram@* develops information based on non-contiguous sequences of associated words (Cheng et al. 2006: 414):

> The development of the notion of a concgram challenges the current view about word co-occurrences that underpins the KWIC display (. . .) word associations become the focus of attention, and a 'node' is not the 'sun' around which collocates orbit in a subordinate relationship.

As can be observed, tool and multimodality play an active role in the development of corpus-based approaches to ELT. The final part of this volume examines availability and multimodality in corpora within the language teaching context, and presents several new devices for corpus processing, introducing tools such as a query program for parallel corpora or a tool for implementing pedagogical annotation. The chapters discuss the opportunities and challenges that multilayered and multimodal corpora may pose to corpus linguistic investigation in ELT.

More specifically, José María Alcaraz et al. (Chapter 17) show a tool that allows annotation for any language and explain how the seven language corpora in the SACODEYL project can be annotated with the same tool, thus providing useful resources in the pedagogical design and analysis of classroom material. In Chapter 18, Josep Roderic Guzmán and Eva Alcón use two corpora made up of TV series in English which are translated into Catalan and Spanish, and narrative works where English is the language of the original texts, also translated into Catalan. They explain how these corpora may be approached by means of the AlfraCOVALT tool in order to design tasks to cater for the use of requests in English, and how to apply the data provided by the corpora to the creation of activities for translation students.

In Chapter 19, Inmaculada Fortanet and Mercedes Querol present their experience in the compilation of a multimodal video corpus recorded and edited for its application to a teacher training course for lecturing in English at Universitat Jaume I. These authors offer an example of the type of tagging or classification of speech events that can be done in video corpora, which can later assist the teaching of pragmatics, grammar and/or vocabulary. They advocate for multimodal corpus analysis, stating that when teaching spoken academic discourse by means of corpus-based learning, corpora transcripts do not always provide enough information about the real situation, lacking of general context and background. They conclude that language is accompanied by prosodic features such as intonation, accent, or stress and kinesics which cannot be exclusively analysed from a transcript.

If there is a promising future in corpus studies in the ELT field, it is that of multimodal corpora and the tools developed to support them. The study and analysis of multimodal corpora could be understood as a critical rethinking and reformulation of the relationship between text and society (Baldry and Thibault 2006: 2). This provides researchers with other language, social and cultural aspects not gathered or embedded within a linear approach or analysis. A not-single theoretical framework such as the analysis of multimodal corpora, can in fact adequately describe the very different semiotic systems (language, music, picture, movement, etc.). By analysing multimodal corpora, researchers do not only aim at the study of plain texts or transcripts, but other modes of discourse are taken into consideration and seen as a unique whole. How all these modes of discourse are interrelated or not, structured, organized and presented, can be studied by means of multimodal corpora. As Baldry and Thibault (2006: 3) point out, 'text users' knowledge of culture and society interact with the internal features of text's organization during the making and interpreting of texts.' It should be added at this point that we, as linguists, understand text not only as a written mode of discourse. With multimodal corpus analysis we are not limited to text analysis; there are many other resources that can be used to create or support texts, a phenomenon which has been referred to as *resource integration principle* (Baldry and Thibault 2006: 4).

Many are the ideas to be drawn from this volume. However, we would like to underscore two central issues. One is the fact that research-oriented corpus tools have still a lot to say in indirect corpus applications, that is, on what and when to teach. This is more so for learner corpora and for spoken corpora, due to the fact that these are the most difficult corpora to compile but are also, or should be, more productive in terms of providing data that may be applied to the classroom in a satisfactory way. The same is true about teaching-oriented corpus tools and their role in direct corpus applications, since the development of these tools together with the analysis of teacher and learner needs will undoubtedly lead to a more active participation of both teachers and learners in the corpus-based learning process, that is, in how we may teach and learn a language. In this sense, we can remain assured that the future of corpus linguistics and language teaching will go hand in hand to provide valuable and much needed pedagogical applications, to improve teaching materials and course syllabi, and ultimately to respond to the needs of both teachers and learners.

A second, final issue concerns the reflection made around concepts introducing the prefix 'multi' in combinations such as 'multilayered', 'multimodal', 'multipurpose', 'multilingual', 'multiple tools', 'multiple annotation', etc. We would like to take the 'multi-combinations' terms used throughout this volume as an emblem towards the new and exciting challenges that the new corpora and updates of the old ones bring on to the stage for corpus linguistics and ELT.

Notes

[1] The term pedagogical corpora might imply study and evaluation of that discourse as pedagogical, without questioning the efficiency of that discourse in learning contexts. Use of teacher and teaching materials corpora may sometimes reveal a bigger or lesser degree of pedagogical inadequacy.

References

Aijmer, K. (ed) (2009), *Corpora and Language Teaching*. Studies in Corpus Linguistics 33. Amsterdam/Philadelphia: John Benjamins.

Amador-Moreno, C., Chambers, A. and O'Riordan, S. (2006), 'Integrating a corpus of classroom discourse in language teacher education: the case of discourse markers'. *ReCALL,* 18, (1), 83–104.

Aston, G., Bernardini, S. and Stewart, D. (eds) (2004), *Corpora and Language Learners.* Amsterdam/Philadelphia: John Benjamins.

Baldry, A. and Thibault, J. (2006), *Multimodal Transcription and Text Analysis.* Cardiff: Equinox Textbooks and Surveys in Linguistics.

Bhatia, V. K. (1993), *Analysing Genre: Language Use in Professional Settings.* London: Longman.

Braun, S. (2006), 'ELISA: A pedagogically-enriched corpus for language learning purposes', in Braun, S., Khon, K. and Mukherjee, J. (eds), *Corpus Technology and*

Language Pedagogy. English Corpus Linguistics. 3. Frankfurt: Peter Lang, pp. 25–48.

Burnard, L. and. McEnery, T. (eds) (2000), *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt am Main: Peter Lang.

Cheng, W. (2007), '"Sorry to interrupt, but . . .": Pedagogical implications of a spoken corpus', in Campoy, M. C. and Luzon, M. J. (eds), *Spoken Corpora in Applied Linguistics*, Linguistic Insights 51. Bern: Peter Lang. pp. 199–216.

Cheng, W., Greaves, C. and Warren, M. (2006), 'From n-gram to skipgram to ConcGram'. *International Journal of Corpus Linguistics*, 11, (4), 411–433.

Connor, U. and Upton, T. A. (eds) (2004a), *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins.

—(2004b), 'The genre of grant proposals: A corpus linguistics analysis', in Connor, U. and T. A. Upton (eds), *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins, pp. 235–255.

Csomay, E. (2007), 'A corpus-based look at linguistic variation in classroom interaction: Teacher talk versus student talk in American University classes'. *Journal of English for Academic Purposes*, 6, (4), 336–355.

Drescher, N. (2007), 'Linguistic variation in U.S. universities: a multidimensional analysis of spoken language', in Campoy, M. C. and Luzón, M. J. (eds), *Spoken Corpora in Applied Linguistics*. Linguistic Insights 51. Bern: Peter Lang, pp. 77–95.

Farr, F. (2007), 'Spoken corpus analysis as a tool for reflective practice in language teacher education: quantitative and qualitative insights', in Campoy, M. C. and Luzón, M. J. (eds), *Spoken Corpora in Applied Linguistics*. Linguistic Insights 51. Bern: Peter Lang, pp. 235–258.

Flowerdew, L. (2004), 'The argument for using English specialized corpora', in Connor, U. and Upton T. A. (eds), *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins, pp.11–33.

García, P. (2007), 'Pragmatics in academic contexts: a spoken corpus study', in Campoy, M. C. and Luzón, M. J. (eds), *Spoken Corpora in Applied Linguistics*. Linguistic Insights 51. Bern: Peter Lang, pp. 97–126.

Gavioli, L. (2005), *Exploring corpora for ESP learning*. Amsterdam: John Benjamins.

Ghadessy, M., Henry, A. and Roseberry, R. (eds) (2001), *Small Corpus Studies and ELT Theory and practice*. Amsterdam: Benjamins.

Granger, S. Hung, J. and Petch-Tyson, S. (eds) (2002), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Language Learning and Language Teaching 6. Amsterdam/Philadelphia: John Benjamins.

Greaves, Ch. (2009), *ConcGram© 1.0. A Phraseological Search Engine*. Studies in Corpus Linguistics Software 1. Amsterdam/Philadelphia: John Benjamins.

Hunston, S. (2002), *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kettemann, B. and Marko, G. (eds) (2002), *Teaching and Learning by Doing Corpus Analysis*. Amsterdam; New York: Rodopi.

Kilgarriff, A. and Grefenstette, G. (2003), 'Web as Corpus. *Introduction to the Special Issue on Web as Corpus*'. *Computational Linguistics*, 29, (3), 1–15.

Luzón, M. J., Campoy, M. C., Sánchez, M. M. and Salazar, P. (2007), 'Spoken Corpora: New Perspectives in Oral Language Use and Teaching', in Campoy, M. C. and

Luzón, M. J. (eds), *Spoken Corpora in Applied Linguistics*. Linguistic Insights 51. Bern: Peter Lang. 3–26.

Mauranen, A. (2007), 'Investigating English as a lingua franca with a spoken corpus', in Campoy, M. C. and Luzón, M. J. (eds), *Spoken Corpora in Applied Linguistics*. Linguistic Insights 51. Bern: Peter Lang, pp. 33–56.

O'Keefe, A., McCarthy, M. and Carter, R. (2007), *From Corpus to Classroom*. Cambridge University Press.

Römer, U. (2005), *Progressives, Patterns, Pedagogy. A Corpus-driven Approach to Progressive Forms, Functions, Contexts and Didactics*. Amsterdam: John Benjamins.

Scott, M. and Tribble, C. (2006), *Textual Patterns: Keywords and Corpus Analysis in Language Education*. Studies in Corpus Linguistics 22. Amsterdam: John Benjamins.

Seidlhofer, B. (2005), 'Key concepts in ELT. English as lingua franca'. *ELT Journal*, 59, (4), 339–341.

Sharoff, S. (2006), 'Open-source corpora. Using the net to fish for linguistic data'. *IJCL*, 11, (4), 435–462.

Sinclair, J. McH. (ed.) (1987), *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins.

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

—(ed.) (2004), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Swales, J. (1990), *Genre Analysis. English in Academic and Research Settings*. Cambridge: Cambridge University Press.

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

Chapter 2

Using General and Specialized Corpora in English Language Teaching: Past, Present and Future

Ute Römer
University of Michigan

2.1 Introduction: Corpus Linguistics and Language Teaching

Over the past 25 years, corpora, corpus tools and corpus evidence have not only been used as a basis for linguistic research but also in the teaching and learning of languages. Tim Johns's data-driven learning (DDL), Dieter Mindt's empirical grammar research, and John Sinclair's work with COBUILD can be considered particularly groundbreaking developments in the field of English corpus linguistics and language pedagogy in the 1980s (see Mindt 1981,1987; Johns 1986, 1991; Sinclair 1987, 1991).

Nowadays, more and more researchers and practitioners treasure what corpus linguistics has to offer to language pedagogy, and the impressive number of recently published monographs and edited collections on the topic clearly indicate the growing popularity of pedagogical corpora use and the need for research in this area (see, for example, Aston 2001; Granger et al. 2002; Sinclair 2004a; Römer 2005; Ädel 2006; Braun et al. 2006; Gavioli 2006; Kettemann and Marko 2006; Scott and Tribble 2006; Campoy and Luzón 2007; and the proceedings of the first six events in the TaLC (Teaching and Language Corpora) series: Aston et al. 2004; Botley et al. 1996; Burnard and McEnery 2000; Hidalgo et al. 2007; Kettemann and Marko 2002; Wichmann et al. eds. 1997).[1]

I would, however, still be hesitant to say that corpora and corpus tools have after all fully 'arrived' on the pedagogical landscape. The practice of ELT (English Language Teaching) to date, at least, seems to be largely unaffected by the advances of corpus research, and comparatively few teachers and learners know about the availability of useful resources and get their hands on corpus computers or concordancers themselves (see

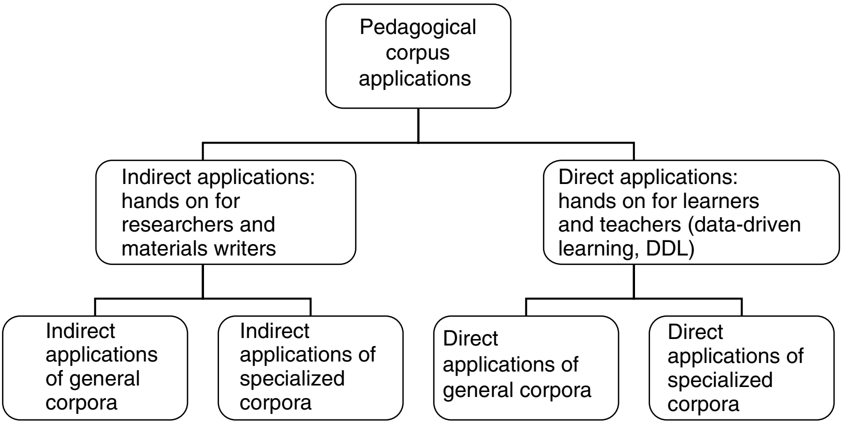

Pedagogical corpus applications
Indirect applications: hands on for researchers and materials writers
Direct applications: hands on for learners and teachers (data-driven learning, DDL)
Indirect applications of general corpora
Indirect applications of specialized corpora
Direct applications of general corpora
Direct applications of specialized corpora

FIGURE 2.1 The use of corpora in language learning and language teaching

e.g. Mukherjee 2004). The aim of the present chapter is to, first, review what has been achieved so far in the field of corpus linguistics and language teaching, and to provide a brief overview of pedagogical applications of general and specialized English language corpora. The chapter then looks at some unresolved issues and future tasks for applied corpus researchers, and discusses what steps could (and should perhaps) be taken in fostering uses of corpora in language learning and teaching. Throughout the chapter, reference will be made to the distinctions illustrated in Figure 2.1, especially the distinction (going back to Leech 1997) between direct and indirect corpora applications.

As Figure 2.1 shows, direct and indirect pedagogical corpus uses apply to both general and specialized corpora. Indirect applications involve hands-on work mainly for corpus researchers as well as, to a limited extent, materials writers and provide answers to questions on *what* to teach and *when* to teach it, whereas direct applications mainly affect *how* something is taught and actively involve the learner and teacher in the process of working with corpora and concordances.

2.2 Corpus Linguistics and Language Teaching: Past

Let us first address the question 'How did it all begin?' How (and when) did corpus linguists and language teachers get together? Perhaps the most important developments in this context took place at the University of Birmingham (United Kingdom) in the early 1980s when John Sinclair (professor in the Department of English) collaborated with Collins publishing on the COBUILD project in pedagogically oriented lexical computing (cf. Sinclair 1987). At the heart of this project, the aim of which was to provide English language learners with better dictionaries and teaching materials that present 'real' English and focus on those items and meanings that learners are most likely to encounter in actual communicative situations, was the Bank of English (BoE), a growing multimillion word corpus of different native-speaker varieties of spoken and written English. The COBUILD learners' dictionaries, grammars and usage guides are fully BoE-based, incorporate findings on frequency distribution and collocations, and contain genuine instead of invented examples. They hence constitute a typical case of indirect application of corpora in ELT.

Other early examples of indirect pedagogical corpus applications (that are perhaps less well-known than the COBUILD dictionaries) are the design of the Collins COBUILD English Course (CCEC, Willis and Willis 1989), a 'lexical syllabus' that focuses on 'the commonest words and phrases in English and their meanings' (Willis 1990: 124), and the work on an empirical grammar of the English verb system by Dieter Mindt (Mindt 1987, 1995, 2000). Mindt and his colleagues at Berlin's Free University (Germany) contributed to both syllabus and materials design and created corpus-driven and frequency-based resources for use by research-oriented teachers and materials designers, mainly addressing the problems of selection of language items and progression in the course. A related but much earlier attempt to improve English language, or mainly English vocabulary teaching based on word-frequency data is Michael West's (1953) *General Service List of English Words* (GSL). Developed in pre-computer corpora times and without the help of software tools for corpus analysis, West's GSL (similar to Willis's CCEC) suggests a syllabus that is based on frequently occurring words rather than on grammatical structures.

A turn from early indirect to the beginnings of direct pedagogical corpus applications leads us back to the University of Birmingham where Tim Johns, inspired by John Sinclair's corpus work and supported by his colleagues Tony Dudley-Evans and Philip King, pioneered the use of concordances in grammar and vocabulary classes in the English for International Students Unit (cf. Johns 1986, 1991). Johns's idea was to put the learner (instead of the teacher) at centre stage and make her/him 'a linguistic researcher' (Johns 2002: 108) who takes on an active role in discovering patterns around and meanings of selected lexical items, often related to problems that were found in learners' academic writing samples. This interaction between the learner and the corpus (or corpus data in the form of concordances) is now usually referred to as 'data-driven learning' (DDL; cf. Johns 1986, 1994). We will get back to the concept of DDL and some concrete examples of its implementation in ELT later (in sections 2.3.1 and 2.3.2).

2.3 Corpus Linguistics and Language Teaching: Present

If we now move on from the 1980s to the early twenty-first century, we notice that much has happened in corpus linguistics and language teaching and that many researchers, language teachers and publishers worldwide have been influenced and inspired by the activities of pioneers like Johns, Sinclair, West and Mindt. The question I would like to address in the following two sections is 'How far have we come, and where are we now in terms of direct and indirect pedagogical corpus applications?'

Corpora and corpus tools come in many different shapes, and not all of them may be equally useful to all groups of learners or for research that can inform teaching resources. It is thus an important task for the applied corpus linguist to guide corpus novices, learners, teachers and materials designers in the selection of the most appropriate resources and to create concordancers and corpora that are easy to use and, ideally, make them freely available, e.g. through online search interfaces. In the following discussion, I will distinguish between general and specialized English corpora throughout. General corpora (also referred to as reference corpora) tend to be fairly large (several million, sometimes even several hundred million words in size) and usually cover a wide range of text types from different registers and different varieties of the language. Typical examples of such corpora are the above-mentioned COBUILD Bank of English (BoE), the British National Corpus (BNC) and the recently launched BYU Corpus of American English (cf. Davies 2008; renamed 'Corpus of Contemporary American English', COCA). The BNC, like COCA and parts of the BoE, are freely accessible online via web-interfaces that allow the user to create concordances and extract lists of collocations for specified words or phrases (see Appendix for a list of web addresses for online-searchable corpora).

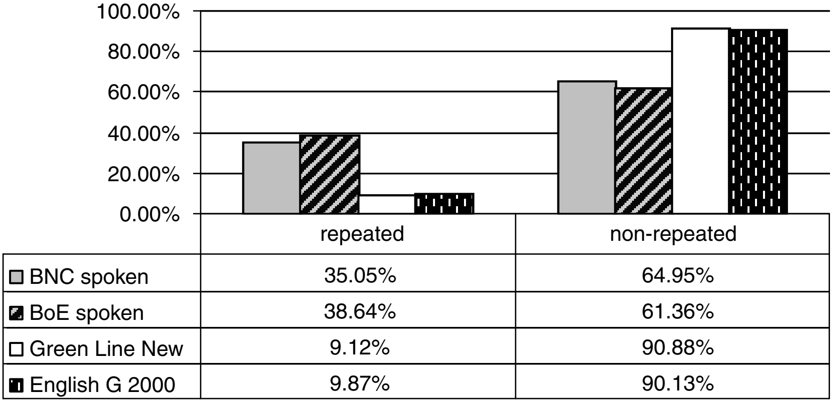
Different from general corpora, most specialized corpora, i.e. collections of texts from a particular field of expertise (e.g. economics), produced by a narrowly defined group of language users (e.g. advanced learners of English whose L1 is Swedish), or produced in a certain setting (e.g. in biology study groups at a US university), are small, often home-made and custom-compiled, and not generally made available by the researchers (or language teachers) who compile them for their own specific research or teaching purposes. One of the few welcome exceptions of a freely accessible specialized corpus that can be searched and browsed through an online interface is MICASE, the Michigan Corpus of Academic Spoken English (cf. Simpson et al. 2002). The MICASE online interface at http://quod.lib.umich.edu/m/micase/ allows users to specify their searches according to a range of criteria such as speaker native-speaker status (e.g. near-native speaker), academic discipline (e.g. chemical engineering), or speech event type (e.g. large lecture), and thus makes it possible to derive authentic speech samples or concordance materials from MICASE that are tailored to particular groups of EAP (English for academic purposes) learners. In the near future (by the end of 2009), MICASE online will be complemented by an academic written English corpus: MICUSP, the Michigan Corpus of Upper-level Student Papers (see project website at http://micusp.elicorpora.info/).

2.3.1 Applications of general corpora

Let us now look at a few examples of current direct and indirect pedagogical applications of general corpora. We have seen above in the discussion of COBUILD work done at the University of Birmingham that corpus evidence can greatly affect course design and the contents of teaching materials, and I would like to join Sinclair (2004c: 271) in stressing the need for evaluating existing pedagogical descriptions in the light of 'new evidence.'

Recent pedagogically oriented studies of the indirect type that take corpus findings seriously, and use language features that are known to cause problems to language learners as their starting point, include those by Barlow (1996) on reflexives, by Conrad (2004) on linking adverbials, by Grabowski and Mindt (1995) on irregular verbs, by Mindt (1997) on future time expressions, by Römer (2004a,b, 2005, 2006) on modal verbs, if-clauses, and progressives, and by Schlüter (2002) on the present perfect. All of these studies found considerable mismatches between naturally occurring English and the English presented in EFL (English as a foreign language) teaching materials (textbooks, grammars), and discuss the need for revised pedagogical language descriptions that take corpus findings into account and present a more adequate picture of language as it is actually used.

A case in point here are the misrepresentations of the preferred functions and contexts of the English progressive in German EFL teaching materials. Figure 2.2, taken from Römer's (2005) comparative study on the use of progressives in genuine spoken British English and in representations

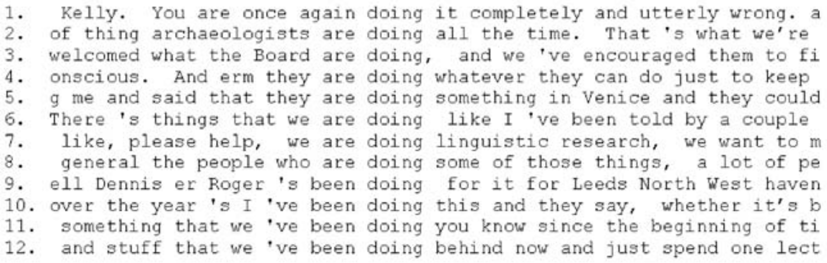


	repeated	non-repeated
▒ BNC spoken	35.05%	64.95%
▨ BoE spoken	38.64%	61.36%
□ Green Line New	9.12%	90.88%
■ English G 2000	9.87%	90.13%

Figure 2.2 Progressives and shares of repeated actions across 'real English' and 'school English' corpora (Römer 2005)

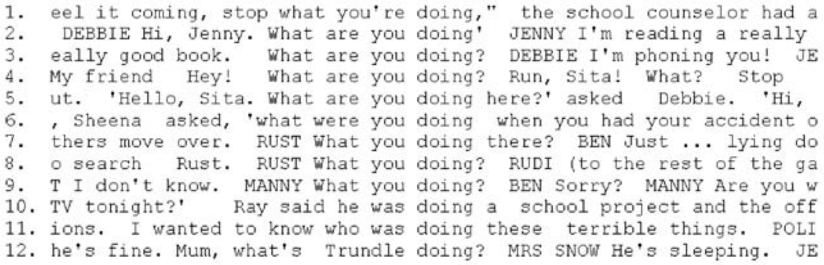
of spoken English in EFL textbooks, displays the shares of progressive verb forms expressing repeatedness in large data sets from four 'real English' and 'school English' corpora (the spoken parts of the BNC and the BoE, and from two small EFL textbook corpora, Green Line New and English G 2000). Examples of progressives (taken from the analysed BNC dataset) which refer to repeated actions or events are given in (1) and (2). As the two right-hand bars in the first bar cluster in Figure 2.2 illustrate, repeatedness is very rarely expressed by progressives in the textbook data, where progressive verb forms refer to single continuous events in more than 90 per cent of the cases, as exemplified in (3) and (4). This clear trend also becomes apparent in the concordance samples in Figures 2.3 and 2.4 (taken from BNC spoken and English G 2000 concordances of 'doing'). While there are a number of progressives that refer to repeated actions or events in the BNC spoken concordance sample (e.g. lines 1, 2, 6, 8), the dominant pattern in the English G 2000 sample in Figure 2.4 is 'What (are) you doing?' – a question that clearly points at an ongoing single event.

1. Do you find tha that that you're you're bringing traditions to people? (BNC spoken)
2. Yes. er it was sold out, you know, and everybody, and people had been ringing up thanking them and everything. (BNC spoken)
3. Robert: Well, no. I'm not listening to Radio 1. It's Radio Nottingham. I'm listening to my mum. (Green Line New)
4. Now, are we playing, or are we packing in? GLORIA Playing! (English G 2000)



```
1.   Kelly.  You are once again doing it completely and utterly wrong. a
2.  of thing archaeologists are doing all the time.  That 's what we're
3.  welcomed what the Board are doing,  and we 've encouraged them to fi
4.  onscious.  And erm they are doing whatever they can do just to keep
5.  g me and said that they are doing something in Venice and they could
6.  There 's things that we are doing  like I 've been told by a couple
7.   like, please help,  we are doing linguistic research,  we want to m
8.   general the people who are doing some of those things,  a lot of pe
9.  ell Dennis er Roger 's been doing  for it for Leeds North West haven
10. over the year 's I 've been doing this and they say,  whether it's b
11.  something that we 've been doing you know since the beginning of ti
12.  and stuff that we 've been doing behind now and just spend one lect
```

Figure 2.3 BNC spoken concordance sample of 'doing' progressives, showing a high share of repeatedness



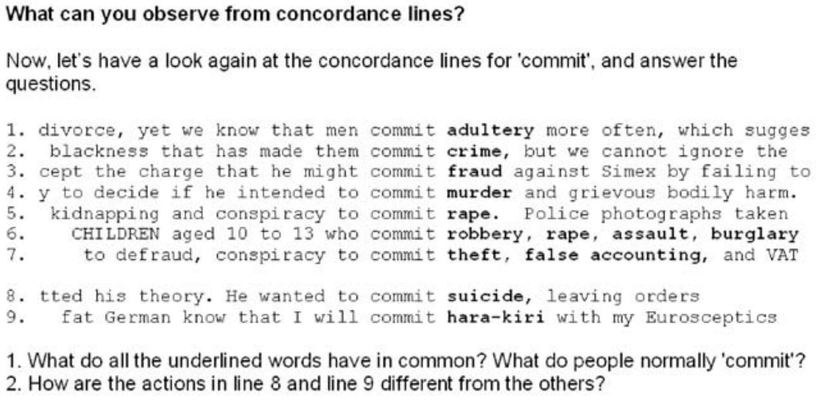
```
1.  eel it coming, stop what you're doing,"  the school counselor had a
2.   DEBBIE Hi, Jenny. What are you doing'  JENNY I'm reading a really
3.  eally good book.   What are you doing?  DEBBIE I'm phoning you!  JE
4.  My friend   Hey!   What are you doing?  Run, Sita!  What?   Stop
5.  ut.  'Hello, Sita. What are you doing here?' asked   Debbie.  'Hi,
6.  , Sheena  asked, 'what were you doing  when you had your accident o
7.  thers move over.  RUST What you doing there?  BEN Just ... lying do
8.  o search   Rust.  RUST What you doing?  RUDI (to the rest of the ga
9.  T I don't know.  MANNY What you doing?  BEN Sorry?  MANNY Are you w
10. TV tonight?'    Ray said he was doing a  school project and the off
11. ions.  I wanted to know who was doing these  terrible things.  POLI
12. he's fine. Mum, what's  Trundle doing?  MRS SNOW He's sleeping.  JE
```

Figure 2.4 English G 2000 concordance sample of 'doing' progressives, showing a low share of repeatedness

It seems that findings like these call for corpus-inspired adjustments in the language teaching syllabus, as far as selection and progression are concerned, and for revised lexical-grammatical descriptions for the hand of the learner. Such new descriptions could, for instance, address the described imbalance of functions and contexts in which progressives are used in real conversations and textbooks, use authentic instead of invented examples, and focus on frequent instead of rarely attested patterns (cf. Römer 2005, ch. 7).

Like indirect applications, direct uses of general corpora in pedagogical contexts have also come a long way since the 1980s. Not only do a growing number of language teachers (especially on advanced levels of instruction) work with corpora and concordances in the classroom, there are now also a couple of websites and textbooks available that provide either ready-made data-driven learning exercises or the tools for teachers to create simple DDL tasks themselves (e.g. gapfill or reorder). An interesting project is CorpusLAB, administered by Michael Barlow (see http://www.corpuslab.com). CorpusLAB is essentially a collection of websites (some of which are still under development) that lists ready-made DDL exercises for teachers and students and, what is more, contains exercise authoring capabilities for an ad-hoc creation of tasks on different genres. A number of ready-to-use DDL exercises based on large general corpora can also be found in Chris Tribble's and Glyn Jones's (1997) book *Concordances in the Classroom*, and several more are downloadable from websites created by Tim Johns, the 'father' of DDL himself, and by one of his former Ph.D. students, Passapong Sripicharn.[2] An example taken from Sripicharn's DDL pages that uses a filtered concordance (a list of selected concordance lines) from the Bank of English to highlight the use and collocates of the verb form 'commit' in written English is displayed in Figure 2.5 below (see also Sripicharn 2003).

Also worth mentioning is Tom Cobb's 'Compleat Lexical Tutor' website (see http://132.208.224.131/), a tool collection '[f]or data-driven language learning on the web,' which offers (among other things) a set of vocabulary quizzes and vocabulary lists linked to concordances that enable users to 'explore the nuances of form, meaning, and collocation' of words (quotes from website blurb). The site also provides cloze test builders that help teachers create cloze passages for words from specified frequency bands. Another promising development in this context is the publication of the first general corpora-based EFL textbooks, such as the volumes published in CUP's *Touchstone* series (e.g. McCarthy et al. 2005). According to CUP's promotional description for the series, *Touchstone* draws 'on research



What can you observe from concordance lines?

Now, let's have a look again at the concordance lines for 'commit', and answer the questions.

```
1. divorce, yet we know that men commit adultery more often, which sugges
2.  blackness that has made them commit crime, but we cannot ignore the
3. cept the charge that he might commit fraud against Simex by failing to
4. y to decide if he intended to commit murder and grievous bodily harm.
5.  kidnapping and conspiracy to commit rape.  Police photographs taken
6.    CHILDREN aged 10 to 13 who commit robbery, rape, assault, burglary
7.     to defraud, conspiracy to commit theft, false accounting, and VAT

8. tted his theory. He wanted to commit suicide, leaving orders
9.   fat German know that I will commit hara-kiri with my Eurosceptics
```

1. What do all the underlined words have in common? What do people normally 'commit'?
2. How are the actions in line 8 and line 9 different from the others?

Figure 2.5 DDL exercise on verb-noun patterns around 'commit' (courtesy of Passapong Sripicharn)

in the Cambridge International Corpus' and 'presents the vocabulary, grammar, and functions students need for effective conversations.'

Common to all these projects and publications is the general understanding that the creation of a data-rich learning environment and working with concordance materials in the classroom in an inductive fashion can have lots of positive effects on the language learning process. Central keywords related to the effects of DDL or 'corpus-aided discovery learning' (Bernardini 2002: 165) are learner motivation, serendipity, communicative competence, language awareness raising and learner autonomous learning. So, corpus work in pedagogical contexts is seen to have a number of advantages since, for the learner, '[c]orpora will clarify, give priorities, reduce exceptions and liberate the creative spirit' (Sinclair 1997: 38). This is of course not only valid for work with general but also for pedagogical applications of specialized corpora.

2.3.2 Applications of specialized corpora

Like large general corpora, smaller collections of more specialized texts or of language produced by a specific group of people can also have a strong direct or indirect impact on language teaching practice. As far as indirect applications are concerned, corpora that capture a particular LSP (language for special or specific purposes) can influence syllabus design for LSP courses. As Gavioli (2006: 23) points out, in ESP (English for specific purposes) 'working out basic items to be dealt with is a key teaching problem.' A keyword analysis based on a corpus that contains the specific text or discourse type in question (e.g. English business letters, medical research articles, or newspaper editorials) can help solve this problem and assist teachers in 'focus[ing] their efforts in terms of selection of language contents' (Pérez-Paredes 2003: 1).

An important issue for ESP teachers (who may or, what is more likely, may not be experts in the specific discourse they have to teach) is that they should give priority to teaching those words and expressions that their learners will need later on to be able to handle texts in their subject area. For instance, having access to a corpus of biology readings and lectures, to give just one example of a science English course described by Flowerdew (1993), can enable teachers to successfully address this issue and make informed decisions about item and text selection for their course. Another example of how corpora can impact general EAP teaching is Coxhead's (2000) Academic Word List (AWL). The AWL, based on a corpus of academic writing, contains those vocabulary items which are most relevant and useful to EAP learners. Simpson-Vlach and Ellis (in preparation) have taken Coxhead's idea from word to phrase level and devised an Academic Formulas List (AFL) which consists of word combinations that occur significantly more often in academic than in non-academic speech and writing. Both studies (AWL and AFL) take an indirect approach to using specialized corpora in language teaching and contribute to improving the teaching of English for academic purposes through informing syllabus design.

Recently, specialized corpora have also come to be used in EAP and ESP classroom concordancing, and materials derived directly from specialized corpora are regarded valuable tools by EAP/ESP instructors worldwide, as indicated by the large number of hits on the MICASE online and MICASE teaching materials websites from people in more than 130 different countries.[3] Resources like the MICASE instructional materials and MICASE kibbitzers (at http://micase.elicorpora.info/micase-kibbitzers) focus on some core items and communicative functions of academic English (e.g. hedging or complaining) and help students gain insights into the phraseology of academic speech.

Gavioli (2001) reports on classroom concordancing with small and specialized corpora and addresses the question 'How can learners analyse corpora without getting lost?' (Gavioli 2001: 109). She suggests that teachers reduce and classify the data and tell their students to look for recurrent features in concordance samples. In my experience this strategy works quite well, and students enjoy working with concordances once they know how to read them and what to focus on. In a different publication (Gavioli 2006), the author provides examples of DDL activities for students of economics, centring (among other things) around the specialist vocabulary of marketing research articles. Her exercises highlight important collocations (e.g. of the word 'market') and keywords in the discourse of marketing. Such exercises can help economics students familiarize themselves with the central vocabulary in their discipline, hence becoming better readers, and maybe also better producers of economics texts. I would, however, argue that the focus on ESP/EAP teaching should not only be on the specific lexis and typical communicative functions of the discourse type in question but that it also makes sense to work with concordances of general high-frequency words like 'price,' 'way' or 'cost' – words that learners already know but that may be used in different ways and take on specific meanings in specialized languages. I would also suggest a shift in focus from lists of frequent individual words or keywords to lists of frequent clusters or n-grams that provide insights into the phraseological profile of a certain genre. The extraction of 3- and 4-word clusters around the word 'way' from a corpus of academic research articles, for instance, results in a list of phraseological items such as ‘in this way’, ‘by way of’, ‘way in which’, ‘in the same way’, ‘in such a way’, ‘gave way to’ or ‘as a way of (V-ing)’ – items that may be very useful for students in their own academic writing.

2.4 Corpus Linguistics and Language Teaching: Future

Having provided a couple of examples of past and present pedagogical corpora use, I would now like to turn to the future of corpus linguistics and language teaching and consider what could or should be done next in applied corpus linguistic research and what items we should put on our agenda. The tasks I envisage are grouped under three topics and discussed in turn below: (i) focus on learner and teacher needs, (ii) foster indirect uses of corpora in language teaching and (iii) foster direct uses of corpora in language teaching.

While many corpus researchers (including myself) claim that corpus linguistics has an immense potential to help improve language pedagogy, I would argue that they do not always make sufficient efforts to reach practitioners with the ‘corpus mission’ and to find out about what teachers actually want and need. My suggestion would therefore be to focus our attention more on language teachers and their needs and see how we could support them in their work. A survey among 78 practising English language teachers that I carried out in 2005 and report on in Römer (2009) brought to light that a number of wishes and everyday problems of German EFL teachers could actually be addressed by applied corpus linguists. Among the things the teachers who participated in my survey called for were, e.g., better teaching materials, support in creating materials, and native speaker advice. One possible response to these wishes would be to introduce more teachers to corpus resources that are already freely available online. If teachers received a basic training in working with corpora and had access to computers with a good internet connection, they could design the required materials themselves whenever they needed them, e.g. a worksheet on the most frequent nouns in the academic journals subsection of the Corpus of Contemporary American English or an exercise around a concordance of the adjective ‘significant’ in all lectures included in MICASE (using the MICASE online interface, see Appendix A). They would also see that corpora, as large collections of native or expert speaker/writer output, can replace the ‘always available native speaker informant’ they asked for, and that questions like ‘What prepositions go with that verb?’

can easily be answered by looking at a right-sorted concordance of the verb in question.

Another task on our list should be to pay more attention to the needs of learners and consider which groups of learners may profit most from which type of materials. Related to this issue are questions centring around the learners' willingness and ability to deal with computer corpora, online search interfaces and concordance exercises prepared by their teachers. DDL may work well with the computer-savvy student who is ready to explore larger amounts of language data, but it may not be the best solution for the techno-phobic student who prefers a teacher-centred, controlled type of instruction. The needs of learners will probably also not only vary considerably by learner type but also by learner level, course type and learner objectives. Depending on whether we are dealing with intermediate or more advanced learners, for instance, our focus in designing corpus-derived materials may shift to more specialized vocabulary and its preferred patterns of usage. Similarly, participants in a business English class or international students of mechanical engineering will probably profit most from working with materials that are tailored to their specific needs and discourse in their field of study. That means that in making decisions on what to teach and how to teach it, it is important to consider the learner's language background and what discourse community she/he eventually wants to be a member of and be able to communicate with.

In terms of fostering indirect uses of corpora in language learning and teaching, more work probably has to be put into the creation of reliable corpus-based language descriptions for learners and teachers, especially descriptions of specialized discourses, such as academic English or business English. This implies that there is a need for more large specialized corpora that can be used as bases for creating dictionaries, usage guides and grammars tailored to the needs of different groups of learners. While corpora of American and British spoken academic English (MICASE, see above, and BASE, see Nesi and Thompson 2006) have been compiled and made available in the past few years, there is not yet a large corpus of academic writing in the public domain that could be exploited by applied corpus linguists and inform pedagogical materials.[4] A promising project in this context is the Aston Corpus Network (ACORN, see http://acorn.aston.ac.uk/) which aims at compiling a multimillion word corpus of academic English and exploiting it for language teaching purposes. Generally, I see more scope for research activities that are inspired by the needs of learners and teachers (as discussed above) and that take the learners' communication needs and common learning problems into account. Language points that tend to be particularly difficult for learners could e.g., be identified through comparative analyses of corpora capturing learner/novice and native-speaker/expert performance data, or through contrastive linguistic analyses based on parallel corpora of the learners' native and target languages. Further comparative studies of lexical-grammatical features in corpora and coursebooks (see section 3.1) could also provide valuable insights into mismatches between 'real language' and 'school language' that need to be remedied.

In terms of fostering direct uses of corpora in language learning and teaching, corpus researchers would do well to help create more DDL exercises and corpus-derived teaching materials in general. In the future, I would hope to see more publications (similar to Tribble and Jones 1997 or Barlow and Burdine 2006) that contain ready-made exercises based on authentic speech and writing from different text types and language varieties and focused on language items that are of central importance and/or troublesome for learners. Web-based projects like CorpusLAB are also likely to promote direct corpora use in language teaching and may help bring more corpus materials into the classroom. Another important step we need to accomplish if we want DDL to gain more ground is create a DDL-friendly environment that encourages learner and teacher involvement. Teachers and learners have to be provided with access to corpora that are available on the internet or to offline corpora and easy-to-use concordance packages. A popularization of corpora and their pedagogical use also requires some basic training in accessing corpora and in working with concordances or collocation lists. Such training is crucial because concordance output, at first glance, may seem hard to handle, and because 'a corpus is not a simple object, and it is just as easy to derive nonsensical conclusions from the evidence as insightful ones' (Sinclair 2004b: 2). For a number of teachers (and learners), the technology behind DDL may, however, be too difficult, even if they are given a basic training, or they simply may not have access to computers and the internet. In this case, DDL materials will have to come in a paper-based format, e.g. as photocopiable worksheets.

2.5 Conclusion

This chapter has looked at the relationship between corpus linguistics and language teaching and has sketched some past, present and possible future developments in the field. It has aimed to demonstrate the importance of the work of applied corpus linguists in helping to improve pedagogical practice and, by means of examples, illustrated the wide range of corpus applications in language teaching. It has also tried to make clear that, despite the progress that has been made over the past two or three decades, much still remains to be done in research and practice to help corpus linguistics fully ‘arrive’ in the classroom, and that general and specialized corpora could be even better exploited to positively affect the life of teachers and learners (see also Römer 2008).

In a questionnaire he sent out to language teachers, teacher educators, and linguistic researchers in spring 2008, Chris Tribble referred to the (direct) use of corpora in language teaching as ‘a minority sport.’ One purpose of this chapter was to describe the potential of pedagogical corpus applications and to provide ideas on what can be done to foster direct and indirect corpus use in language teaching and to bring corpora and corpus tools to a larger group of learners and teachers. I think that, if we take up some of the ideas mentioned above and if we are successful in improving communication among researchers, teachers and materials designers, we can get more people involved in DDL and related activities and thus perhaps make applied corpus linguistics more of a majority sport.

Acknowledgements

I would like to thank Mari Carmen Campoy and Begoña Bellés Fortuño for inviting me to present a version of this chapter at the first international conference on corpus-based approaches to ELT in Castellón, Spain, 6–8 November 2007. I am also grateful to Keith McDade and Stefanie Wulff for helpful comments on an earlier draft of this chapter.

Notes

[1] TaLC conferences take place every other year in different European countries. TaLC 1 through 8 were held in Lancaster (1994 and 1996), Oxford (1998), Graz (2000), Bertinoro (2002), Granada (2004), Paris (2006) and Lisbon (2008). TaLC 9 is scheduled to take place in Brno in 2010. Since 1999 conferences with a similar focus to TaLC have also been held in North America (in Ann Arbor, MI in 1999 and 2005, in Flagstaff, AZ in 2000 and 2006, in Boston, MA in 2001, in Indianapolis, IN in 2002, in Montclair, NJ in 2004, and in Provo, UT in 2008), organized by the American Association for Applied Corpus Linguistics (now the American Association for Corpus Linguistics, AACL).

[2] See http://www.eisu2.bham.ac.uk/johnstf/ddl_lib.htm and http://www.ajarnton.com/DDLunits/units.htm.

[3] There were more than 150,000 hits on the MICASE online pages in 2007 (http://quod.lib.umich.edu/m/micase/). For the MICASE project website, including links to online MICASE-based instruction materials (http://lw.lsa.umich.edu/eli/micase/index.htm, now at http://micase.elicorpora.info), we tracked 14,230 visits from 133 countries between September 2007 and April 2008.

[4] Researchers can, of course, use the academic English subsections of the BNC and BYU Corpus of Contemporary American English, but these collections may not be large enough and not cover a wide enough range of text types and academic disciplines.

References

Ädel, A. (2006), *Metadiscourse in L1 and L2 English.* Amsterdam: John Benjamins.

Aston, G. (ed.) (2001), *Learning with Corpora.* Bologna: CLUEB and Houston, TX: Athelstan.

Aston, G., Bernardini, S. and Stewart, D. (eds) (2004), *Corpora and Language Learners.* Amsterdam: John Benjamins.

Barlow, M. (1996) ‘Corpora for theory and practice’. *IJCL,* 1, (1), 1–37.

Barlow, M. and Burdine, S. 2006. *American Phrasal Verbs (CorpusLAB Series).* Houston, TX: Athelstan.

Bernardini, S. (2002), ‘Exploring new directions for discovery learning’, in Kettemann, B. and Marko, G. *Teaching and Learning by Doing Corpus Analysis. Proceedings of the fourth international conference on Teaching and Language Corpora, Graz 19–24 July, 2000.* Amsterdam: Rodopi, pp. 165–182.

Botley, S., Glass, J., McEnery, T. and Wilson, A. (eds) (1996), *Proceedings of Teaching and Language Corpora 1996.* Lancaster: University Centre for Computer Corpus Research on Language.

Braun, S., Kohn, K. and Mukherjee, J. (eds) (2006), *Corpus Technology and Language Pedagogy.* Frankfurt: Peter Lang.

Burnard, L. and McEnery, T. (eds) (2000), *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt: Peter Lang.

Campoy, M. C. and Luzón, M. J. (eds) (2007), *Spoken Corpora in Applied Linguistics.* Bern: Peter Lang.

Conrad, S. (2004), ‘Corpus linguistics, language variation, and language teaching’, in Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins, pp. 67–85.

Coxhead, A. (2000), ‘A new academic word list’. *TESOL Quarterly,* 34, (2), 213–238.

Davies, M. (2008), ‘The 360 million word BYU Corpus of American English (1990–2007)’. Paper presented at the American Association of Corpus Linguistics conference, 13–15 March 2008, Brigham Young University, Provo, UT, USA.

Flowerdew, J. (1993), ‘Concordancing as a tool in course design’. *System,* 21, (2), 231–244.

Gavioli, L. (2001), 'The learner as researcher: Corpus concordancing in the classroom', in Aston, G. *Learning with Corpora.* Bologna: CLUEB and Houston, TX: Athelstan, pp. 108–137.

—(2006), *Exploring Corpora for ESP Learning.* Amsterdam: John Benjamins.

Grabowski, E. and Mindt, D. (1995), 'A corpus-based learning list of irregular verbs in English'. *ICAME Journal*, 19, 5–22.

Granger, S., Hung, J. and Petch-Tyson, S. (eds) (2002), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Amsterdam: John Benjamins.

Hidalgo, E., Quereda, L. and Santana, J. (eds) (2007), *Corpora in the Foreign Language Classroom.* Amsterdam: Rodopi.

Johns, T. F. (1986), 'Microconcord: A language-learner's research tool'. *System*, 14, (2), 151–162.

— (1991), 'Should you be persuaded – Two samples of data-driven learning materials', in Johns, T. F. and King, P. (eds), *Classroom Concordancing* (*ELR Journal*, 4, 1–16.

—(1994), 'From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning', in Odlin, T. (ed.), *Perspectives on Pedagogical Grammar.* Cambridge: Cambridge University Press, pp. 27–45.

—(2002), 'Data-driven learning: The perpetual challenge', in Kettemann, B. and Marko, G. (eds), *Teaching and Learning by Doing Corpus Analysis. Proceedings of the fourth international conference on Teaching and Language Corpora, Graz 19–24 July, 2000.* Amsterdam: Rodopi, pp. 107–117.

Kettemann, B. and Marko, G. (eds) (2002), *Teaching and Learning by Doing Corpus Analysis. Proceedings of the fourth international conference on Teaching and Language Corpora, Graz 19–24 July, 2000.* Amsterdam: Rodopi.

—(2006), *Planing, Gluing and Painting Corpora. Inside the Applied Corpus Linguist's Workshop.* Frankfurt: Peter Lang.

McCarthy, M., McCarten, J. and Sandiford, H. (2005), *Touchstone Student's Book 1.* Cambridge: Cambridge University Press.

Mindt, D. (1981), 'Angewandte Linguistik und Grammatik für den Englischunterricht', in Kunsmann, P. and Kuhn, O. (eds), *Weltsprache Englisch in Forschung und Lehre: Festschrift für Kurt Wächtler.* Berlin: Schmidt, pp. 175–186.

—(1987), *Sprache – Grammatik – Unterrichtsgrammatik. Futurischer Zeitbezug im Englischen I.* Frankfurt: Diesterweg.

—(1995), *An Empirical Grammar of the English Verb. Modal Verbs.* Berlin: Cornelsen.

—(1997), 'Corpora and the teaching of English in Germany', in Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds), *Teaching and Language Corpora.* London: Longman, pp. 40–50.

—(2000), *An Empirical Grammar of the English Verb System.* Berlin: Cornelsen.

Mukherjee, J. (2004), 'Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany', in Connor, U. and Upton, T. A. (eds), *Applied Corpus Linguistics. A Multi-dimensional Perspective.* Amsterdam: Rodopi, 239–250.

Nesi, H. and Thompson, P. (2006), *The British Academic Spoken English Corpus Manual.* Warwick: The University of Warwick.

Pérez-Paredes, P. (2003), 'Small corpora as assisting tools in the teaching of English news language: A preliminary tokens-based examination of Michael Swan's Practical English Usage news language wordlist'. *ESP World* 6, 2. Available online at http://www.esp-world.info/articles_6/pascual.htm (accessed 30 April 2008).

Römer, U. (2004a), 'Comparing real and ideal language learner input. The use of an EFL textbook corpus in corpus linguistics and language teaching', in Aston, G., Bernardini, S. and Stewart, D. (eds), *Corpora and Language Learners.* Amsterdam: John Benjamins, pp. 151–168.

—(2004b), 'A corpus-driven approach to modal auxiliaries and their didactics', in Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins, pp. 185–199.

—(2005), *Progressives, Patterns, Pedagogy: A Corpus-Driven Approach to Progressive Forms, Functions, Contexts and Didactics.* Amsterdam: John Benjamins.

—(2006). 'Looking at *looking*: Functions and contexts of progressives in spoken English and "school" English', in Renouf, A. and Kehoe, A. (eds), *The Changing Face of Corpus Linguistics.* Amsterdam: Rodopi, pp. 231–242.

—(2008), '7. Corpora and language teaching', in Lüdeling, A. and Kytö, M. (eds), *Corpus Linguistics. An International Handbook.* Berlin: Mouton de Gruyter, pp. 112–130.

—(2009), 'Corpus research and practice: What help do teachers need and what can we offer?', in Aijmer, K. (ed.), *Corpora and Language Teaching.* Amsterdam: John Benjamins, pp. 83–98.

Schlüter, N. (2002), *Present Perfect. Eine korpuslinguistische Analyse des englischen Perfekts mit Vermittlungsvorschlägen für den Sprachunterricht.* Tübingen: Narr.

Scott, M. and Tribble, C. (2006), *Textual Patterns. Key Words and Corpus Analysis in Language Education.* Amsterdam: John Benjamins.

Simpson, R. C., Briggs, S. L., Ovens, J. and Swales, J. M. (2002), *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan.

Simpson-Vlach, R. C. and Ellis, N. C. (In preparation), 'An academic formulas list (AFL)'.

Sinclair, J. McH. (ed.) (1987), *Looking Up: An Account of the COBUILD Project in Lexical Computing.* London: Collins ELT.

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

—(1997), 'Corpus evidence in language description', in Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds), *Teaching and Language Corpora.* London: Longman, pp. 27–39.

—(ed.) (2004a), *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins.

—(2004b), 'Introduction', in Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins, 1–10.

—(2004c), 'New evidence, new priorities, new attitudes', in Sinclair, J. McH. (ed.), *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins, 271–299.

Sripicharn, P. (2003), 'Evaluating classroom concordancing: the use of concordance-based materials by a group of Thai students'. *Thammasat Review,* 8, (1), 203–236. Available online at http://www.tu.ac.th/resource/publish/interview/vol8.html (Retrieved 30 April 2008).

Tribble, C. and Jones, G. (1997), *Concordances in the Classroom.* Houston, TX: Athelstan.

West, M. (1953), *A General Service List of English Words.* London: Longman.

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

Willis, D. (1990), *The Lexical Syllabus. A New Approach to Language Teaching.* London: HarperCollins.

Willis, D. and Willis, J. (1989), *Collins COBUILD English Course.* London: HarperCollins.

Appendix: A Selection of Online-Searchable Corpora (All Accessed 30 April 2008)

BNC World online service for simple corpus searches: http://sara.natcorp.ox.ac.uk/lookup.html

Brigham Young University (BYU) Corpus of Contemporary American English (COCA): http://www.americancorpus.org/

Collins Concordance and Collocations Sampler: http://www.collins.co.uk/Corpus/CorpusSearch.aspxp

Michigan Corpus of Academic Spoken English (MICASE) http://quod.lib.umich.edu/m/micase/

Phrases in English (PIE) interface to the BNC: http://pie.usna.edu/

This page intentionally left blank

Part Two

Corpora and English for Specific Purposes

This page intentionally left blank

Chapter 3

Using Corpora to Teach Academic Writing: Challenges for the Direct Approach

Annelie Ädel
Stockholm University

3.1 Introduction

The digital revolution is bringing with it a range of new tools for teaching and learning. Some of these tools are already widely known and used, such as spellcheckers for written texts. Others are less widely used, such as corpora and concordancers. We do not know at this point in time exactly what impact this latter category will eventually have on teaching and learning, but it is interesting to consider the current challenges and how they could be met in the future.

The focus of this chapter is on the use of corpora and corpus tools in higher education, specifically to teach academic writing. The aim is to give an overview of the challenges involved in using corpus-based approaches in teaching writing in an EAP setting, specifically using a direct, hands-on approach. The motivation for doing so is not to discourage the teaching community from using corpora in the writing classroom, but rather to provide a starting point for evaluating the hurdles involved in order to make it possible to overcome some of them. Ways in which the challenges can be met, or at least alleviated, are suggested below.

Although the chapter exhibits a clear English-language bias, the general ideas presented are applicable to the teaching of any (academic) language. Furthermore, although the target student population involves both native and non-native speakers, the latter group perhaps more obviously stands to gain from access to corpus data. After all, this population lacks native-speaker intuitions and frequently (inappropriately) carries over writing patterns from the L1 into a rhetorically different L2, as shown by research in contrastive rhetoric; see e.g. Connor (1996).

One thing which will not be discussed here is why we might want to use corpora in teaching in the first place. Others have already presented very good reasons (e.g. Gavioli and Aston 2001; Sinclair 2004; Yoon and Hirvela 2004; O'Sullivan and Chambers 2006). Furthermore, I will not be able to offer any empirical support in favour of corpus-based approaches to teaching and learning. The potential gains of corpus-informed teaching and learning have generally been much touted in the corpus research camp, despite the fact that, as Granger (2004: 136) puts it, 'the number of concrete corpus-informed achievements is not proportional to the number of publications advocating the use of corpora to inform pedagogical practice.' Although the efficiency of such approaches is rarely tested, there are some studies specifically on writing applications, such as Creswell (2007), Gaskell and Cobb (2004), Chambers and O'Sullivan (2004), O'Sullivan and Chambers (2006), Yoon and Hirvela (2004) and Henry (2007).

We can make a simple distinction between two different approaches to corpus-informed teaching: (i) using corpus-informed materials, versus (ii) using corpora in the classroom. This distinction is a basic one made by many, for instance Hunston (2002: 137) and Römer (2006: 124–126 and this volume). I will adopt Römer's labels for this distinction and refer to the former as the 'indirect approach' and the latter as the 'direct approach'. In the indirect approach, the student is given access to corpus-informed teaching materials, which might have been produced by the teacher, or which could be in the form of a reference tool, such as corpus-based dictionaries and grammars. In the direct approach, by contrast, the student is given hands-on access to corpora; this may be done in more guided or more unguided ways. The guided way is sometimes referred to as 'directed learning', while a typical example of the unguided way is 'Data-Driven Learning', a method developed by Tim Johns (1991). At this extreme, students are said to act as language detectives or researchers. I will focus here on the direct, hands-on corpus approach (whether guided or unguided), specifically as applied to the teaching of writing. The majority of EAP practitioners and corpus linguists agree that the *indirect* approach is desirable and useful (albeit to varying degrees), but the *direct* approach is clearly both more controversial and less explored, which makes it a very interesting topic for discussion.

If we look specifically at the context of writing instruction, we will see that very little attention has been paid to the potential of corpora. Although it is easy to imagine the benefits of corpus consultation for students' writing skills, such consultation seems not to be widely practised. If we discount classroom sessions on vocabulary and collocation in general or adverbial connectors in particular (see e.g. Creswell 2007), there is not much to report on with respect to writing – neither in the literature nor on the internet.

As an example, in the 16-year history of the *Journal of Second Language Writing*, by March 2008 only two articles had ever been concerned with the use of corpora. Also, judging from titles and abstracts from TaLC (Teaching and Language Corpora) conferences since 1994, only a very small number of papers about writing have been presented. The very few published examples of hands-on use of corpora in writing instruction (e.g. Lee and Swales 2006) tend to take the form of one-time pedagogical experiments. Taken together, these data points show that corpus-based courses on academic writing are still at an exploratory stage – and furthermore, that at this point in time, it takes a corpus linguist to offer a corpus-based writing class.

The question that arises is why there are so few examples of teaching academic writing using corpora, especially above the level of vocabulary and collocation. In section 4, a tentative answer is presented in the form of a list of seven challenges to the teaching of writing in the corpus classroom. However, before considering these challenges, in section 2, I will set the scene by briefly outlining the traditional topics taught in writing. This is followed by an example, in section 3, of how corpus tools have been used in the EAP writing classroom.

3.2 Topics in Writing

Table 3.1 lists some of the topics traditionally dealt with in the teaching of academic writing.

Table 3.1 Traditional topics in academic writing

Effective persuasion (higher-level rhetorical strategies)
Audience analysis/awareness
Using sources and citation (how to use and refer to other sources, including avoiding plagiarism)
Text types (e.g. argumentative, expository)
Genres (e.g. research papers, business letters)
Coherence and cohesion (e.g. development of ideas, structuring of information, paragraph writing, use of connectives)
Organization (e.g. types of information to put in various sections; strategies for planning the text such as mind-mapping)
Level of formality/style
Rhetorical actions (e.g. paraphrasing, concluding, introducing topic)
Drafting, proofreading and revising
Vocabulary
Lexicogrammar and phraseology
Formatting (e.g. how to write a bibliography following *The Chicago Manual of Style*)

The table is organized from general to specific, where the top of the list represents higher-level skills and the bottom lower-level skills. Topics at the bottom of the table, such as lexicogrammar, collocation and academic vocabulary are important features which have been those most explored in the corpus classroom. This is quite understandable, since they are easier to study in a corpus-based paradigm than the higher-level features.

Writing involves many more topics than lexicogrammar and vocabulary, however, many of which cannot be taught using direct corpus-linguistic approaches, at least not considering the current state of the art. There are possibilities, though, of covering more of these topics in the corpus classroom, and of doing so in a more systematic way.

The above overview of typical topics taught in academic writing is meant to serve as a reminder that the concerns of composition are often complex, abstract and focused on extended stretches of language. It is an open question to what degree corpus-based tools may be fruitfully applied to these different topics in the corpus classroom.

3.3 A Hands-On Example

To illustrate one way in which corpus tools have been used in the EAP classroom, let me give an example of hands-on use of corpora in writing instruction from the University of Michigan's English Language Institute. In a small-scale experiment in 2006 (repeated a few times with different groups) I offered corpus consultation aiming to improve beginner students' writing skills. The pre-planned three-hour session involved guided corpus work in a relatively controlled environment.

The students were beginners in two ways: not only were they novice academic writers, but this was also their first exposure to corpus tools. They were first-year undergraduates at the end of their first year and they were non-native speakers of English, primarily from an Asian background, with relatively high English language proficiency levels. They were taking an introductory writing course. The corpus instructor (myself) had been informed about recent class activities by the regular writing instructor. The target writing of this particular group was referred to as 'general academic writing'.

The group first received a quick introduction to the concordancer. We started with what they already knew, and then contrasted tools such as dictionaries and grammars to the type of information one can get from a corpus (cf. Gavioli 2005). The focus of the class was on rhetorical functions and phraseology in academic writing. The specific research questions of the class were: What do academic writers say when they . . .? (a) give an example, (b) refer to other texts or researchers, (c) introduce the topic, (d) start their Conclusion section.

Suggestions for linguistic features to locate such actions were continuously elicited from students, although the instructor had done the searches beforehand and knew where to go for faster conclusions and for interesting qualitative analysis. The concordancer used was WordSmith (Scott 2004), and the specific functions used were word frequency, sorting to find patterns, collocational patterns and tracking distribution in texts. The corpus used was MICUSP, consisting of highly proficient, A-grade student writing (see section 3.1).

Based on the research questions, some of the discoveries we made were the following: Concerning (a), we noticed differences in use between 'for example' (tends to be sentence-initial), 'e.g.' (tends to precede a brief list, usually within parentheses) and 'such as' (tends to be preceded by general nouns, e.g. 'characteristics', 'factors', 'variables'). Concerning (b), we collected a list of useful reporting verbs, many of which were observed to be used in the passive. Concerning (c), we learned that writers rarely refer to the 'topic' explicitly, but rather tend to introduce the topic in subtler ways. By a simple search for 'introduction', we looked at Introduction sections and found how prevalent the importance or urgency factor was to justify research and to attract the attention of the reader. We located useful phraseology having to do with quantity and spread (e.g. involving 'countless', 'many', 'several', 'large', 'widely') and phraseology used to boost the topic (e.g. involving 'critical', 'not trivial', 'fundamental', 'central', 'important', 'key', 'greatest'). Concerning (d), finally, the students found phraseology for mentioning remaining problems or issues particularly useful (especially involving 'further', 'future', and 'need/ed').

The outcome of the experiment was mixed, although primarily positive. Among the pros can be mentioned that the analyses worked well in a controlled setting; that a large number of the students were quite enthusiastic about inductive learning and about the use of the computer (note the novelty factor); that the corpus session was nicely supported by instruction from the regular writing instructors (e.g. there was a follow-up lesson taking the indirect approach, using offline data from Conclusions). It can be added that note-taking on handouts was encouraged during the session and that, after class, the students were sent slides which summarized the answers we found to the research questions.

Among the cons, on the other hand, can be mentioned two specific drawbacks: first, the fact that it was a one-time event where students were given future access neither to the corpus (which was under development) nor to the concordancer and, second, the fact that the searches were restricted to surface forms of rhetorical functions, which naturally only works when explicit signals are in fact given. We were working at the bottom of Table 3.1, referring primarily to phraseology and rhetorical actions. Only to a small extent did we manage to touch on higher-level persuasive strategies: in finding the importance/urgency framing in Introductions and the remaining problems framing in Conclusions.

For insights into a more large-scale experiment also involving the direct corpus approach to writing instruction, see Lee and Swales (2006). In this case, the target group consisted of non-native-speaking doctoral students, also at the University of Michigan. The target writing involved dissertations and academic papers in specific sub-fields. Since this experiment involved highly advanced students and lasted an entire term, it was possible also to explore unguided corpus work. Like the hands-on example reported here, this was a one-time pedagogical experiment, devised by corpus linguists, not by writing instructors. Note that neither of the experiments referred to here were institutionalized; 'exploratory' seems indicative of the state of the art as a whole.

3.4 Corpora and Writing: Challenges

The extent to which we will be able to apply corpus methods in treating a greater range of topics depends on our ability to meet certain challenges. In the following section, I discuss seven different challenges involved in using corpus-based approaches in teaching writing in an EAP setting. The discussion will move from the general to the specific, with the more general points presenting problems that hold for any kind of corpus-based teaching, not just writing.

3.4.1 Lack of corpus availability

The first challenge is the lack of corpus availability. If one considers the scarcity of available corpora, it becomes evident that it is not easy to teach corpus-based academic writing. Although some corpora of academic writing exist, they tend not be generally available. Furthermore, when corpora of academic writing *are* available, they often consist of text fragments rather than full texts (e.g. the academic writing part of the British National Corpus), which makes it impossible to study many patterns in writing. Anyone interested in corpus-informed writing will have to agree with Sinclair's (1995: 27) call from more than a decade ago that whole texts deserve a much stronger position in corpus design. Since then, there has indeed been a shift in interest from large and general to small and specialized corpora (see e.g. Ghadessy et al. 2001), leading to a decreased use of text extracts, although this trend has not resulted in greater availability to any great degree.

However, there are recent developments that paint a more hopeful picture, especially with respect to student writing. In addition to the university-level student essays of the International Corpus of Learner English (e.g. Granger 1998), some new student-oriented corpus resources are becoming available. One is the Michigan Corpus of Upper-level Student Papers (MICUSP; see Ädel and Garretson 2006), which is under compilation at the University of Michigan's English Language Institute; another is the British Academic Written English Corpus (BAWE; see Nesi et al. 2004), being compiled by a group of universities in the United Kingdom.[1]

While unpublished student writing tends to be difficult to access, published academic writing tends to be relatively easy to obtain, especially research articles deposited in an electronic format on the internet. However, despite the easy access, the possibilities of using such material in a (generally available) corpus are limited by the fact that it is necessary to obtain permission from publishers to avoid breach of copyright laws. Indeed, copyright law is one of the greatest obstacles to the compilation of freely available corpora (Ädel 2007), even though practitioners tend to ignore copyright issues (see, e.g. Gavioli 2005). One potential approach to this problem would be for linguists and teachers to create a lobby group with the aim of achieving a universal, 'fair use' solution for using published work for purposes of research and teaching.

3.4.2 The corpus as a maze

Assuming that one has access to a corpus of academic writing, the second challenge to be addressed is what we can call the 'corpus as a maze' problem, referring to the difficulty of knowing what to look for in a corpus. A corpus will often seem easy to get lost in, especially to the uninitiated, and a great deal of linguistic and rhetorical knowledge is required in order to be able to use a corpus in interesting ways. As Stubbs (2005: 21) (in the context of corpus stylistics) puts it: 'Pure induction will never get you from empirical observations to interesting generalizations. You have to know where to look for interesting things.' How can a learner, or even a teacher, know in advance which corpus searches and hits will be worth pursuing and analysing?

While it is important to remember that the direct approach involves both the excitement of having a large body of language at one's fingertips, and the potential of the oft-cited *serendipity* of the corpus experience (see e.g. Bernardini 2000), all of the above strongly suggests that the direct approach needs to be used with great care. One way out of the maze is through teacher-guided settings and clearly defined tasks in connection with corpus searches. It could be added that corpus linguists face an urgent task in helping teachers and students through the maze, for example by creating smarter tools which provide better guidance.

3.4.3 Drowning in data

Assuming that one knows what to look for, the third challenge to be addressed is the potential risk of 'drowning in data'. A search which returns a large amount of data – perhaps hundreds, or even thousands, of hits – risks simply overwhelming the student.

This is a challenge which could be met relatively easily by building random sampling techniques into search tools (some concordancers already have this feature). This would enable students to access linguistic data in manageable quantities and in a statistically sound way. If this technical solution is not available, the teacher may work around this problem by assigning different samples to different students, thereby increasing the total number of analysed examples.

3.4.4 Interpretation

The fourth challenge constitutes the difficulty of interpreting corpus data. It has been said that corpus evidence is 'essentially indirect, which means that it cannot be taken at face value but must go through a process of interpretation' (Sinclair 2004: 7). Many different factors and conditions (from grammatical ones to sociolinguistic ones) influence language use, so trying to figure out why a linguistic feature is used the way it is can place considerable demands on the corpus user.

Not only students can find it difficult to deal with complex linguistic data; corpus linguists occasionally debate over the interpretation of a particular piece of corpus evidence. This difficulty is simply part of the analysis of linguistic data, so it could in fact be an advantage for (advanced) students to see that data can be interpreted in different ways and that knowledge is contentious, as suggested by Mauranen (2004). Although a teaching situation necessarily involves simplification, classroom teaching would benefit from an increased acceptance of the contentiousness of knowledge and greater awareness of the complexity of language.

The difficulty of interpretation could be avoided to some degree if corpus users were to familiarize themselves with the corpus by actually reading some of the corpus texts, perhaps even annotating as they went along. It is easier, as Aston (2002: 11) points out, to interpret concordances or numerical data 'if you know exactly what texts a corpus consists of, since this allows a greater degree of top-down processing'. This challenge is related to what Rissanen (1989) calls 'the philologist's dilemma', i.e. the concern that the use of corpus methods may supplant in-depth knowledge of the corpus texts themselves. This point is specifically worth making in the context of writing, where higher-level and contextualized knowledge is at a premium.

Another way in which to meet this challenge is to use comparison as a basic method. This can be recommended especially with non-native-speaking students, who tend to have poor register awareness. When considering a data point, it is often hard for this student population to know whether it is *a lot* or *a little; surprising* or *expected; typical* or *atypical.* This possibility already exists, but we need more sophisticated corpus tools with simple user interfaces that students can use to compare vocabulary, collocations and annotated functional features such as hedges across corpora of different genres. This would raise students' awareness of typical distinctions between genres. The same population would also benefit from, for example, comparing corpora of native-speaker and non-native-speaker writing, which could reveal patterns inappropriately carried over from the L1 to the L2 by the non-native writers.

3.4.5 Evaluation

Assuming that one has found and interpreted a pattern, the next challenge consists in evaluating that pattern. In the context of teaching writing, corpora tend to be used with the purpose of presenting models or patterns for students to emulate. It would be absurd to argue that a student should emulate every single usage that occurs in a corpus of target writing; the question is how a student can tell which patterns to adopt, or which patterns contribute to efficient writing. Take, e.g., the case of writer visibility in research papers: the fact that 'I' occurs at all may lead a student to think that it can be used in all kinds of discursive contexts and academic disciplines, which is certainly not the case.

A good writing instructor is both a descriptive linguist and a language policy maker. In using corpora, we study how people actually write and thus naturally take a descriptive approach. However, writing instructors also have a responsibility to students to give recommendations and good arguments for adopting one pattern and not another. Hunston (2002: 177, emphasis added) states that '[d]istinguishing between *what is said* and *what is accepted as standard* may need the assistance of a teacher or a grammar book', or even a manual on writing. The difficulty of evaluation can be reduced by teacher-guided sessions and the use of complementary materials.

3.4.6 Decontextualized data

Another potentially serious challenge in the use of corpus data in EAP settings is the fact that the corpus offers largely decontextualized data. This is a criticism directed at corpus linguistics in general by Widdowson (1998), but it could be argued that it is most severe in the case of corpus-based instruction on writing in particular. The criticism is that the social context is largely absent – or at least hidden – from the concordance line. Social context can be said to involve elements such as communicative context, typical writer-reader roles, cultural values and intertextual knowledge, all of which are crucial in the development of appropriate writing skills.

While the lack of context is a drawback to corpus-linguistic methods, it should not cause us to abandon corpus-informed teaching, even in the context of writing. The corpus and the computer should not be required to give the whole picture; e.g., they can not take on the main responsibility for giving the communicative purpose of a text, or for socializing students into the academic world or a specific academic field. Since 'a corpus is not going to offer all the resources learners and teachers require' (Tribble 2002: 145), corpus-based teaching works best as a *complement*. Especially in EAP writing instruction, it is important to ensure contextualization by bringing in other methods and giving students other perspectives on the target genres.

3.4.7 Focus on surface form

The final, and most serious, challenge to be addressed here is the inevitable focus on surface forms in corpus work. We are basically restricted to studying features that our corpus tools can find, which means that we run the risk of focusing exclusively on the word and the phrase level when using computer-assisted methods. The challenge lies in connecting surface forms (which are easy to search for by computer) to meaning (which tends to require human analysis) – whether lexical, collocational, pragmatic or discursive. Related concerns have been voiced by Swales (2002), who argues that the computer-based orientation of corpus studies leads to atomized, bottom-up investigations of language use (see also Flowerdew 2005).

In a writing context, we are often interested in exploring specific functions of language, which do not stand in a one-to-one relation to formal realizations. For example, compare the retrieval of (a) *modal verbs* and (b) *hedges* in a corpus search. While a modal verb represents a specific lexicogrammatical *form*, a hedge represents a linguistic *function* that can be realized through several different surface forms.

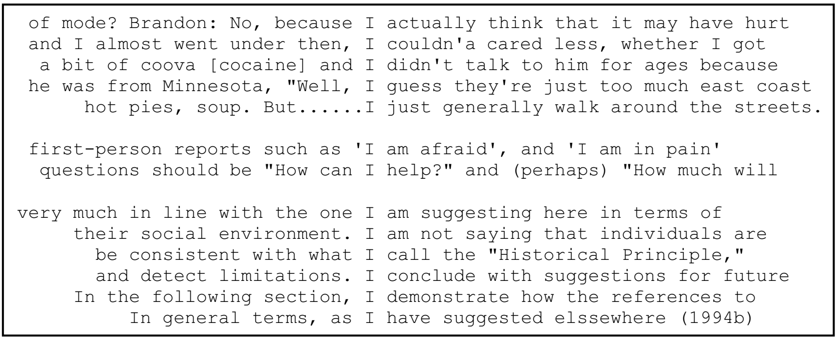
In order to find instances of (a) in a corpus, we can simply list all existing modal verb forms using a grammar for reference, and then search the corpus. It would help if the corpus were tagged for part of speech, otherwise cases of homonymy would have to be checked (e.g. the verb ‘can’ versus the noun ‘can’). In order to find instances of (b) in a corpus, we would first have to try to compile a list of all the possible linguistic forms that could function as hedges. Such a list would not be found in a grammar, let alone in a dictionary. Furthermore, different people’s lists might be quite different. If we assume that this initial hurdle can be overcome, though, all the instances in the corpus would have to be retrieved (e.g. modals ‘may’ and ‘could’), and then every single one would have to be checked in order to exclude those examples that do not function as hedges, such as *We may now turn to the following aspect of the problem . . .* or *We could not detect any statistically significant difference . . .* (examples from Salager-Meyer 2001). Although hedging, or the degree of (un)certainty toward the propositions expressed by a writer, is a feature of great interest in writing instruction, retrieving actual instances in the classroom can be prohibitively time-consuming.

Another example, through which we can illustrate some simple solutions to this problem, is attribution, or references to other sources. How does one locate examples of attribution in the corpus classroom? Three ways that have been suggested for finding simple indicators of, or proxies for, attribution are the following: (1) Tribble (2002: 143–144) reports that asking learners to ‘look at where and how parentheses are used is an excellent way of beginning an investigation of citation practices – especially in that once the parenthesised citations have been identified, it is then easy to follow up how (and with which verbs, in which structures) the proper nouns which occur in such lists are used in the text'. (2) Provided that the texts involved are not too quantitative in character, very simple search strings such as *200** and *19*** would retrieve common citation years.[2] (3) Proper names can also be retrieved, based on the names listed in the bibliographies of the corpus texts (cf. Ädel and Garretson 2006). There is only so much time we can ask our students to put into interpretation and analysis, so there is an urgent need to be inventive in retrieving relevant examples.

The examples above illustrate the fundamental issue at stake in the seventh challenge: the form-function split in human language. The challenge for corpus work is to find mappings between functional categories (such as politeness, evaluation or metadiscourse), which are very important in writing, and surface forms. Corpus linguists need to consider this split more thoroughly in order to make progress in corpus-based analysis of text and discourse.

One way in which we can make progress is through the use of annotation, or mark-up. The annotation of corpus text to allow for searches above the word level has been suggested in corpus stylistics (see Wynne 2005), and it also represents a promising avenue for creating corpora that are more useful in the classroom. A corpus annotated for features such as hedging and attribution, exemplified above, would be highly useful in the EAP classroom.

One relatively simple example of annotating a written corpus is marking all quoted material (typically explicitly marked by quotation marks, or set off in block quotes) as distinct from the running text. It is a serious restriction in present-day written corpora that there is no automatic way of making a distinction between the current writer's text on the one hand and quoted text on the other. This presents an unnecessary obstacle to many studies of writing, specifically those concerning the choices that writers make in their texts. In the case of students researching corpus frequencies, for example, it means having to spend time looking at the co-text of every single occurrence in order to determine whether it is a case of the current writer's text or quoted material. To give an example, when investigating writer visibility in research papers, one may end up with thousands of hits for 'I', particularly in papers from the social sciences, which tend to include text from informants, questionnaires, etc. Figure 3.1 shows a concordance sample of 'I' from a collection of research articles in the humanities, where the occurrences have been grouped into three categories.



```
   of mode? Brandon: No, because I actually think that it may have hurt
   and I almost went under then, I couldn'a cared less, whether I got
    a bit of coova [cocaine] and I didn't talk to him for ages because
   he was from Minnesota, "Well, I guess they're just too much east coast
        hot pies, soup. But......I just generally walk around the streets.

   first-person reports such as 'I am afraid', and 'I am in pain'
    questions should be "How can I help?" and (perhaps) "How much will

 very much in line with the one I am suggesting here in terms of
       their social environment. I am not saying that individuals are
         be consistent with what I call the "Historical Principle,"
         and detect limitations. I conclude with suggestions for future
       In the following section, I demonstrate how the references to
            In general terms, as I have suggested elssewhere (1994b)
```

FIGURE 3.1 Concordance sample of 'I' from research articles in the humanities

In the first group, what you see is not an academic writer, but an informant or interviewee, speaking. While the second group consists of metalinguistic quotes, the third group is the only relevant one for anyone researching academic writer visibility. This illustrates how a relatively simple annotation of quoted material would enable more efficient corpus searches.

3.4.8 Summary of Challenges

The seven challenges are listed in Table 3.2, together with a number of potential ways of meeting these challenges.

As I hope to have shown in this discussion, there are many pitfalls and a great deal of work remains to be done. With greater awareness of the difficulties involved, however, it should be possible to make faster progress. After all, the underlying assumption is that corpus-based inductive methods have a great deal to offer to both students and teachers of academic writing.

3.5 Conclusion

The impact of the digital revolution on teaching practices is an exciting but slow-moving process. While the use of corpus tools in researching writing and discourse became somewhat common in the early 2000s (e.g. Ädel and Reppen 2008), such tools can at best be described as marginal in the teaching of writing. This chapter has shown that there are several challenges that need to be addressed before we can cover a larger number

Table 3.2 Summary of challenges

Challenges	Suggested measures
❶ Availability of corpora	Greater efforts among corpus linguists are needed to make corpora generally available. Copyright issues need to be resolved, e.g. by forming a lobby group of linguists and educators.
❷ Corpus as a maze	Teacher-guided settings, clearly defined tasks and smarter tools would all contribute to making it easier to find ways out of the maze.
❸ Drowning in data	Random sampling needs to be built into search tools (this would also result in improved representativeness). A less desirable option is to have different students work on different parts of the corpus.
❹ Interpretation	Students need to be somewhat familiar with the texts included in the corpus. Comparison as a basic method makes it easier for students to gauge the typicality etc. of linguistic features found in a corpus.
❺ Evaluation	Reference materials can be used in conjunction with corpus data. Teachers need to take their advisory role seriously while at the same time embracing the descriptive corpus approach.
❻ Decontextualized data	Contextualization of the corpus data is needed, e.g. by giving students other perspectives on the target writing. Corpus-based instruction is one of many methods to be used in teaching writing.
❼ Focus on surface form	The most serious challenge. Annotation of functional features is a possible solution, although it is time-consuming and often subjective. The possibility of finding simple proxies exists in some cases.

of the traditional topics in academic writing using the direct corpus approach. It is impossible to know at this stage how far corpus-based writing instruction can be taken, but it is clear that so far we have explored few of the possibilities.

It seems appropriate to conclude this chapter by returning to the topic of annotation, especially the addition of pedagogically useful functional interpretations to texts, as one way of making progress. Annotation is needed in order to move beyond the surface level of text, which is often necessary in the context of the teaching of writing. Researchers into writing and discourse are beginning to explore the annotation of rhetorical or discoursal features, for example rhetorical moves in specific genres (e.g. job applications; see Connor et al. 2002). There also seems to be a growing pedagogical awareness among these researchers. For example, Flowerdew (1998: 549) suggests that annotating texts for generic move structures 'would have wide pedagogical applications'.

There are, of course, plenty of challenges involved in order to achieve successful corpus annotation. The primary among these is the need for solid taxonomies in the areas of discourse and rhetoric. These taxonomies need to procure high rates of inter-rater agreement while not avoiding categories where there is less agreement. The actual units we work with also need to be more explicitly defined (e.g. where do they begin and end?). Another issue at stake here is a technical one: the need for more user-friendly annotation software that enables powerful searches that still appear simple.[3] On that note, we can conclude that the bulk of the work outlined above makes demands on corpus linguists rather than teachers themselves; these improvements have to begin in research before they can trickle down to teaching.

Notes

[1] Information on the internet about MICUSP is found at http://lw.lsa.umich.edu/eli/eli1/micusp/Index.htm and about BAWE at http://www2.warwick.ac.uk/fac/soc/al/research/projects/resources/bawe/ (retrieved May 2008).

[2] The asterisk represents a wildcard, i.e. it represents any single character.

[3] One example of a useful piece of annotation software is Dexter (Garretson 2006; www.dextercoder.org).

References

Ädel, Annelie & Randi Reppen (2008). 'The challenges of different settings: An overview'. In Ädel, Annelie & Randi Reppen (eds), *Corpora and discourse: The challenges of different settings*. Amsterdam/Philadelphia: John Benjamins. 1–6.

Ädel, Annelie (2007). Review of Gavioli, L. (2005) 'Exploring corpora for ESP learning' published in *Studies in Second Language Acquisition*, 29, (4), 623–624.

Ädel, Annelie & Gregory Garretson (2006). Citation practices across the disciplines: The case of proficient student writing. In Pérez-Llantada Auría, Carmen, Ramon Pló Alastrué & Peter Neumann (eds), *Academic and professional communication in the 21st century: genres, rhetoric and the construction of disciplinary knowledge. Proceedings of the 5th International AELFE Conference*. 271–280.

Aston, Guy (2002). 'The learner as corpus designer'. In Ketteman, Bernhard & Georg Marco (eds), *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi. 9–25.

Bernardini, Sylvia (2000). 'Systematising serendipity: Proposals for concordancing large corpora with language learners'. In Burnard, Lou & Tony McEnery (eds), *Rethinking language pedagogy from a corpus perspective: Papers from the third international conference on teaching and language corpora*. Frankfurt am Main: Peter Lang. 225–234.

Chambers, Angela & Ide O'Sullivan (2004). 'Corpus consultation and advanced learners' writing skills in French'. *ReCALL*, 16, (1), 158–172.

Connor, Ulla, Kristen Precht & Thomas Upton (2002). 'Business English: Learner data from Belgium, Finland and the US'. In Granger, Sylviane, Joseph Hung &

Stephanie Petch-Tyson (eds), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam/Philadelphia: John Benjamins. 175–194.

Connor, Ulla (1996). *Contrastive rhetoric: Cross-Cultural Aspects of Second-Language Writing*. Cambridge: Cambridge University Press.

Creswell, Andy (2007). 'Getting to "know" connectors? Evaluating data-driven learning in a writing skills course'. In Hidalgo, Encarnación, Luis Quereda & Juan Santana (eds), *Corpora in the foreign language classroom: Selected papers from the sixth international conference on Teaching and Language Corpora (TaLC 6)*. Amsterdam/New York: Rodopi. 267–287.

Flowerdew, Lynne (2005). 'An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies'. *English for Specific Purposes*, 24, 321–32.

Flowerdew, Lynne (1998). 'Corpus linguistic techniques applied to textlinguistics'. *System*, 26, 541–552.

Garretson, Gregory (2006). Dexter: Free tools for analyzing texts. In Pérez-Llantada Auría, Carmen, Ramon Pló Alastrué & Peter Neumann (eds), *Academic and professional communication in the 21st century: genres, rhetoric and the construction of disciplinary knowledge. Proceedings of the 5th International AELFE Conference.* 659–665.

Gaskell, Delian & Thomas Cobb (2004). Can learners use concordance feedback for writing errors? *System*, 32, (3), 301–319.

Gavioli, Laura (2005). *Exploring corpora for ESP learning*. Amsterdam/Philadelphia: John Benjamins.

Gavioli, Laura & Guy Aston (2001). Enriching reality: Language corpora in language pedagogy. *ELT Journal*, 55, (3), 238–246.

Ghadessy, Mohsen, Alex Henry & Robert L. Roseberry (2001). *Small Corpus Studies and ELT: Theory and Practice*. Amsterdam: John Benjamins.

Granger, Sylviane (2004). 'Learner corpus research: Current status and future prospects'. In Ulla Connor & Thomas Upton (eds), *Applied Corpus Linguistics: A multidimensional perspective*. Amsterdam: Rodopi. 123–146.

Granger, Sylviane (ed.) (1998). *Learner English on Computer*. London & New York: Addison Wesley Longman.

Henry, Alex (2007). 'Evaluating language learners' response to web-based, data-driven, genre teaching materials'. *English for Specific Purposes*, 26, 462–484.

Hunston, Susan (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Johns, Tim (1991). 'Should you be persuaded—two samples of data-driven learning materials'. *English Language Research Journal*, 4, 1–13.

Lee, David & John Swales (2006). 'A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora'. *Journal of English for Specific Purposes*, 25, 56–75.

Mauranen, Anna (2004). 'Speech corpora in the classroom'. In Aston, Guy, Sylvia Bernadini & Dominic Stewart (eds), *Corpora and Language Learners*. Amsterdam/Philadelphia: John Benjamins. 195–211.

Nesi, Hilary, Gerard Sharpling & Lisa Ganobcsik-Williams (2004). 'Student papers across the curriculum: Designing and developing a corpus of British student writing'. *Computers and Composition*, 21, (4), 401–503.

O'Sullivan, Ide, & Chambers, Angela (2006). 'Learners' Writing Skills in French: Corpus Consultation and Learner Evaluation'. *Journal of Second Language Writing*, 15, (1), 49–68.

Rissanen, Matti (1989). 'Three Problems Connected With the Use of Diachronic Corpora'. *ICAME Journal*, 13, 16–20.

Römer, Ute (2006). 'Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments'. *Zeitschrift für Anglistik und Amerikanistik*, 54, (2), 121–134.

Scott, Mike (2004). *Wordsmith Tools version 4.* Oxford: Oxford University Press.

Sinclair, John (ed.) (2004). *How to use corpora in language teaching.* Amsterdam/ Philadelphia: John Benjamins.

Sinclair, John (1995). 'Corpus Typology: A Framework for Classification'. In Melchers, Gunnel & Beatrice Warren (eds), *Studies in Anglistics.* Stockholm: Almquist & Wiksell International. 17–34.

Stubbs, Michael (2005). 'Conrad in the computer: Examples of quantitative stylistics methods'. *Language and Literature*, 14, (1), 5–24.

Swales, John (2002). 'Integrated and fragmented worlds: EAP materials and corpus linguistics'. In Flowerdew, John (ed.), *Academic Discourse.* Harlow: Longman. 150–164.

Tribble, Christopher (2002). 'Corpora and corpus analysis: New windows on academic writing'. In Flowerdew, John (ed.), *Academic discourse.* Harlow: Longman. 131–149.

Widdowson, Henry (1998). 'Context, community and authentic language'. *TESOL Quarterly*, 32, (4), 705–716.

Wynne, Martin (2005). 'Stylistics: Corpus approaches'. In Brown, Keith (ed.), *Encyclopaedia of Language and Linguistics, 2nd edition.* Oxford: Elsevier.

Yoon, Hyunsook & Alan Hirvela (2004). 'ESL student attitudes towards corpus use in L2 writing'. *Journal of Second Language Writing* 13: 257–283.

Chapter 4

'I sort of feel like, um, I want to, agree with that for the most part . . .': Reporting Intuitions and Ideas in Spoken Academic Discourse

Begoña Bellés-Fortuño and Mari Carmen Campoy-Cubillo
Universitat Jaume I, Spain

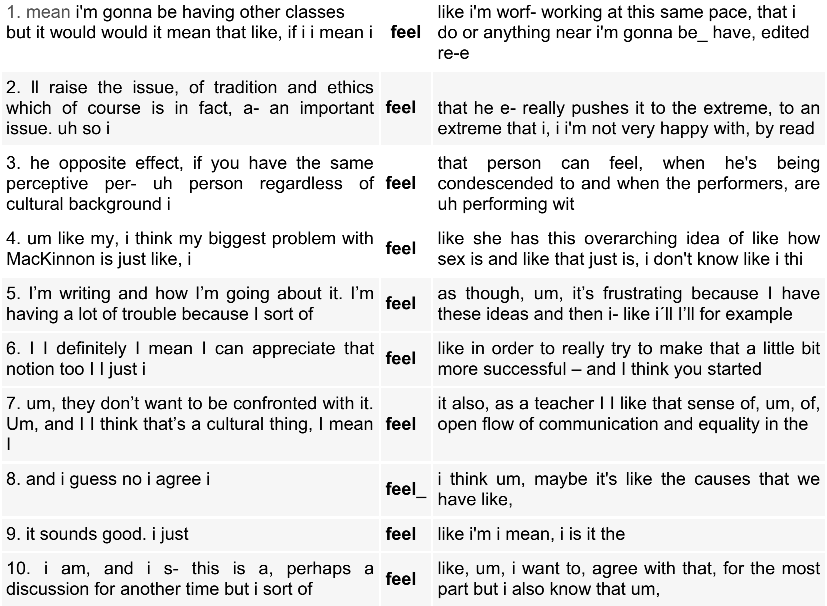
4.1 Introduction

The study of spoken academic discourse on discourse analysis and corpus linguistics has raised the interest of researchers in the past 20 years. One of the main focuses of corpus-based discourse analysis has been the study of the linguistic features that are common in specific discourse genres (Bellés-Fortuño 2004, 2007; Crawford 2004; Swales 2004; Biber 2006, Campoy and Luzón 2007). In this context, corpus-based analysis to examine interaction in professional contexts is one of the most productive areas of research. This chapter aims at analysing the use of *I feel* in the *Michigan Corpus of Academic Spoken English* (MICASE) as a means of expressing stance in natural discourse. Other aspects on the use of *I feel* that have been taken into consideration are its disciplinary and gender uses. Our analysis has revealed that *I feel* in the MICASE corpus has a considerable number of instances which are not evenly distributed among the corpus speech event types. Following the semantic classification of lexical verbs proposed by Biber et al. (1999), *feel* may be categorized as a mental verb. Among other verbs in this category, mental verbs include perception verbs (*see, taste*) and verbs reflecting attitudinal states (*feel, prefer*). Another useful approach to grammar on semantic principles is provided by Dixon (1991) who classifies the verb *feel* as a primary transitive verb within the THINKING semantic verb type. Biber et al.'s and Dixon's proposals are contrasted with the MICASE data in order to explore how the expression *I feel* is used in spoken academic English by different speakers and across different genres. Usage patterns emerging from the spoken academic discourse analysed in this chapter suggest ways to inform learners on how to use *I feel* and its variants in an academic context as different from other usage contexts.

4.2 Method

To carry out the analysis we have used the MICASE corpus (Simpson et al. 2002). MICASE (Michigan Corpus of Academic and Spoken English) is an online search engine containing a collection of transcripts of academic speech events recorded at the University of Michigan, in Ann Arbor (MI, USA). It has approximately 1.8 million words transcribed from a variety of speech events that goes from February 1998 up to now. Recently, the original MICASE search interface has been updated enabling users to get some additional information about descriptive statistics across speech events and speakers relating factors like academic roles and gender.

We used the MICASE corpus searching for *I feel* and some other collocational patterns in which the elements in the phrase *I feel* are non-contiguous (e.g. '*I sort of feel, you know, that . . .*'), analysing data by sorting results to the left (2L, 3L) of the word *feel*. Positions to the right of *feel* (2R, 3R, 4R) were also analysed for subordination patterns. A total of 38 non-contiguous uses of *I feel* were found (see for instance examples 9 and 10 in Figure 4.1 below). Within non-contiguous *I feel* there is a preference for position 2L as shown in example 10 in Figure 4.1, although non-contiguous *I feel* in 3L is also common.



1. mean i'm gonna be having other classes but it would would it mean that like, if i i mean i	**feel**	like i'm worf- working at this same pace, that i do or anything near i'm gonna be_ have, edited re-e
2. Il raise the issue, of tradition and ethics which of course is in fact, a- an important issue. uh so i	**feel**	that he e- really pushes it to the extreme, to an extreme that i, i i'm not very happy with, by read
3. he opposite effect, if you have the same perceptive per- uh person regardless of cultural background i	**feel**	that person can feel, when he's being condescended to and when the performers, are uh performing wit
4. um like my, i think my biggest problem with MacKinnon is just like, i	**feel**	like she has this overarching idea of like how sex is and like that just is, i don't know like i thi
5. I'm writing and how I'm going about it. I'm having a lot of trouble because I sort of	**feel**	as though, um, it's frustrating because I have these ideas and then i- like i´ll I'll for example
6. I I definitely I mean I can appreciate that notion too I I just i	**feel**	like in order to really try to make that a little bit more successful – and I think you started
7. um, they don't want to be confronted with it. Um, and I I think that's a cultural thing, I mean I	**feel**	it also, as a teacher I I like that sense of, um, of, open flow of communication and equality in the
8. and i guess no i agree i	**feel_**	i think um, maybe it's like the causes that we have like,
9. it sounds good. i just	**feel**	like i'm i mean, i is it the
10. i am, and i s- this is a, perhaps a discussion for another time but i sort of	**feel**	like, um, i want to, agree with that, for the most part but i also know that um,

Figure 4.1 Examples of contiguous and non contiguous I feel matches taken from the MICASE corpus

4.3 Analysis

A general search for the phrase *I feel* in MICASE (contiguous position) has given a total number of 147 matches in 59 transcripts. The word *feel* has also been searched in combination with *I* in non-contiguous instances where *I* precedes *feel* (e.g. I certainly feel . . .). Examples of some of these matches can be read in Figure 4.1.

There are some recurrent collocational patterns for contiguous *I feel.* The most frequent collocate to the right of *I feel* is *like.* The function of this particle in the bundle *I feel like_*(as seen in the examples 1, 4 and 6 in Figure 4.1) is different from its meaning 'wanting/desiring to do something' expressed in the following examples:

(11) an artist's musical journey, that remains, an inspiration for me. gee, **i feel like** clapping <*LAUGH*> that was really, great.

Although the bundle *(I) feel like* is usually taught in English language learning contexts by introducing examples similar to (11), in academic speech it is clear that the typical use of this bundle is that of combining the thinking meaning of 'feel' with the word 'like' to introduce a particular opinion or attitude, in fact only a couple of examples in the corpus have the meaning of 'wanting or desiring to do something'. This use of *like* in (1), (4) and (6) emphasizes the hedging nature of *I feel* in academic discourse. This may then be seen as a genre and context specific use of the bundle *I feel like.*

The other most frequent collocates are a group of verb modifiers such as *just, kinda, sort of, still* and the phrase *I mean* to indicate reformulation or hesitation. Some examples also illustrate this hesitation by reformulation of utterances where the speaker uses several thinking verbs, as may be seen in example (8) in Figure 4.1 above.

A recently updated engine on the MICASE interface provides statistical counts for speaker and speech event categories. These include information on the number of matches for a specific search in each speech event, the number of hits arranged according to academic division, interactivity rating, gender and academic role. In this chapter only statistics for contiguous use of *I feel* will be given in the tables that follow, though it should be noted that the preferences for non-contiguous *I feel* is the same as its contiguous counterpart, in terms of percentage of use in the different disciplines and when used by different speaker roles and genres. Likewise, there is a marked preference for highly interactive modes.

Table 4.1 Hits by academic division

Types	Hits
Biological and Health Sciences	13
Humanities	38
Not Applicable/Other	16
Physical Sciences and Engineering	19
Social Sciences and Education	61

Regarding the academic division factor, *Social Sciences and Education* is the academic division in which the highest number of contiguous *I feel* tokens is found (61), followed by *Humanities* with 38 hits and the rest of academic divisions with similar results: *Physical Sciences and Engineering, Biological and Health Sciences.* There are some academic divisions that do not apply to any of the categories established in the MICASE corpus and these only represent 16 hits out of the total number (see Table 4.1).

Engineering, Physical, Technical, Biological and Health Sciences are generally believed to show an objective and positive discourse, whereas the discourse in the *Humanities* and *Social Sciences and Education* (which belong to more closely related areas of research) tends to be more creative and subjective. Moreover, the academic division boundaries within MICASE are somehow blurred since if we take into consideration the concept of academic disciplines, we could amalgam *Humanities, Social Sciences and Education* under the broader branch of *Humanities.* Thus, as may be seen in Table 4.1, it is clear that there is a preference for the use of this hedging device (*I feel*) in the Humanities disciplines as opposed to the other sciences.

The semantic nature of the verb *feel* has been studied in Biber et al. (1999) who included *feel* within the category of verbs referring to mental/attitudinal states or activities (mental verbs), where stance on the part of the speaker is expressed. Thus, within the major grammatical devices conveying attitudinal stance, we find verbs followed by complement clauses, although Biber et al. do not include *feel* as a typical example in this category. Dixon (1991) provides a more complex verb type semantic classification explaining how the verb *feel* may belong to different semantic and grammatical classifications. Following Dixon's typology, *feel* may be included under CORPOREAL, ATTENTION, THINKING (KNOW sub-type) and SEEM. The analysis of *feel* in this chapter takes into consideration the use of *feel* as described in Dixon within the THINKING category. We want to point out a difference between THINK and KNOW sub-types within the THINKING verb types (see Table 4.2). While in THINK the speaker's perspective is internal and the speaker takes the role of Cogitator focusing on one person, thing, state or happening, in KNOW

Table 4.2 Dixon's (1991) THINKING Primary-B semantic verb type

Semantic Type	**Semantic Subtypes**
Primary-B verbs	
THINKING **(roles: Cogitator & Thought)**	THINK (the cogitator's mind just focusing on one person, thing, state or happening; the cogitator being aware of some fact, or body of information or method of doing something. e.g. think about / of / over, consider, imagine KNOW (referring to the Cogitator being aware of some fact, or body of information, or method of doing something) e.g. know, sense, learn, understand, teach.

'the Cogitator is aware of some fact, body of information or method of doing something' (Dixon 1991: 133). As stated in Dixon (1991: 133) *feel* could be included within the KNOW category with the meaning: 'to know something intuitively'.

When we want to report intuitions and ideas we may observe how the verb *feel* is frequently used as a KNOW verb type in examples where clusters expressing attitude such as *very acutely, a little bit more, more talented, really important, really thoughtful, is probably worth it,* are frequent. *Feel* as a KNOW verb sub-type also seems to be used when speakers are aware of some piece of information and try to convey it but their knowledge of that information is intuitive or the relationship with the other speaker is such that this knowledge cannot be stated as such (i.e., it has to be presented as an intuition so as not to be face-threatening).

A closer look at the use of the phrase *I feel* regarding Speech Event Title shows that the highest number of occurrences (20 tokens) are found in a Senior Thesis Study Group, followed by 7 and 6 matches respectively, in *Intro Biology Study Group* and in *Women in Science Conference Panel* and *Astronomy Peer Tutorial* with the same token result. Table 4.3 shows the total amount of *I feel* matches (147 matches) organized by Speech Event Title, number of matches and frequency of *I feel* occurrences per 1,000 words.

Taking a look at the frequency of words, we can see a close match between this frequency and the total number of *I feel* matches in the above-mentioned speech events (Table 4.4). Therefore, the high frequency word rate could be contributing in some way to the high number of matches in the *Senior Thesis Study Group* or the *Intro Biology Study Group.* It is worth mentioning that the highest number of matches for *I feel* occur in group speech events and not in proper lecture and seminar genres. Study groups are probably less fossilized academic speech events that allow a more subjective type of discourse, as opposed to lecture or seminar genres, which tend to be more monologic.

Table 4.3 *I feel* matches distributed by Speech Event Title and word count

Speech Event Title	Matches	Word count	Frequency / 10,000 words
Honors Advising	2	9,519	2.1
Academic Advising	3	28,160	1.06
Provost Public Lecture	1	9,116	1.09
Women's Studies Guest Lecture	1	10,370	0.96
Women in Science Conference Panel	6	20,099	2.98
Career Planning and Placement Workshop	1	14,842	0.67
Peking Opera Colloquium	1	12,152	0.82
Christianity and the Modern Family Colloquium	2	12,666	1.57
Social Psychology Dissertation Defence	1	12,280	0.81
Music Dissertation Defence	1	15,516	0.64
Artificial Intelligence Dissertation Defence	2	21,594	0.92
Philosophy Discussion Section	3	8,939	3.35
Intro to American Politics Discussion Section	4	7,751	5.16
Graduate Student Research Interview 2	1	2,963	3.37
Interview with Botanist	1	5,159	1.93
Biology of Fishes Field Lab	1	11,370	0.87
Cognitive Psychology Research Lab	3	14,839	2.02
Behaviour Theory Management Lecture	2	14,385	1.39
Fantasy in Literature Lecture	2	13,545	1.47
Separation Processes	1	5,438	1.83
Radiological Health Engineering Lecture	1	13,658	0.73
Visual Sources Lecture	1	12,526	0.79
Intro to Psychopathology Lecture	2	8,375	2.38
Rehabilitation Engineering and Technology	1	7,374	1.35
Intro to Groundwater Hydrology Lecture	1	14,151	0.7
Sex, Gender, and the Body Lecture	4	14,629	2.73
Ethics Issues in Journalism Lecture	2	16,291	1.22
Technical Communications Tutorial	4	4,178	9.57
Astronomy Peer Tutorial	6	21,798	2.75
Computer Science Office Hours	1	19,977	0.5
Anthropology of American Cities Office Hours	2	31,268	0.63
American Culture Advising	4	8,511	4.69
Linguistics Independent Study Advising	1	6,943	1.44
Economics Office Hours	1	14,050	0.71
Graduate Education Advising	3	9,224	3.25
Intro to Poetry Office Hours	1	12,317	0.81
Art History Office Hours	3	9,233	3.24
Graduate Philosophy Seminar	1	22,214	0.45
Graduate Buddhist Studies Seminar	1	26,075	0.38
Graduate Public Policy Seminar	1	25,414	0.39
First Year Philosophy Seminar	1	13,906	0.71
English Composition Seminar	3	21,442	1.39
Math Study Group	1	17,753	0.56
Objectivism Student Group	3	22,416	1.33
Biochemistry Study Group	1	17,530	0.57

(*Continued*)

Table 4.3 Continued

Speech Event Title	Matches	Word count	Frequency / 10,000 words
Organic Chemistry Study Group	1	18,124	0.55
Intro Biology Study Group	7	24,514	2.85
American Family Group Project Meeting	5	14,116	3.54
Senior Thesis Study Group	20	15,483	12.91
Second Language Acquisition Student Presentations	5	10,365	4.82
Bilingualism Student Presentations	1	15,956	0.62
Multicultural Issues in Education Student Presentations	2	13,078	1.52
Architecture Critiques	2	24,228	0.82
Brazilian Studies Student Presentations	5	12,905	3.87
Nursing Student Presentations	3	25,251	1.18
Black Media Student Presentations	3	10,540	2.84
Media Union Service Encounters	1	19,072	0.52
Science Learning Centre Service Encounters	1	8,613	1.16
Art Museum Tour	2	9,190	2.17

Table 4.4 Interactivity rating

Type	Hits
Highly interactive	87
Highly monologic	2
Mostly interactive	31
Mostly monologic	14
Mixed	13

Table 4.5 Gender tokens

Gender	Hits
Female	113
Male	34
Unknown	0

As to the type of speech event interactivity rate, i.e. interactive or monologic, the results for *I feel* show that this collocate is regularly used when the speech event is of a highly interactive or mostly interactive nature, as opposed to the low number of *I feel* matches in mostly monologic and highly monologic speech events (see Table 4.5).

Other aspects under study concern the speakers profile on subjects such as gender or academic status. Regarding the latter, MICASE distinguishes

Table 4.6 Academic role tokens

Role	**Hits**
Faculty	26
Graduate	14
Other	21
Undergraduate	86

among Faculty, graduate, undergraduate and other categories. When searching for *I feel* matches related to these factors, the results have given a higher number of hits for *I feel* among female speakers in comparison to male speakers. Regarding the academic role, the undergraduate category stands out. Undergraduate speakers tend to use *I feel* more than other academic roles, followed by Faculty with the highest number of tokens (see Table 4.6).

That female speakers in MICASE stand out in the use of *I feel* could be explained as the way females express attitudes and emotions and the degree of commitment towards what is being said. The specific use of personal pronouns determines the distance between speakers and listeners, the association of the verb *feel* with the pronoun *I* cannot be interpreted as a coincidence. In fact, some research on how personal pronouns behave in spoken academic discourse carried out (Morell 2001; Fortanet 2004b) show that the use of the first person singular pronoun (*I*) excludes the audience and creates a distance between speaker and hearer, as opposed to the most common meaning of the first person plural (*we*), in which speaker and hearer are usually included (Fortanet 2004b). That does not mean that males deliberately avoid the use of the collocate *I feel*, rather that they may use other hedging devices and collocates.

An observation should be made regarding gender and academic role in MICASE. Since the distribution of these factors is not balanced in the corpus, it could occasionally lead to some misinterpretations.

Similar mental verbs in phrases referring to first person speaker yielded the following results in MICASE: *I think*[1] (4,4028 matches, 149 transcripts; 2,246 female and 1,782 male); *I believe* (110 matches, 57 transcripts; 69 female, 41 male); *my belief* (1 instance, no examples found for *it is my belief*); *it seems to me* (56 matches, 27 transcripts; 22 female, 34 male); *it is my opinion* (20 matches, 14 transcripts; 10 female, 10 male). They all seem to prefer highly and mostly interactive speech and are most widely used in the *Humanities* and *Social Sciences and Education* academic divisions. It can be seen from these data that *I feel* shows a marked preference of use by female speakers (113 female, 34 male) over the other verbs and phrases, at least in the corpus used here.

4.4 Implications for EAP

Statistical information provided in MICASE reveals data which might be of interest when applying research results to the classroom. As stated in Luzón et al. (2007: 18):

> Previous corpus use for language teaching purposes relied to a great extent on concordances and the identification of collocates in order to focus on lexical and grammatical patterns. But pedagogically oriented corpora are currently experiencing a shift from focus on concordance to focus on text and text selection: the challenge now lies in the enhancement of the connection between the concordance, or bottom-up approach to corpus use, and the top-down, text-oriented approach. In the intersection, understanding and application of both approaches lays the future of corpora in learning environments.

In this sense, the data discussed here on the use of *I feel* can be explored in the classroom in connection to various speaker roles, gender differences, speech events, various academic disciplines and different interaction modes. MICASE online availability makes it a useful tool for the classroom and students may analyse aspects of the use of <u>I feel</u> by accessing concordances, keywords in context and full texts. Moreover, they may carry out their own restricted searches. For instance, they may want to see if there are any differences in use by graduates and undergraduates, or among the disciplines included in MICASE. This analysis may also be prompted by the teacher by posing questions such as:

1. Why do you think some disciplines use *I feel* frequently while others don't?
2. In which situations do faculty speakers use *I feel* in the Humanities?
3. Why do you think most people use *I feel* in highly interactive speech events and only a small number of speakers use this phrase in monologic events?
4. After analysing the concordances and answering the above questions, what can you say about the use of *I feel*?

4.5 Conclusion

We have seen how *I feel* behaves in spoken academic discourse in the MICASE corpus when we want to report intuitions and ideas. Gender differences were observed in the use of *I feel.* Female speakers make a higher use of *I feel* in their discourse utterances which may be interpreted as a female tendency towards an emotional and attitudinal academic discourse.

The co-text of *I feel* instances in MICASE has shown that it tends to co-occur with words and utterances with a high level of hedging. *I feel* rarely appears in isolation in spoken academic discourse and it usually co-occurs with other modifiers that form collocates such as *I sort of feel, I mean I feel, I feel like,* which deserve further and deeper analysis as multi-word units or lexical bundles. This is also evident in the occurrences of *I feel* in highly interactive speech events. Socio-cultural factors show the preference for the use of *I feel* as a stance marker in female discourse conducted in the *Humanities* branch and within highly interactive events.

This chapter has also shown how improvements in the design of multilayered corpora allow the combination of the study of grammatical and lexical features along with a wide range of socio-cultural features such as different kinds of speech events, social roles, gender and level of interactivity in a specific academic situation. Corpus-driven and corpus-based applications in the field of teaching and learning spoken speech should focus on the interaction of these different levels.

Recent advances in technology have thus been crucial in the compilation and improvement of large database corpora. These advances have made it possible to computerize socio-cultural features in order to assist corpus linguists in the understanding of discourse.

Notes

[1] The high number of I think occurrences also provides for a more extensive range of functions as stated in Fortanet (2004a), who classifies them into opinion, vagueness, uncertainty, politeness, approximator and hesitation.

References

Bellés-Fotuño, B. (2004), 'The spoken academic discourse of the Social Sciences: Discourse Markers within the university lecture'. Unpublished MA Thesis. Universitat Jaume I, Castelló, Spain.

—(2007), *Discourse Markers within the University Lecture Genre: A contrastive study between Spanish and North-American lectures.* Ph.D. dissertation. Publicacions Universitat Jaume I, Castelló: Spain. On-line archive: http://www.tesisenxarxa.net/TDX-0526108–134615/index.html

Biber, D. (2006), 'Stance in spoken and written university registers'. *Journal of English for Academic Purposes*, 5, 97–116.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999), *The Longman Grammar of Spoken and Written English.* London: Pearson Education.

Campoy, M. C. and Luzón, M. J., (eds), (2007), *Spoken Corpora in Applied Linguistics.* Linguistic Insights 51. Bern: Peter Lang.

Crawford, B. (2004), 'Interactive discourse structuring in L2 guest lectures: Some insights from a comparative corpus-based study'. *Journal of English for Academic Purposes*, 3, 39–54.

Dixon, R. M. W. (1991), *A New Approach to English Grammar on Semantic Principles.* Oxford: Clarendon Press.

Fortanet, I. (2004a), 'I think: opinion, uncertainty or politeness in academic spoken English?'. *RAEL: Revista Electrónica de Lingüística Aplicada*, 3, 63–84.

—(2004b), 'The use of "we" in university lectures: reference and function'. *English for Specific Purposes*, 23, 45–66.

Luzón, M. J., Campoy, M. C., Sánchez, M. M. and Salazar, P. (2007), 'Spoken corpora: New perspectives in oral language use and teaching', in Campoy, M. C. and Luzón, M. J. (eds), *Spoken Corpora in Applied Linguistics.* Linguistic Insights 51. Peter Lang: Bern, pp. 3–30.

Morell, T. (2001), 'The role of discourse markers and personal pronouns in lecture discourse', in Moreno, A. I. and Cowell, V. (eds.), *Perspectivas recientes sobre el discurso.* León: Universidad de León, pp. 202–206.

Simpson, R. C., Briggs, S. L., Ovens, J. and Swales, J. M. (2002), *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan.

Swales, J. (2004), *Research Genres. Exploration and Application.* Cambridge: Cambridge University Press.

Chapter 5

Hong Kong Engineering Corpus: Empowering Professionals-in-Training to Learn the Language of Their Profession

Winnie Cheng
The Hong Kong Polytechnic University, China

5.1 Introduction

To be successful communicators in the future workplace, ESP learners, i.e. professionals-in-training, need to enhance their language awareness through learning how to analyse language use, making informed decisions about their language choices, and being creative in their use of language. Bhatia (2004: 146) describes three components of 'professional expertise', namely disciplinary knowledge, professional practice and discursive competence. Discursive competence in professional contexts, which is relevant to the purpose of the present study, can operate at the levels of textual competence, generic competence and social competence (Bhatia 2004: 144). Textual competence refers to the ability for professionals to both master language (i.e. sounds, words, grammar, word meanings, discourse) and to use textual, contextual and pragmatic knowledge to construct and interpret texts. Generic competence is the ability for professionals to respond to recurrent and new communicative situations by producing, interpreting and using generic conventions in the accountancy disciplines to achieve professional goals. The last competence, social competence, according to Bhatia (2004), refers to the ability for professionals to use language more widely to participate effectively in a wide variety of social and institutional contexts to give expression to their social identity.

The study reported in this chapter is unique and innovative in two ways. First, it describes a new English ESP corpus, the Hong Kong Engineering Corpus,[1] and how it can be interrogated to empower professionals-in-training to learn the language in their profession. Second, it introduces a new computer-based methodology, 'concgramming' (Cheng et al. 2006;

Greaves and Warren 2007) and discusses how it can be used to facilitate the introduction of phraseology to ESP learners. The study argues that the phraseology and phraseological profile of the language contained in a text or a corpus that is specific to a discipline constitutes the 'aboutness' (Phillips 1983, 1989) of the discipline or profession. 'Phraseological profile' refers to the identification of the meaningful word co-occurrences in a text or a corpus (Cheng et al. 2006). The purpose of the study is, therefore, to empower professionals-in-training to learn the phraseology and phraseological profile characteristic of their profession, and most importantly, to learn the techniques in order to make inquires into the language, as a step towards achieving 'textual competence', and hence 'discursive competence', and eventually 'professional expertise' (Bhatia 2004: 146).

The chapter describes the strategies and methods to make corpus-based and web-based inquiries which will support the learning of the contextual use, and patterns of forms and meanings of language. The concgramming methodology is then described. For illustration purposes, the chapter will describe the aims and contents of Hong Kong Engineering Corpus, followed by presenting examples taken from the Hong Kong Engineering Corpus to investigate the key words and phraseology of distinctive engineering fields and text-types, in order to find out their local grammars and meanings in context. It also describes how ConcGram was used by students in corpus linguistics studies in a university in Hong Kong.

5.2 Phraseology

The chapter argues that phraseology is a major area of English language study that has yet been given sufficient attention. Two different approaches to classifying the product of the phraseological tendency in language (Sinclair and Renouf 1991; Sinclair 1996; Biber et al. 1999) are briefly outlined below. The first approach is expounded by Biber et al. (1999: 989–1025) and distinguishes among four different types of word associations: idioms, collocations, lexico-grammatical associations and lexical bundles. Examples of 'idioms' are *crop up, put up with, get away from* (Ibid.: 988); 'collocations' are associations between lexical words when the collocates co-occur more frequently than expected by chance (Ibid). An example of 'lexico-grammatical associations' (Ibid.: 989) is when verbs such as *think* and *know* are strongly associated with *to*-complement clauses (Ibid.: 989). Lastly, 'lexical bundles' (Ibid.: 989–1025) can be 'regarded as extended collocations: bundles of words that show a statistical tendency to co-occur' (Ibid.: 989), e.g. *do you want me to, the nature of, has not been* and *put it in.*

The second approach to classifying the product of the phraseological tendency is Sinclair's two types of extended unit of meaning, namely lexical core and grammatical core. A lexical core is an obligatory element in Sinclair's (1996) 'lexical item'. Lexical item is based on 'a lexical core and extended to incorporate grammatical as well as other lexical choices' (Ibid.: 105). The second type of extended unit of meaning is based on a grammatical core which constitutes Sinclair and Renouf's (1991) 'collocational framework' that is 'extended to incorporate lexical choices' (Ibid.: 105). In Sinclair's notion of extended unit of meaning, there is a core unit (either lexical or grammatical) around which are co-selected lexico-grammatical and semantic choices. The centrality of phraseology in language use propounded by Sinclair has led those working in the field of pattern grammar (e.g. Francis 1993; Hunston and Francis 2000; Hunston 2002) to argue that eventually corpus linguists will be able to describe all lexical items in relation to their syntactic preferences, and all grammatical structures with regard to their lexis and phraseology (Francis 1993: 155).

Sinclair (1996, 2004) describes the five categories of co-selection in accounting for the internal structure of a lexical item or a unit of meaning, namely the obligatory invariable core and semantic prosody, and the optional collocation, colligation and semantic preference. The lexical item 'reconciles the paradigmatic and syntagmatic dimensions of choice at each choice point' (Sinclair 2004: 141), using the five descriptive categories to describe both dimensions (Ibid.: 148). The first obligatory category is the 'core', which is 'invariable, and constitutes the evidence of the occurrence of the item as a whole' (Ibid.: 141). The other obligatory category is the 'semantic prosody'. It is 'the determiner of the meaning of the whole' lexical item, expresses the 'function' of the lexical item and shows 'how the rest of the item is to be interpreted functionally' (Ibid). A word may be said to have a particular semantic prosody if it can be shown to co-occur typically with other words that belong to a particular semantic set and display 'a subtle element of attitudinal, often pragmatic meaning' (Ibid.: 145). The three optional categories are collocation (Firth 1935, 1957), colligation (Firth 1935, 1957) and semantic preference. Collocation refers to 'the co-occurrence of words with no more than four intervening words' (Sinclair 2004: 141). Colligation is 'co-occurrence of grammatical choices', and semantic preference is 'the restriction of regular co-occurrence to items which share a semantic feature, e.g. about sport or suffering' (Ibid.: 141). The three optional categories 'realize co-ordinated secondary choices within the item, fine-tuning the meaning and giving semantic cohesion to the text as a whole' (Ibid).

5.3 The Hong Kong Engineering Corpus

The Hong Kong Engineering Corpus used in this study is comprised of 1,066,602 words and is the product of the first large-scale research project to collect corpus texts representative of the English language of the engineering sector in Hong Kong. It is compiled to enhance our understanding of real-world language use in the engineering industry in Hong Kong, in particular the study of the patterns of language use and their meanings. The Hong Kong Engineering Corpus consists primarily of the texts retrieved from the following sources: the Hong Kong Institution of Engineers (HKIE) website, the CD-ROM holding the *10th Anniversary HKIE Transactions* (a professional journal), newsletters from the *i-version Journal Hong Kong Engineer Online*, and other engineering-related websites, primarily Hong Kong government departments, academic institutions, and engineering companies.

The Hong Kong Institution of Engineers (HKIE) website provides a large variety of engineering texts and genres for all the engineers in Hong Kong. Corpus texts and genres were collected under the following headings: Civil Discipline, Conference Proceedings, Disciplinary Advisory Reports, Ethics in Practice, Guidance Notes, Mandatory Basic Safety Training, News, Press Release, Practical Guide, Newsletter Article, and Rules of Conduct. Table 5.1 summarizes the number of words for each genre downloaded from the HKIE website, and the total number of words is 335,080.

Table 5.1 Texts and genres in Hong Kong Engineering Corpus downloaded from HKIE website

HKIE website	Words
Civil Discipline	1,865
Conference Proceedings	131,265
Discipline Advisory Report	4,854
Ethics in Practice	11,816
Guidance Notes	1,301
Mandatory Basic Safety Training	465
News	3,677
Press Release	16,112
Practical Guide	14,538
Newsletter Article	147,814
Rules of Conduct	1,373
Total:	335,080

The second source of corpus texts is *HKIE Transactions* which is a quarterly periodical of the Hong Kong Institution of Engineers. *HKIE Transactions* publishes papers concerning engineering in all aspects, and are useful and interesting to practising engineers and academics. The four main genres are Abstracts, Research Papers, Technical Notes and Discussion Articles. The *10th Anniversary HKIE Transactions* CD-ROM contains an archive of ten years (1994–2003) of back issues of *HKIE Transactions,* amounting to 731,522 words.

Apart from the HKIE website and *HKIE Transactions*, newsletters from the *Hong Kong Engineer* i-version, October 2005 to May 2007, were included in the corpus. *Hong Kong Engineer* is the official monthly journal of HKIE and contains articles on engineering topics and news about HKIE activities. The fourth and last source of texts for the Hong Kong Engineering Corpus is the websites of Hong Kong government departments, institutions and private companies. To date, nine engineering-related government departments, six institutions, and fifty private engineering companies have been selected for data collection. At the time of writing, data collection for corpus compilation is in its final stages.

5.4 Concgramming the Hong Kong Engineering Corpus

'Concgramming' (Cheng et al. 2006; Greaves and Warren 2007) is a new computer-based methodology that has as its primary aim the automatic identification of the phraseological profile and hence the 'aboutness' (Phillips 1983, 1989) of a text or a corpus. With the use of the search engine ConcGram© (Greaves 2005), it is possible to extract recurrent 'concgrams', i.e. sets of between 2 and 5 co-occurring words, fully automatically, within a wide span (up to 12 words on either side of the origin), and which include all of a concgram's configurations irrespective of any constituent variation (e.g. AB and A*B) and positional variation (e.g. AB and BA) present. Cheng et al. (2006) suggest that identifying the concgrams in a corpus facilitates a fuller understanding of Sinclair's (2004) idiom principle, by revealing the word co-selections made by the speakers and writers represented in the corpus. Concgrams are, therefore, a useful starting point for quantifying the extent of phraseology in a corpus, and thus determining the phraseological profile of the language contained within it.

Applying the function of 'exclusion list' on the computer program ConcGram, the present study examines only the lexically-rich words in the 1,066,602-word Hong Kong Engineering Corpus. Membership of the lexically-rich words is determined by the majority of the co-occurring words, being what are traditionally termed 'lexical words'. In other words, in this study, words that are traditionally termed 'grammatical words' were excluded from the concgram searches.

5.5 Findings and Discussions

In the following, some findings of the study are presented and discussed. The purpose is also to demonstrate the methods that ESP learners can employ and the kinds of activities that they can carry out in order to understand the aboutness of the texts specific to their disciplines and professions. The author teaches an undergraduate subject ENGL303 Corpus-driven Language Learning, which was taken by students from different disciplines. In two computer laboratory seminars, the students worked on some tasks which were designed to search for the 'aboutness' of the HKEC, using ConcGram. The following describes the steps students followed in the seminars:

First, a basic search function found a total of 10,713 2-word concgrams in the HKEC. Table 5.2 shows the top twenty 2-word concgrams, i.e. phrases or word co-occurrences, with their frequencies of occurrence.

Table 5.2 Top twenty 2-word concgrams in Hong Kong Engineering Corpus

First word	Second word	Frequencies
1. Hong	Kong	2,457
2. as	such	577
3. as	well	458
4. concrete	strength	342
5. more	than	318
6. al	et	218
7. carried	out	218
8. Dr	Jr	202
9. concrete	high	189
10. high	strength	172
11. figure	shown	164
12. engineers	Hong	158
13. engineers	Kong	158
14. Civil	Engineering	156
15. Kong	University	153
16. Hong	University	152
17. power	supply	150
18. less	than	148
19. power	system	143
20. control	system	138

A glance at Table 5.2 suggests the aboutness of the texts contained in the corpus, which is one that is concerned with engineering. An illuminating example is 'Dr/Jr', occurring 202 times, which is the top 8th 2-word concgram. 'Jr.' is the short form for 'ingenieur' in French which means 'engineer *[noun-masculine]'. 'Jr.' is a conventional title, and in the case of an engineer with a doctorate degree, the conventional address form is 'Jr. Dr'.

The next step is to divide the twenty concgrams into discipline- and profession-specific (Table 5.3) and non-specific concgrams (Table 5.4). Table 5.3 shows that ten of the twenty (50 per cent) concgrams are discipline- and profession-specific. They are, in descending order of

Table 5.3 2-word concgrams from top twenty 2-word concgrams that are characteristic of engineering

First word	Second word	Frequencies
1. concrete	strength	342
2. Dr	Jr	202
3. concrete	high	189
4. high	strength	172
5. engineers	Hong	158
6. engineers	Kong	158
7. Civil	Engineering	156
8. power	supply	150
9. power	system	143
10. control	system	138

Table 5.4 2-word concgrams from top twenty 2-word concgrams that are not specific to engineering

First word	Second word	Frequencies
1. Hong	Kong	2,457
2. as	such	577
3. as	well	458
4. more	than	318
5. al	et	218
6. carried	out	218
7. figure	shown	164
8. Kong	University	153
9. Hong	University	152
10. less	than	148

frequencies, ‘concrete/strength’, ‘Dr/Jr.’, ‘concrete/high’, ‘high/strength’, ‘engineers/Hong’, ‘engineers/Kong’, ‘Civil/Engineering’, ‘power/supply’, ‘power/system’ and ‘control/system’.

Table 5.4 shows the other ten two-word concgrams that are not specific to the engineering discipline and profession; in other words, they are generic concgrams which are expected to be found in other disciplines and professions.

The students then further divided the ten general 2-word concgrams (Table 5.4) into two groups: those that are indicative of the nature of the corpus under study, and those that are not. Table 5.5 lists the first group which consists of ‘Hong/Kong’, ‘Hong’ or ‘Kong’, showing that the corpus is related to Hong Kong. There are three such examples among the ten.

The remaining six 2-word concgrams are interesting to analyse in order to find out the nature of these 2-word concgrams (Table 5.6).

Sinclair’s (2004: 141) semantic preference refers to ‘the restriction of regular co-occurrence to items which share a semantic feature, e.g. about sport or suffering’. The respective ‘indicative or predicted’ semantic preferences for the six concgrams are shown in Table 5.6 above. In other words,

Table 5.5 2-word concgrams from top twenty 2-word concgrams that indicate other aspects of the corpus

First word	Second word	Frequencies
1. Hong	Kong	2,457
2. Kong	University	153
3. Hong	University	152

Table 5.6 2-word concgrams from top twenty 2-word concgrams that are neither characteristic of the disciplinary language nor the nature of the textual corpus

First word	Second word	Frequencies	Indicative or predicted semantic preference
1. as	such	577	quality or characteristic intrinsic of a person or thing
2. as	well	458	additive relation
3. more	than	318	comparison of two items
4. al	et	218	academic referencing
5. carried	out	218	steps or procedures
6. figure	shown	164	quantitative reporting
7. less	than	148	comparison of two items

the concordances for the concgrams would need to be carefully studied to see whether the regular co-occurring words to the 2-word concgrams share a semantic feature, and then whether the shared semantic feature is what is predicted.

Regarding teaching input, in a lecture the following was explained to prepare students for a similar task. The following analysis taken from Cheng et al. (2006: 424–25) shows all of the concordance lines for the 2-word concgram 'high/low', with 'high' as the single origin. The corpus being examined is the 2-million-word Hong Kong Corpus of Spoken English. The purpose is to illustrate the procedure of analysing the phraseological profile of a concgram.

Example 1: 2-word concgram 'high/low'

1 proved my er hypothesis is correct it's um **high** proficiency students got better results than **low**

2 low profit and **low** earnings and and not very **high** stock prices and that makes people

3 authority versus the **low** authority versus the **high** structure versus the **low** structure

4 people that's a **low** authority society a **high** authority society is just the

5 try and buy when it's **low** and sell when it's **high** otherwise doesn't matter how

6 they're taking advantage of the **low** cost and **high** quality of production facilities in

7 a **low** individualism society or or or if you're **high** collectivist say Hofstede

8 okay can you (inaudible) (.) individualism **high** and **low** individualism if you are **low**

9 because it is too erm the whatever it's too **high** and too [**low** then erm this is

10 ays and bad the great changes the moments of **high** peaks and **low** troughs but I always

11 lity to be able to feel that at all that's a **high** EQ person a **low** EQ person's one

12 period of over fifty months of deflation **high** unemployment **low** levels of consumer

13 ividual er relationships are emphasized in a **high** individ- in a **low** individualism

14 students in order to know whether they are **high** proficiency or **low** proficiency

15 tside and there's imbalance because it's too **high** and this is too **low** (.) and that

16 del is applicable for any company dealing in **high** tech middle tech **low** tech and even

17 business model applies for any company doing **high** tech middle tech **low** tech and even

18 individualist versus the collectivist between **high** authority versus the **low** authority

19 it's the group will take care (.) whereas a **high** individualism society or a **low**

20 wer-fixed correction if the voltage ratio is **high** in the case of the boost or **low** in

21 he buckle converter is not preferred for its **high** peak current especially when **low**

22 (laugh)) B: but it needs to doesn't [come up **high** enough a1: [well it's too **low** it

All the instances of 'high/low' are non-contiguous, meaning there are always intervening words between 'high' and 'low'. The concgram 'high/low' has both constituency and positional variations. The positional variant 'low . . . high' in lines 2–7 has between 2 and 7 intervening words. In lines 1 and 8–22, the other positional variant 'high . . . low' has between 1 and 7 intervening words.

Three uses of ‘high/low’ are observed. First, speakers juxtapose points on a scale of ‘high <–> low’, and the item or attribute being juxtaposed include proficiency (line 1, 14), authority society (line 4), individualism (line 8, 13), EQ person (line 11), tech(nology) (line 16, 17), authority (line 18) and voltage ratio (line 20). Second, speakers present a relationship between two related items or qualities, e.g. ‘low earnings’ and ‘not very high stock prices’ (line 2), ‘low cost and high quality of production facilities’ (line 6), ‘low individualism society’ and ‘high collectivist’ (line 7), ‘high peaks’ and ‘low troughs’ (line 10), ‘high unemployment and low levels of consumer confidence’ (line 12), and ‘high peak current’ and ‘low voltage ration’ (line 21). The last usage is that the concgram ‘high/low’ extends across two speakers, and is an example of paraphrasing in which one speaker’s ‘doesn’t come up high enough’ is another speaker’s ‘it’s too low’ (line 22) (Cheng et al. 2006: 424–25).

5.6 Conclusion

This chapter has introduced a new computer-based methodology, ‘concgramming’ that was designed primarily for automatic identification of the phraseological profile of a text or a corpus. The methodology has been outlined, and examples from the Hong Kong Engineering Corpus have been discussed. The chapter argues that the linguistic information generated by ConcGram is not available in a dictionary or a thesaurus, but, with the right training provided to ESP learners, or the professionals-in-training, it is available through searching ESP corpora, such as the Hong Kong Engineering Corpus. The chapter has also described concgramming learning and teaching activities that highlight key elements in the understanding and production of phraseology in English, and which have been used in a corpus linguistics subject in a university in Hong Kong. The students found the concgram analysis task challenging but learned a lot about phraselogical patterns, particularly when the corpus texts are related to their disciplines.

The chapter strongly recommends employing the concgramming methodology and activities in ESP learning and teaching to raise the awareness of students and teachers regarding the role, nature and importance of the phraseological tendency in English language. They can also enhance both learners and teachers’ critical and creative thinking through the understanding, analysis, comparison and application of phraseology that is specific to ESP texts and genres.

Acknowledgement

The work described in this chapter was substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administration Region (Project No. G-YG10).

Note

[1] The Hong Kong Engineering Corpus now contains 9,224,384 words and is available by entering 'search a profession-specific corpus' on the website of the Research Centre for Professional Communication in English (RCPCE), Department of English, The Hong Kong Polytechnic University (http://www.engl.polyu.edu.hk/RCPCE/).

References

Bhatia, V. K. (2004), *Worlds of Written Discourse.* London and New York: Continuum.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Edward F. (1999), *Longman Grammar of Spoken and Written English.* Harlow: Longman.

Cheng, W., Greaves, C. and Warren, M. (2006), 'From n-gram to skipgram to Concgram'. *International Journal of Corpus Linguistics,* 11, (4), 411–433.

Firth, J. R. (1935), 'The technique of semantics', reprinted in J. R. Firth (1957), *Papers in Linguistics 1934–1951* (pp. 7–33). London: Oxford University Press.

—(1957), *Papers in Linguistics 1934–1951.* London: Oxford University Press.

Francis, G. (1993), 'A corpus-driven approach to grammar: Principles, methods and examples', in Baker, M., Francis, G. and Tognini-Bonelli, E. (eds), *Text and Technology: In Honour of John Sinclair.* John Benjamins, Amsterdam, pp. 137–156.

Greaves, C. (2005), *Introduction to ConcGram©.* Tuscan Word Centre International Workshop. Certosa di Pontignano, Tuscany, Italy, 25–29 June 2005.

Greaves, C. and Warren, M. (2007), 'Concgramming: A computer-driven approach to learning the phraseology of English'. *ReCALL Journal,* 17, (3), 287–306.

The Hong Kong Institution of Engineers Website, http://www.hkengineer.org.hk/program/home/index.php

The Hong Kong Institution of Engineers, *i-version Journal Hong Kong Engineer Online,* http://www.hkie.hk/html/Publication/pinkpageguide.htm

Hunston, S. (2002), *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Hunston, S. and Francis, G. (2000), *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English.* Amsterdam: John Benjamins.

Phillips, M. (1983), 'Lexical Macrostructure in Science Text'. (Unpublished Ph.D. thesis, Department of English, Faculty of Arts, University of Birmingham).

—(1989), *Lexical Structure of Text.* Discourse analysis monographs: 12. English Language Research: University of Birmingham.

Sinclair, John McH. (1996), 'The search for units of meaning'. *Textus*, 9, (1), 75–106.

—(2004), *Trust the Text.* London: Routledge.

Sinclair, J. M. and Renouf, A. (1991), 'Collocational frameworks in English'. Reprinted in Sinclair, J. McH. (ed.), *Lexis and Lexicography*, National University of Singapore: Unipress, pp. 55–71.

Chapter 6

Analysis of Organizing and Rhetorical Items in a Learner Corpus of Technical Writing

María José Luzón Marco
Universidad de Zaragoza, Spain

6.1 Introduction

The analysis of expert and learner corpora is key in determining the items of the language code that are the most worth teaching in ESP courses. The increasing number of corpus-based studies which analyse the features of English for Academic and Professional Purposes (e.g. Oakey 2002; Biber 2004; Biber et al. 2004; Groom 2005; Ward 2007) provide valuable information for the design of teaching materials. Some of these corpus-based studies (e.g. Biber et al. 1999; Oakey 2002; Biber 2004; Biber et al. 2004) have revealed the high frequency of some word combinations which fulfil organizing or rhetorical functions, such as introducing a topic, contrasting, explaining, summarizing, concluding, e.g. *the aim of this study, in addition, for example, it has been suggested, as a result of.* Therefore, in order to improve their textual competence and produce effective texts, students need to be trained to get familiar with the clause-combining and text-organizing items that tend to occur in professional and academic genres.

However, as several researchers (e.g. Flowerdew 1998, 2000; Aston 2000) have pointed out, studies based on the analysis of expert or native corpora are not enough to inform the design of EAP teaching materials and must be complemented with studies which analyse students' interlanguage. Aston (2000: 10) mentions two reasons why the analysis of corpora of native-speaker texts does not provide on its own an adequate basis for syllabus design. First, the findings from such analysis, 'provide no information as to the relative difficulty and learnability of particular features to be taught'; second, these findings do not provide help in identifying the productivity of particular features from the learner's perspective. He proposes using learner corpora in order to get relevant information on interlanguage development, 'which can be used to code and classify recurrent errors, along with over-uses and under-uses, with a view to identifying features which teaching should perhaps emphasise and to evaluating their difficulty'.

In this chapter we report the results of the analysis of a computerized corpus of technical English texts, written by Engineering students from the university of Zaragoza. The purpose was to find typical features of these students' interlanguage and identify unconventional elements that might not be accepted or might be misunderstood by professionals in their discipline. In this chapter we focus specifically on linguistic items used by students to organize the text, signal rhetorical relations between parts of the text or combine clauses within the text. The analysis will help us to reflect on how to improve the teaching of technical genres in the ESP classroom and to identify the aspects that should be focused on when teaching.

6.2 Pragmatic and Textual Errors in Learner Corpora of English for Academic or Professional Purposes

The analysis of learner corpora of academic writing has revealed recurrent errors and unconventional features in the writing of EFL students. A substantial part of this research has focused on errors occurring at the pragmatic and textual levels. In some cases there is a striking difference in the frequency of use of some items or word combinations in the discourse of students and experts. Students underuse some types of connectors, while they overuse other connectors that are frequent in their mother tongue and many typical EAP multi-word sequences (Granger and Tyson, 1996; Bolton et al., 2003; De Cock, 2003; Cortés, 2004; Carrió, 2006). For example, Granger and Tyson (1996), in their study on connector usage, found that French learners of English used more frequently those connectors which add to, exemplify or emphasize a point, but underuse connectors which are used for contrast or to take an argument logically forward. Carrió (2006) found that Spanish students tend to overuse additive, result and contrastive connectors, and underuse apposition connectors, which are considered repetitive and more informal in Spanish. Flowerdew (2000) examined both referential errors (faulty collocations and word forms) and pragmatic errors in a learner corpus of recommendation-based survey reports written by Hong Kong undergraduates and found that pragmatic markers (e.g. boosters and downtoners or hedging devices) were underused and sometimes semantically misused. The overuse of some items or word combination is sometimes a result of the influence of the mother tongue, as shown by Granger (1998) in her study of phrases which function as macro-organizers with a pragmatic function. She discovered that French learners massively overused the frame 'we/one/you can/cannot/may/could/might say that'.

Corpus-based research has also revealed how learners' use of some lexical items is different from experts' use (Flowerdew, 1998; DeCock, 2003). Flowerdew (1998) observed that Cantonese speakers always used the connectors *so, thus* and *then* as markers of local coherence, while in an expert corpus these connectors usually wrap up a previous stretch of text. Similarly, De Cock (2003) noted that French learners misuse some English sequences that have French deceptive cognates (e.g. *on the contrary* ≈ 'au contraire'). Cortés (2004) found that in the few cases in which students used certain lexical bundles,[1] their use did not correspond to the uses of such bundles by professional authors. L. Flowerdew (2003) also pointed to important differences between expert and novice writing in the use of *problem* in the problem-solution pattern. The results of Flowerdew's study show that students have trouble with a very frequent type of organising items in academic and professional writing: signalling nouns. Signalling nouns are defined by J. Flowerdew (2003: 329) as 'potentially any abstract noun, the meaning of which can only be made specific by reference to its context', e.g. *difficulty, process, reason, result.* The importance of these nouns in academic writing has led some researchers to analyse how they are used by students (e.g. Aktas, 2005; Flowerdew, 2006). Flowerdew (2006) examined a corpus of argumentative essays written by Cantonese L1 learners of English to produce a taxonomy of error types and frequency data of the different types.

6.3 Method

The learner corpus for this research was made up of 111 student assignments of approximately 800 words each, totalling 88,835 words. 71 texts were written by Computer Science students and 40 by Chemical Engineering students. Students wrote their assignments as response to a real-life task, where they were asked to produce either a report or a proposal for a specific audience. Although most learner corpora are compiled from the writing of high-intermediate level students, the texts in our corpus were produced both by students with a high-intermediate and a low-intermediate level, since our main purpose was to identify the errors that our students (no matter what level) make.

The first step of the analysis involved producing a wordlist, ordered by frequency. The list was analysed manually in order to identify items, including signalling nouns, which could potentially fulfil organizational or rhetorical functions, such as contrasting (*on the other hand, in contrast to*), explaining (*therefore, so*) exemplifying (*for example*), concluding (*in conclusion*), expressing purpose (*The purpose of this report is*), signalling transitions between parts of the text, combining clauses and linking paragraphs (*These results show*), and so on. We did not compare the frequency of the items in the learner corpora with their frequency in an expert corpus. Only in the cases where we considered that an item occurred with an extremely high frequency, did we check our impressions by searching the item in the written component of the BNC corpus. We will therefore make few statements concerning overuse or underuse and focus on the study of misuse or atypical, unconventional use.

In the second step we resorted to concordancing and qualitative analysis of the different organizing items to find information on positioning, patterns and function. As several researchers have pointed out (e.g. Flowerdew 2000), qualitative analysis is necessary in research concerned with pragmatic and textual aspects because concordances do not always show whether certain pragmatic and rhetorical devices are being used appropriately or not.

6.4 Results

The analysis of the corpus revealed that the students' misuse of English organizing items was sometimes due to a poor command of English and sometimes due to lack of genre awareness and of familiarity with generic conventions. The cases of misuse or atypical use of organizing items were classified into the following categories: (i) errors regarding the meaning or function of an item, (ii) use of lexical bundles or multi-word combinations which are atypical or inexistent in English, (iii) errors involving confusion in word class, (iv) position in the sentence different from expert's use, (v) use of words typical of informal discourse or oral discourse, (vi) errors regarding genre phraseology and (vii) errors involving signalling nouns.

(i) Errors regarding the meaning or function of an item

The learners had a tendency to use some items with a meaning or function which does not correspond to the uses of the item in native writing. This is the case, for instance, of *according to* or *on the other hand*. The Spanish prepositions *según* y *de acuerdo con* share several meanings or functions with *according to (*e.g. both the English and the Spanish prepositions are used to indicate the source from which the speaker got a piece of information, or to indicate that something is done according to a particular set of principles), but there is not an exact match in meaning between *according to* and the Spanish terms. However, due to transfer from the mother tongue, *según/ de acuerdo con* are just mapped to *according to,* and this item is used in the corpus with all the meanings of the Spanish terms, for instance to introduce the data which the speaker uses to draw a conclusion or to express an opinion. Learners often used *according to* with the meaning of *taking into account/ considering* (see examples 1a and 1b).

(1) a. *According to* the information that has been given above, we think that the Government should use booms and skimmers in order to remove the biggest part of the spillage.
b. *According to* the characteristics of the Prestige disaster zone, the best methods to clean the oil spilled are . . .

Another rhetorical item that was often misused in the learner corpus is *on the other hand.* This item is used in English to introduce 'a second argument that contrasts with what has just been said' (Chalker 1996: 33). However, the most frequent use in our corpus is not a contrastive, but an additive one, to signal that a further argument is being added, with a very similar meaning to *besides.*

(2) Although the industry of VR is growing faster and faster, there are technological problems that prevent a massive application in games (*list of the problems*). *On the other hand,* there are ethical and psychological barriers.

A group of items that were frequently misused by students are listing words. Although some of these items are only used to signal the different stages of a process, learners also used them to refer to the different parts of a text, with a metadiscursive function. In the corpus there is a lack of lexical bundles or words which refer to the following text (e.g. *in the following paragraph/section, below*), and the learners tended to use instead adverbs such as *next* or *subsequently.* For instance, *subsequently,* a time connector used to 'indicate that an event takes place at a later time than a previously mentioned event or time' (Chalker 1996: 17), was sometimes used in the learner corpus to refer to some other part of the text (e.g. 3).

(3) *Subsequently*, I'm going to make a description of the advantages and disadvantages of these renewable energies.

Another type of organizing items that posed problems were those used for stating a topic or referring to a different but related topic (e.g. *as for, as regards, with reference to*). Most students are not familiar with these items and tend to always use the expression *with respect to,* more similar in form to the Spanish *con respecto a* (e.g. 4). We analysed a random sample of 100 occurrences of *with respect to* in the written component of the *BNC*, and only in one case was it used to state a topic.

(4) With respect to the future of virtual reality, due to the current technological limits, only the sight and sound are the two senses that can be implemented and improved.

(ii) Use of lexical bundles or multi-word combinations which are atypical or inexistent in English

This is the case of multi-word items used to express cause in the structure "preposition+ this/that (pronoun)" (*due to this/ that*) (e.g. 5) which occur in the text more frequently than it would be expected. Both *due to this/that* (pronoun) and *because of this/that* (pronoun) occurred in the learner corpus eight times.

(5) . . . so a high-pressure tank is not necessary. *Due to this*, the tank can have an appropriate size for use in vehicles.

Because of this is relatively frequent in the written component of the *BNC* (470 occurrences), both with *this* as a pronoun or as an adjective, but there is no occurrence of *due to that* (pronoun) and only three occurrences of *due to this* (pronoun). In two of these occurrences the pattern is "be due to this" (e.g. 'The relatively sudden rise of xenophobic parties is largely due to this'). In the remaining 25 occurrences of *due to this* in the written component of the *BNC, this* is an adjective, followed by signalling nouns or by nominalizations (e.g. *due to this influence/condensation/process/factor*). Nominalizations and signalling nouns are powerful cohesive devices but they seem to be problematic for our students. The only signalling nouns/ nominalizations occurring in our corpus after *due to this/because of this* are *fact, change* and *aspect. Because of/due to this fact* and *because of this aspect* are used instead of common causal bundles, such a*s for that reason* or *this is the reason why* (e.g. 6).

(6) These simpler chemicals continue the damage of the environment although in a smaller amount and, *because of this aspect,* the substances that can be used in the area are regulated.

In some cases interlingual transfer leads learners to change the form of multi-word items. For instance, *to start* is used in our corpus instead of *to start with* as a listing connector, to emphasize the points that the speaker is making both to refer to time or reason. In Spanish no preposition would be placed after items such as *para comenzar, para empezar* and students seem to have transferred to English the structure of the lexical bundle in their mother tongue.

(7) Game devices deliver side effects. *To start,* extremely long playing can confuse some users, and make them believe reality and VR are the same thing

There are also several examples in the corpus of unusual expressions to indicate the different sections of the text. For instance, *I would bet for* instead of *I would recommend,* or *It is for all this that I recommend.*

(8) a. *I would bet for* study better the possibilities
b. *It's for all this that I recommend* the investigation for the immediate future of the Virtual Reality and . . .

Nesselhauf (2004: 141) points out that 'the unavailability of pragmatic chunks for the learners (. . .) appears to be the underlying reason for a number of deviant collocations which are used to structure the body of the essay', e.g. *Only have a look at; A first argument I want to name for this.* These deviant collocations are frequent in our learner corpus when learners use metadiscourse to present the topic/points that are going to be dealt with in the text.

(9) a. *The first point that should be clear* is that hydrogen is not an energy source but a storage method.
b. Some *solutions to this problem are going to be introduced* now.

c. First of all, current and future *consequences are going to be raised.* Afterwards *available solutions* which can be adopted *are going to be described.*
d. Global warming consequences *are our second point.*

(iii) Errors involving confusion in word class
Some prepositions, specially those expressing cause (*due to, because of*) and contrast/concession (*despite, in spite of, unlike*) are used as if they were conjunctions, as a result of transfer from the Spanish mother tongue, where many conjunctions take the form "preposition+ that" (e.g. *a pesar de que, debido a que*).

(10) a. We can only make tanks that are not practical for their use in vehicles, *due to* they are too heavy and too large
b. *In spite of* hydrogen has good density per weight, occupies too much volume at atmospheric pressure.

(iv) Position in the sentence different from expert's use
In her analysis of the expression of causality in expert and learner corpora, Flowerdew (1998) found marked differences in the positioning of cause-result connectors. In the expert corpus *therefore* occurred in sentence initial position in just 1 line out of 13. By contrast, in the learner corpus, *therefore* occurred in sentence initial position in 31 out of the 32 lines. Granger and Tyson (1996: 25) also pointed to the students' inexperience in 'manipulating connectors within the sentence structure'. This is also the case in our corpus. *Therefore* occurred in sentence initial position in 69 out of 74 lines. In three of the five remaining lines *therefore* occurred in the construction *and therefore*+verb with the subject ellipted.

(11) High-pressure tanks achieve 6,000 psi, and therefore must be periodically tested and inspected

Sentence initial position is preferred with virtually any type of connector, not just with causal connectors. For instance, *however* occurred in sentence initial position in 71 out of 72 lines. *Also* is another item very frequently used at the beginning of a sentence, when another reason is given, or when extra information is added (e.g. 12). Out of the 238 occurrences of *also* in the corpus, 50 were in sentence initial position (in 4 cases the initial item was *But also*). These figures contrast with those yielded by searches of these items in the *BNC*. In a random sample of 100 occurrences of *therefore, however* and *also*, we found that *therefore* occurred in sentence initial position only in

4 cases, *however* in 52 cases and there was only one occurrence of *also* in sentence initial position, but in the structure 'also+ gerund' ('Also benefiting from the fund (. . .) is Courtaulds Engineering pensioner Frank Burslem').

(12) Wind is becoming the most popular renewable source of energy but only few places have enough wind to work. *Also* windmills can fail due to the fact that some birds crash with the blades.

(v) Use of items typical of informal discourse or oral discourse

Although the task required the students to produce a written text (a report or a proposal) many of the items used to structure the text were more typical of oral academic genres (e.g. *so, by the way*). This may be due to the fact that students got explicit instruction on oral presentations and on how to organize and deliver them.

(13) Europeans have to be proud because Europe is currently the global leader in wind energy exploitation. *By the way,* I consider very important the evolution of Aragon in last years.

So is sometimes used in the corpus at the end of a text, with a conclusive function (*So, in my opinion; so, as a conclusion*). In oral discourse, *so* is sometimes used to 'suggest that what has been said is understood and therefore the next statement follows' (Chalker 1996: 122). There are a few occurrences of *so* with this function in the learner corpus.

(14) *So,* recapping the main points of this text up.

Learners frequently used items and lexical bundles typical of oral discourse when they wanted to introduce the next topic or point to be discussed in the text (e.g. 15). For instance, although *next* is used in oral discourse to mention what the speaker intends to discuss or do when he/she has finished discussing or doing something else (*Collins Cobuild English Dictionary*), *e.g. Next, I'll explain . . .*, it is not common to use *next* in academic writing for this purpose. Thus, example (15b) could be more correctly rephrased as "Some of those equipments will be described *below*".

(15) a. *To start,* I tell the history of the e-books (The correct form is *to start with*)
b. But there are other VR equipment quite developed. *Next,* some of those equipments are going to be described.
c. *Moving to the next point,* generation IV reactors are one type of . . .

The presence in learners' essays of items that are more typical of speech than of writing has already been noted by several researchers (e.g. Granger and Tyson, 1996; Gilquin and Paquot, 2007), who point out that the lack of register awareness is shared by learners from several mother tongue (L1) backgrounds (Gilquin and Paquot, 2007).

(vi) Errors regarding genre phraseology

The texts written by our students exhibit an absence of phraseology used to mark the different moves or parts of academic and professional written genres. This suggests that students are not familiar with generic conventions or with the phraseology of academic and professional discourse. For instance, very few texts state their purpose or goal explicitly. Lexical bundles such as *the objective/purpose/aim/ goal (of this report) is* are missing. The word *objectiv*e occurs 10 times, but only 3 in the first part of the text. In 4 cases it occurs in the last paragraph, when the learner wants to summarize. Similarly, *goal* tends to occur at the end of the text, when stating the conclusion or recommendation.

(16) Our *goal* should be to convert our firm in the first one releasing a playable virtual console.

The lexical signalling of specific moves in a genre is related to what Hoey (2005) calls 'textual colligation': the fact that words/items are primed to occur in or avoid certain textual positions. Words such as *goal, purpose,* etc. would be expected to occur in the first part of the text, where the purpose needs to be made explicit. But this is not the case in our corpus.

A striking example of deviant use in textual colligation is the use of *First* and *First of all* in text initial position (in the first sentences of the text) (e.g. 17). In a few cases *First/First of all* is the first item in the text (e.g 17a).

(17) a. *First of all,* let's define what virtual reality is.
b. The text deals with global warming. *First of all,* current and future consequences are going to be raised.

The use of *First/ First of all* in text initial position suggests that students are not familiar with the moves of the report genre, and therefore do not state the purpose of the report or provide background information before presenting the different parts of the report. When *First of all* occurs in text initial position the first move in the text is 'indicating text structure', a move that does not occur as the first one in academic or professional written genres. As pointed out above, the use of items such as *First/First of all* may be due to the influence of the instruction they have got on how to structure oral discourse.

Finally is another item which was used to signal report moves in a different way as it would be expected in expert writing. Although *finally* is normally used to indicate that the writer is reaching the end of a list (Chalker 1996: 56), in the learner corpus *finally* is sometimes used at the end of the text, before presenting a conclusion.

(18) *Finally*, everybody must contribute to decrease CO_2 emissions. We must become more environmentally aware.

From my point of view is also frequently found in the learner corpora at the end of the text, as an item introducing a recommendation (e.g. 19). This item was used in the learner corpora more frequently than it would be expected in academic or professional writing. In the corpus there were 10 occurrences of *from my point of view*, 30 occurrences of *in my opinion* and 2 occurrences of *in my view*. A search in the written component of the *BNC* yielded the following results: *from my point of view* (49), *in my opinion* (484), *in my view* (434*)*. Thus, while in native written discourse *from my point of view* is much less frequent than *in my opinion* or *in my view*, in our learner corpus *in my opinion is* only three times more frequent than *from my point of view* and *in my view* is much less frequent. These results are in agreement with Gilquin and Paquot's (2007) finding that the expression *from my point of view* is overused in learner academic writing.

(19) In conclusion, *from my point of view* the company should develop videogames with all the potential that the technology permits us to exploit the sound and the sight.

(vii) Errors involving signalling nouns

Signalling nouns such as *reason, conclusion, consequence, result, effect, advantage*, etc. are powerful cohesive devices in academic and professional writing, but students seem to have problems in using them to structure the text. This may be due to the fact that, as Flowerdew (2006: 345) points out, lexical cohesion has been neglected in the teaching of English as a second language. Signalling nouns are usually realized either earlier in the text, in a previous clause or clauses (e.g. 20), or later in the text, in the following clause(s) (e.g. 21) or as post-modification within the noun group.

(20) Although ebooks appeared years ago they have never really got off the ground. *This situation* is due to the lot of problems that ebook publishing entails.

(21) In this group we can underline *the following effects*: a change in the weather in several countries, . . .

J. Flowerdew's (2003) analysis of native corpora shows that in native writing the meaning of signalling nouns is most often realized across-clause than in the clause. However, our students do not seem to master the use of signalling nouns to create cohesive relations and, although there are cases of signalling function across clause, signalling nouns are most frequently used with their realization within the clause. Learners over-rely on the in-clause use of signalling nouns (e.g. 22) and underexploit them at the across-clause level. The analysis of our corpus also reveals that some learners have problems with the form of the post-modification. In some cases the problem arises from the fact that learners post-modify the signalling noun with a prepositional phrase instead of with a that-clause (e.g. 23).

(22) *The most interesting aspect* of this method is that the resulting compounds are carbon dioxide, water and other compounds that do not damage the environment.

(23) Also, sometimes there are special gloves to simulate the sense of touch, but they have the *drawback* of sweat with them

In his study of errors in signalling nouns in a corpus of essays written by Cantonese L1 learners of English, Flowerdew (2006) found four types of errors which were also common in our corpus: colligation, collocation, incorrect signalling nouns and missing signalling nouns. For instance, a frequent error is the use of the wrong lexical noun to label a stretch of text. In example (24) it is not clear what 'that point' encapsulates, and in example (25) it would be more appropriate to use 'this type of storage'.

(24) If hydrogen economy were viable, the transportation sector would be clean and carbon dioxide emissions would be less. But *that point* is at a developing stage.

(25) Hydrogen can be stored in fossil fuels. *This way* is known as liquid carrier storage.

A signalling noun that is sometimes misused is *fact*, which occurs 98 times in our corpus in different patterns. Thirty-eight occurrences of the word function as signalling noun. *Fact* seems to be an all-purpose word, very frequently used by learners instead of an appropriate nominalization. For instance, in example (26) *rise/increase (in temperature)* might be more appropriate alternatives.

(26) According to some models and sources, it is expected that global temperatures will rise between 1.4 and 5.8°C (2.5 to 10.5°F) between 1990 and 2100. *This fact* will cause climatic and environmental changes in the ecosystem.

The misuse (and overuse) of some signalling nouns seems to be due in part to the fact that learners have problems with nominalizations and that they rely on a limited repertoire of signalling nouns. In this respect, an interesting remark made by Flowerdew (2006: 352) is that the use of a wrong signalling noun 'might be considered developmental in terms of second language acquisition'. The learners understand how signalling nouns can be used to encapsulate a concept in the preceding or following co-text, but they lack the knowledge of the specific abstract noun to be used in a given cotext.

Another common error in our corpus, also pointed out by Nesselhauf (2004) and Flowerdew (2006) was the use of some typical EAP nouns (e.g. *action, aim, attitude, problem, question, statement, step* and *conclusion*) with deviant verbs (e.g. *lay out facts*).

(27) *Given all the facts laid out above,* everybody must do something to improve this situation.

Another frequent atypical grammatical pattern with signalling nouns is '*There are* (several)+ signalling noun'. Although with some nouns the pattern is correct, in other cases a different structure would be more appropriate. For instance, in examples (28a) and (28b) appropriate alternatives would be 'current technology poses several problems'/or 'solar energy has several advantages'.

(28) a. There are several problems in the current technology
b. Solar energy (. . .) There are several advantages, for instance, it's an inexhaustible supply of energy . . .

6.5 Conclusions

The purpose of this study was to identify cases of misuse or atypical use of items with organizational or rhetorical functions. In addition to confirming many of our intuitions concerning learners' problems when using these items, the results have also drawn our attention to some problems we had not previously noticed or whose importance we had overlooked.

The analysis has revealed that some of these items are used repeatedly with meanings or functions they do not have in the writing of native speakers. One of the aspects that seemed to be more problematic for our students was the use of metadiscursive multi-word sequences to organize the text and indicate transitions. Likewise, many students lacked the skill to correctly use nominalizations and signalling nouns as cohesive devices, which led them to produce problematic or erroneous language when trying to use other cohesive devices instead. Even when signalling nouns were used to encapsulate a preceding or following text, students frequently stuck to a few signalling nouns and so tended to use nouns that were not appropriate in the co-text. The study has also shown students' difficulty to distinguish formal/informal and written/oral registers. Many of the items used to structure the text were more typical of oral academic or professional genres than of written genres.

The cases of misuse or atypical use of organizing and rhetorical items are due either to the learners' inadequate linguistic competence or to their lack of genre awareness and familiarity with genre conventions. Thus, we agree with Connor et al. (2002) when they claim for the need to develop genre-specific learner corpora which facilitate the analysis of student writing for specific purposes and show the difficulties that learners have to acquire genre knowledge. The results from analysing this kind of corpora could inform learning materials both to enhance learner's writing competence in general and to improve their generic competence.

Finally, it should be pointed out that some of the errors identified in our study seem to be due to the influence of the mother tongue, but others (e.g. the overuse of some items, the difficulty to ascribe an item to a specific register) seem to result from the instruction they have received. Students often tend to oversimplify and use a limited set of items for expressing some functions, without realizing that some apparently similar items may have different pragmatic meanings or may express different degrees of formality. More classroom time should be devoted to teaching organizing and rhetorical items at the discourse level, focusing on how they can be used to produce coherent, effective and pragmatically appropriate texts. Materials designed for this purpose should be based on the analysis both of native and learner corpora, since only learner corpus data can reveal problematic areas for students and help to identify learner patterns that differ from expert writing patterns.

Acknowledgements

The present study is part of a research project financed by the University of Zaragoza. Ref: UZ2006-HUM01.

Notes

[1] Cortés (2004: 400) defines lexical bundles as 'extended collocations, sequences of three or more words that statistically co-occur in a register', e.g. *as a result of, on the other hand, in the case of the, the context of the, it is likely to.*

References

Aktas, N. (2005), 'Functions of "Shell Nouns" as cohesive devices in academic writing: A comparative corpus-based study'. Paper presented at ICAME 26 (International Computer Archive of Modern and Medieval English) – AAACL 6 (American Association of Applied Corpus Linguistics), University of Michigan, 12–15 May 2005.

Aston, G. (2000), 'Corpora and language teaching', in Burnard, L. and McEnery, T. (eds), *Rethinking language pedagogy from a corpus perspective: Papers from the third International Conference on Teaching and Language Corpora.* Hamburg: Peter Lang, pp. 7–17.

Biber, D. (2004), 'Lexical bundles in academic speech and writing', in Lewandowska-Tomaszczyk, B. (ed.), *Practical Applications in Language and Computers* (PALC 2003). Frankfurt am Main: Peter Lang, pp. 165–178.

Biber, D., Conrad, S. and Cortés, V. (2004), '*If you look at . . .*: Lexical bundles in university teaching and textbooks'. *Applied Linguistics*, 25, (3), 71–405.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999), *Longman Grammar of Spoken and Written English.* Harlow: Longman.

Bolton, K., Nelson, G. and Hung, J. (2003), 'A corpus-based study of connectors in student writing: Research from the International Corpus of English in Hong Kong (ICE-HK)'. *International Journal of Corpus Linguistics*, 7, (2), 165–182.

Carrió, M. L. (2006), 'The use of connectors in scientific articles by native and non-native writers'. *English for Specific Purposes World*, 13, 5. Retrieved: 20 September 2007, from: http://www.esp-world.info/Articles_13/Connectors.htm

Chalker, S. (1996), *Collins COBUILD English Guides 9: Linking Words.* London: HarperCollins

Connor, U., Precth, K. and Upton, T. (2002), 'Business English: Learner Data from Belgium, Finland and the U.S', in Granger, S., Hung, J. and Petch-Tyson, S. (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language*

Teaching. Language Learning and Language Teaching 6. Amsterdam and Philadelphia: Benjamins, pp. 175–194.

Cortés, V. (2004), 'Lexical bundles in published and student disciplinary writing: Examples from history and biology'. *English for Specific Purposes*. 23, (4), 397–423.

De Cock, S. (2003), *Recurrent Sequences of Words in Native Speaker and Advanced Learner Spoken and Written English: A Corpus-Driven Approach*. Unpublished Ph.D. dissertation. Louvain-la-Neuve: Université catholique de Louvain.

Flowerdew, L. (1998), 'Integrating expert and interlanguage computer corpora findings on causality: Discoveries for teachers and students'. *English for Specific Purposes*, 17, (4), 329–345.

—(2000), 'Investigating referential and pragmatic errors in a learner corpus', in Burnard, L. and McEnery, T. (eds), *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt am Main: Peter Lang, pp. 145–154.

—(2003), 'A combined corpus and systemic-functional analysis of the problem-solution pattern in a student and professional corpus of technical Writing'. *TESOL Quarterly*, 37, (3), 489–512.

Flowerdew, J. (2003), 'Signalling nouns in discourse'. *English for Specific Purposes*, 22, 329–346.

—(2006), 'Use of signalling nouns in a learner corpus'. *International Journal of Corpus Linguistics*, 11, (3), 345–362.

Gilquin, G. and Paquot, M. (2007), 'Spoken features in learner academic writing: Identification, explanation and solution', in Davies, M., Rayson, P., Hunston, S. and Danielsson, P. (eds), *Proceedings of the Fourth Corpus Linguistics Conference, University of Birmingham, 27–30 July 2007*. Retrieved: 20 March 2008, from: http://ucrel.lancs.ac.uk/publications/CL2007/paper/204_Paper.pdf

Granger, S. (ed.) (1998), *Learner English on Computer*. London & New York: Addison Wesley Longman.

Granger, S. and Tyson, S. (1996), 'Connector usage in the English essay writing of native and non-native EFL speakers of English'. *World Englishes*, 15, 19–29.

Groom, N. (2005), 'Pattern and meaning across genres and disciplines: an exploratory study'. *Journal of English for Academic Purposes*, 4, (3), 257–277.

Hoey, M. (2005), *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Nesselhauf, N. (2004), *Collocations in a Learner Corpus*. Amsterdam: Benjamins.

Oakey, D. (2002), 'Formulaic language in English academic writing: A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines', in Reppen, R., Fitzmaurice, S. M. and Biber, D. (eds), *Using Corpora to Explore Linguistic Variation*. Amsterdam and Philadelphia: Longman, pp. 111–129.

Sinclair, J, Hanks, Patrick, Fox, Gwyneth, Moon, Rosamund, and Stock, Penny (eds) (1987 1st ed.) (1995), *Collins Cobuild English Dictionary*. London: HarperCollins Publisher.

Ward, J. (2007), 'Collocation and technicality in EAP engineering'. *Journal of English for Academic Purposes*, 6, (1), 18–35.

Chapter 7

A Corpus-Informed Approach to Teaching Lecture Comprehension Skills in English for Business Studies

Belinda Crawford Camiciottoli
Università degli Studi di Firenze

7.1 Introduction

The influence of corpus linguistics on ESP research has grown steadily in recent years. The availability of increasingly user-friendly software for text analysis has created greater incentive and more opportunities for ESP researchers to compile and analyse specialized corpora. As a result, there are now abundant studies that provide insights into how authentic spoken and written language is used in various professional settings (Partington 2003; Bhatia et al. 2004; McCarthy and Handford 2004 – to cite only a few). However, what has received less attention is how this knowledge can be fruitfully applied in the ESP classroom. In other words, while a corpus-based approach to ESP research is now well consolidated, the same cannot be said with regard to ESP pedagogy. There are several possible explanations for this gap. As pointed out by Gavioli (2005: 6), a corpus of specialized language must first be carefully designed not only to represent the domain of specialization, but also with a particular learning objective in mind. Then the corpus must be collected and thoroughly analysed before it can be used in ESP contexts. This process is obviously quite lengthy and complex. Moreover, when the corpus is based on spoken discourse, other complications can emerge. First of all, it is necessary to identify speech events that are suitable for recording or, if already recorded, suitable for inclusion in the specialized corpus. In either case, the more specialized the corpus is, the fewer options there will be. Second, there may be issues of accessibility when collecting spoken data, i.e. not all speakers may agree to be recorded. Finally, the recorded data must be transcribed for successive analysis with corpus tools. Transcription is a notoriously time-consuming task, with one hour of speech taking up to 20 hours to transcribe (McCarthy 1998: 12). These are all plausible reasons why the use of corpora in ELT teaching contexts, typically based on pre-existing corpora or easily accessible texts of general language, is clearly a step ahead of the use of corpora in the ESP classroom.

Starting in the 1990s, language researchers and ELT practitioners began to explore the benefits of using corpora with L2 learners. One way that this has been done is through 'corpus-informed' (McCarthy 2004: 18) approaches where corpus research findings and corpus data itself can be used to design curricula and materials. For example, Carter (1998) discussed the need to exploit the findings from the CANCODE corpus of everyday spoken English to make L2 learners aware of important features of speech that are often neglected in standard grammars (e.g. ellipsis and vague expressions). Thurstun and Candlin (1998) described a project to develop self-study materials for learning academic vocabulary based on concordance lines drawn from the *Microconcord corpus of academic texts*. Tribble (1997) proposed the creation of small informally produced corpora derived from CR-ROM encyclopaedias or internet sources on which materials to teach lexical patterning and collocations can be based. Corpora can also prove useful when teaching features of language that are particularly problematic for L2 learners. On the basis of a corpus of newspaper texts, Partington (1998) showed that *if*-constructions were much more varied and complex in this authentic usage with respect to traditional forms presented in most ELT materials. Römer (2004) found the same discrepancy in the way that *if*-clauses were presented in a corpus of ELT textbooks in comparison with spoken language from the BNC corpus.

Corpora may also be used directly by learners. This application is inspired by the concept of data-driven learning (Johns 1991, cited in Hadley 2002) in which L2 learners make their own discoveries about language. Rather than first learning grammar rules and then trying to apply them, in this inductive process learners formulate rules by themselves through the analysis of databases of real-life language with corpus tools, such as concordancers and wordlists. According to Stevens (1995), using corpus techniques in the classroom has some important advantages. Corpora imbue the learning process with authenticity and allow learners to take control, acting as language 'researchers' in their own right. For example, Gavioli and Aston (2001) described an episode when Italian learners of English used concordances to carry out a contrastive analysis of *food* and the Italian equivalent *cibo* and succeeded in pinpointing some marked differences in usage. Hadley (2002) used concordance lines of *eaten* with Japanese EFL students to enhance their awareness of how the word is actually used in authentic language (i.e. the COBUILD Bank of English), which was quite different from how it was presented somewhat unnaturally in their grammar books (i.e. *it was eaten by me*). Cobb's (1997) study showed how concordance software was used by Arab learners to acquire new English vocabulary through a series of progressively difficult activities. In Davies' (2004) study, learners of Spanish enrolled in online courses in the USA used large Spanish language corpora to research and 'discover' forms of syntactic variation, thus moving beyond the prescriptive approach of most foreign language textbooks.

Relatively few studies have focused on how to use corpora specifically in ESP settings. Flowerdew (2001) used concordancing techniques in a specialized corpus of biology texts in order to select the most salient linguistic features to include in the ESP syllabus, to extract language examples for instructional materials and to evaluate the authenticity of currently used materials. Fuentes and Rokowski (2003) described how a specialized corpus of texts in the area of business and information technology was used to design tasks for Business English students. Gavioli (2005) has used two 'hand-made' sets of specialized corpora extensively in ESP courses on medical translation and on European issues. Respectively, the medical corpora contain research papers on various topics of medical research, while the economic-political corpora comprise research papers and speeches about the European monetary union, marketing and business management. The students successfully engaged in concordance-driven learning activities based on the two corpora.

As a further step in this direction, this chapter shows how a small specialized corpus can be utilized to benefit learners in English for Business Studies (hereinafter EBS) teaching contexts, with particular reference to content lecture listening skills. Previous studies indicate that understanding content lectures in English is often quite challenging for L2 learners, even those at relatively high proficiency levels (Thompson 1994; Mulligan and Kirkpatrick 2000), with one of the major causes of difficulty being unfamiliar vocabulary (Kelly 1991; Rost 1994). As an interdisciplinary field, Business Studies lectures contain particularly challenging lexis for L2 learners. Indeed, we find terminology not only from the core subjects of economics, accounting and marketing, but also from other related domains such as law, statistics, information technology and human resource management. In addition, the interdisciplinarity of Business Studies lectures is reflected in different knowledge orientations which are all integrated into the language: theoretical vs. practical and empirically based vs. humanities-based. In the era of globalized education with more opportunities for international exchanges, it is of paramount importance that L2 business students acquire a solid grasp of the multi-dimensional lexis associated with their discipline.

Using both quantitative and qualitative techniques, I analysed the specialized lexis of a small corpus of business studies lectures in order to determine a set of items that L2 students need to acquire for successful comprehension. I then show how the findings can be incorporated into learning activities in the EBS classroom, thus lending further support to the value of corpus-based research for the development of more effective methodologies and authentic materials in ESP pedagogy.

7.2 The Business Studies Lecture Corpus

This study is based on a specialized corpus which consists of the transcripts of twelve lectures in the area of business studies. The Business Studies Lecture Corpus (hereinafter BSLC) was specifically designed and compiled to gain more understanding of the linguistic and discursive features of this particular type of lecture, with particular reference to those that play a fundamental role in L2 listening comprehension.

Table 7.1 provides an overview of the BSLC. As can be seen, the corpus includes lectures that deal with a variety of business studies topics. They were delivered by both native and non-native speakers of English to both undergraduate and postgraduate students in large (>40) and small (<40) class sizes, thereby representing the types of lectures to which international business students are typically exposed.

Lectures 1–6 were derived from a guest lecture series offered at the University of Florence Faculty of Economics. Guest lectures are a valid way

Table 7.1 The Business Studies Lecture Corpus (BSLC)

Topic	Source/Setting	Speaker status	Level	Class size	Word count
1. SMEs in the UK	Florence/L2 guest	NS/BR	UG	Small	5,566
2. The Japanese Economy	Florence/L2 guest	NS/BR	PG	Small	11,460
3. UK Business Strategies	Florence/L2 guest	NS/BR	UG	Small	9,444
4. Productive Systems in Spain	Florence/L2 guest	NNS	UG	Small	14,667
5. SMEs in Aachen (Germany)	Florence/L2 guest	NNS	UG	Small	3,665
6. UK Industrial Policy	Florence/L2 guest	NNS	UG	Small	12,905
7. Labor Economics	MICASE/L1 class	NS/US	UG	Small	12,005
8. Macroeconomics	MICASE/L1 class	NS/US	PG	Large	8,046
9. Economic Principles	NYU/L1 class	NS/US	UG	Large	6,138
10. Ethics and Economics	NYU/L1 class	NS/US	UG	Large	7,410
11. Microeconomics	Iowa/L1 class	NS/US	UG	Large	7,006
12. Industrial Organization	Ohio/L1 class	NS/US	UG	Large	11,137
				Total words	109,449

to provide L2 learners with new listening experiences and to achieve an 'international perspective' that is desirable in many disciplines (Crawford Camiciottoli 2007: 33). These six lectures were attended by NNS students (both Italian and various other nationalities), and therefore constitute an L2 setting.

Lectures 7–12 were procured from various sources. Two (University of Iowa and North Central State College in Ohio) were available via internet. Two lecture transcripts and corresponding audio files were retrieved from the MICASE online corpus (Simpson et al. 1999). Finally, two lectures were recorded at New York University. The transcriptions and audio tapes were courteously provided by a colleague from another Italian university. These six lectures were attended by NS students who shared the same speech community with the NS lecturers, and therefore constitute an L1 setting.

7.3 The Analysis of Specialized Lexis in the BSLC

Specialized lexis can comprise more than the technical terms associated with a given domain, including items that are neither strictly technical nor domain-exclusive, but still have an important role in the context of interaction (Drew and Heritage 1992: 29). This might comprise *semi-technical* or *sub-technical* lexis, meaning 'words which are not specific to a subject specialty but which occur regularly in scientific and technical texts' (Kennedy and Bolitho 1984: 57). In business studies, these would include terms like *market* or *price* that are neither strictly technical nor discipline-exclusive and may easily be found in everyday language. Nevertheless, they are fundamental concepts in business studies and can thus be categorized as specialized lexis.

The analysis was undertaken using a two-pronged methodology that integrates both quantitative and qualitative analytical techniques. As a first step, a wordlist was generated using *Wordsmith Tools* (Scott 1998). To facilitate the elimination of large series of unwanted items, the wordlist function was set to include only words of three or more letters, thereby excluding many frequently appearing grammatical items. In addition, the minimum frequency for inclusion was set at 10 tokens. This procedure produced a list of 1,078 different word types. I then manually sorted through the alphabetically ordered list to distinguish between specialized and non-specialized items. Non-specialized items were considered to be words from general English that do not have any particular semantic link to business contexts (e.g. *you, here, yesterday, talk*) and could usually be quickly identified as such.

An item was instead classified as specialized if it could be interpreted as belonging to one of five semantic categories inspired by previous research (Poncini 2004: 152; Nelson 2005: 223):

1. Business/economics concepts (e.g., *turnover, innovation, competition)*
2. Business entities and actors (e.g., *firms, companies, partner*)
3. Business activities (e.g., *production, manufacturing, input*)
4. Description of business activities and economic trends (*profitable, failing, deal*)
5. Measurement of business performance (*price, cost, rate*)

Table 7.2 shows a short sample of the alphabetized and lemmatized wordlist. The items in bold were classified as specialized, e.g. *account,*

Table 7.2 Sample of wordlist generated from the BSLC

N	Word	Frequency	Lemmas	N	Word	Frequency	Lemmas
1	AACHEN	31		31	ALWAYS	38	
2	ABLE	32		32	AMOUNT	50	
3	ABOUT	496		33	ANALYSIS	32	analyze (8)
4	ABOVE	12		34	AND	2,721	
5	ABROAD	17		35	ANNUAL	12	
6	ACCORDING	10		36	ANOTHER	80	
7	**ACCOUNT**	**18**		37	ANSWER	16	
8	ACROSS	24		38	ANTICIPATED	10	
9	ACT	14		39	ANY	111	
10	ACTIVITIES	116	activity (77)	40	ANYTHING	40	
11	ACTUAL	13	actually (149)	41	ANYWAY	18	
12	ADD	25		42	APARTMENT	14	
13	AFTER	47		43	APART	2	
14	AGAIN	80		44	APPROACH	20	
15	AGAINST	21		45	ARE	910	
16	AGE	53		46	AREA	120	areas (48)
17	**AGGREGATE**	**23**		47	ARGUE	28	argument (15)
18	AGO	31		48	AROUND	69	
19	AHEAD	13		49	ASK	54	asked (23)
20	AID	11		50	**ASSETS**	**21**	
21	**ALERTNESS**	**28**		51	ASSISTANCE	10	
22	ALL	342		52	ASSUME	44	assumption (32)
23	ALLOW	14	allows (2)	53	ATTRACT	22	attractive
24	ALMOST	18		54	AUTHORS	10	
25	ALONG	16		55	AVAILABLE	12	
26	ALREADY	24		56	**AVERAGE**	**37**	
27	ALRIGHT	38		57	AWAY	43	
28	ALSO	153		58	BACK	78	
29	ALTERNATIVE	3	alternatives (10)	59	BAD	29	
30	ALTHOUGH	15		**60**	**BANK**	**96**	**banks (50)**

aggregate, alertness. As might be expected, there was some overlapping among the categories. For instance, a term such as *assets* can be interpreted as both a concept and a measurement of business performance. However, because the aim here was to identify specialized lexis for teaching purposes and not to produce absolute quantitative data, any overlapping was not seen as a problem.

To check for the reliability of the item classification, I asked a colleague who also teaches EBS to review the lists of specialized and non-specialized items, using the same classification criteria described above. Her assessment largely concurred with mine; we discussed the very few discrepancies of opinion until we came to an agreement. This procedure resulted in a final list of 174 lemmas which I believe can be considered as 'core' specialized lexis in business studies lectures (see Table 7.3).

7.4 Classroom Applications

Once the essential specialized lexis of the business studies lectures had been identified, it was then possible to look for ways to use it in the EBS classroom on two different yet interconnected levels: as guidelines for teachers in course/materials selection and preparation, and as tools for hands-on use by learners.

7.4.1 Corpus-informed instructional materials

Following Flowerdew (2001), perhaps the initial usefulness of the list of core specialized lexis is as a reference source for teachers when evaluating existing instructional materials and selecting new ones. It is important to choose published EBS textbooks with an overall vocabulary input that largely covers the items in the list. Moreover, the list serves as a valuable guide for teachers who wish to prepare their own materials for students. For example, teachers could formulate sets of semantically related items (e.g. *innovate, research, cooperation, collaborate, investment, dynamic*) and match them to the topics of ad-hoc texts or audio/visual materials.

The BSLC corpus could serve as a source for tasks that target both vocabulary acquisition and lecture listening skills. Using text excerpts prepared from the lecture transcripts, various types of exercises could be devised to help learners recognize and assimilate the meanings of key specialized vocabulary: gap-filling, matching or multiple choice. Extracts from the transcripts could also be used to help students become aware of the multi-di-

Table 7.3 Specialized lexis in the BSLC ranked according to frequency (number of tokens in parentheses)

1	firm (423)
2	companies (378)
3	product (362)
4	economy (292)
5	price (235)
6	percent (231)
7	work (223)
8	market (189)
9	retire (186)
10	investment (176)
11	model (174)
12	industry (168)
13	dollar (160)
14	employ (159)
15	policy (157)
16	sector (154)
17	growth (149)
18	value (144)
19	rate (139)
20	business (135)
21	pay (120)
22	stock (120)
23	GDP (117)
24	money (111)
25	profit (107)
26	wage (103)
27	innovation (108)
28	manufacturing (104)
29	competition (100)
30	service (97)
31	bank (96)
32	interest (96)
33	increase (91)
34	corporate (83)
35	bond (78)
36	data (76)
37	cost (73)
38	labour (70)
39	rent (70)
40	develop (68)
41	finance (68)
42	tech (68)
43	capital (67)
44	income (66)
45	foreign (63)
46	figure (62)
47	export (61)
48	quantity (60)
49	measure (59)
50	sell (59)
51	equilibrium (57)
52	supply (56)
53	fund (55)
54	utility (55)
55	index (51)
56	spend (50)
57	collaborate (47)
58	exchange (46)
59	objective (46)
60	option (45)
61	buy (44)
62	partner (44)
63	union (44)
64	unit (44)
65	tax (43)
66	saving (42)
67	nominal (39)
68	overseas (38)
69	average (37)
70	decline (37)
71	derivative (37)
72	goods (37)
73	job (37)
74	SME (37)
75	demand (36)
76	management (36)
77	trade (35)
78	net (34)
79	bucks (33)
80	structural (32)
81	curve (31)
82	enterprise (30)
83	fixed (30)
84	number (30)
85	calculate (29)
86	risk (29)
87	share (29)
88	alertness (28)
89	output (28)
90	quality (28)
91	transnational (28)
92	cluster (27)
93	inflation (27)
94	input (27)
95	basket (24)
96	consumption (24)
97	household (24)
98	aggregate (23)
99	worth (23)
100	fiscal (22)
101	assets (21)
102	domestic (21)
103	fail (21)
104	research (21)
105	strategic (21)
106	dynamic (20)
107	equation (20)
108	group (20)
109	turnover (20)
110	deal (19)
111	discount (19)
112	entrepreneur (19)
113	housing (19)
114	international (19)
115	monetary (19)
116	pension (19)
117	technology (19)
118	yen (19)
119	account (18)
120	recession (18)
121	windfall (18)
122	deflator (17)
123	range (17)
124	skill (17)
125	board (16)
126	impact (16)
127	pounds (16)
128	probability (16)
129	divide (15)
130	estimate (15)
131	institution (15)
132	parameter (15)
133	reduce (15)
134	borrow (14)
135	credit (14)
136	gain (14)
137	gross (14)
138	raise (14)
139	survey (14)
140	earnings (13)
141	falling (13)
142	horizontal (13)
143	organisation (13)
144	proportion (13)
145	variable (13)
146	drop (12)
147	inefficient (12)
148	owners (12)
149	package (12)
150	statistics (12)
151	surplus (12)
152	capacity (11)
153	cooperation (11)
154	database (11)
155	euro (11)
156	exogenous (11)
157	expand (11)
158	expenditure (11)
159	external (11)
160	flowing (11)
161	loss (11)
162	meeting (11)
163	micro (11)
164	minus (11)
165	report (11)
166	residual (11)
167	security (11)
168	dividend (10)
169	ethical (10)
170	flexible (10)
171	liability (10)
172	maximize (10)
173	opportunity (10)
174	sample (10)

mensional character of some key items, such as *firm*, which is used in both theoretical and practical contexts, as shown below:

> *Theoretical context*
> So the goods that are flowing away from the *firm* are the products that the *firm's* producing. In other words GDP. *(Lecture 11 – Microeconomics)*
>
> *Practical context*
> I mean that small *firms* are an important layer in the economy. *(Lecture 6 – UK Industrial Policy)*

Perhaps at a later stage, students could read excerpts of transcripts and be asked to identify specialized lexis themselves. To prepare a pre-lecture listening activity, teachers could use the wordlist function to determine the key specialized lexis of one lecture from the corpus. The list could be then presented to students who could predict what the lecture is about before they listened to it. While listening to or watching recorded lectures, students could refer to various sets of specialized vocabulary drawn from computer-generated wordlists and select the one that best corresponds to the topic of the lecture. As a more advanced task, students could listen to a lecture excerpt, list all specialized items that they hear and then check to see if this matches a list prepared by the teacher (based on the master list of specialized lexis). As a follow-up task, students could prepare a short oral or written summary that incorporates the items that they identified.

7.4.2 Corpus-based learning activities

If students have been familiarized with the basics of corpus methodology, with careful guidance they can be encouraged to use BSLC themselves.[1] As a beginning task, students could be given some sets of concordance lines generated from specialized lexical items, but with the search word blanked out. They could then be asked to complete the concordances appropriately choosing from a list of options. To increase awareness of collocations and colligations, students could also be presented with complete concordances and asked to study and identify patterns in the items that surround the specialized lexis. For example, a few concordances of *fund* (see below) could help students discover its usage as both a noun and a verb, common collocations/colligations and especially the important noun + noun pattern often found in business English (e.g. *capital* *fund* and *retirement fund).*

```
N Concordance
1 und they created a regional venture capital fund together with some
2 w it's too risky the what the banks want to fund uh is large companies
3 d to take them under government control to  fund them and restructure t
4 blah blah blah and he says you own a mutual fund? Yeah I've got a
5 n a mutual fund? Yeah I've got a retirement fund I've got a mutual fund
6 bad corporations If they have a retirement  fund and insurance policy
7 they own any stock or if they own a mutual  fund I love to- I love to t
8 a link between him and uh uh that endowment fund that he and his wife put
```

As a more advanced discovery-driven task, students could be asked to generate a wordlist of one lecture and then decide for themselves which items are specialized or not. Some interesting discussion is likely to emerge as they justify their choices. They could also carry out follow-up concordancing and cluster analysis of the items they selected to discover usage tendencies.

7.5 Conclusion

In this chapter, I have shown how a small specialized corpus can be analysed to determine its key discipline-related lexis and how these findings can be developed into corpus-informed, corpus-based and corpus-driven activities to help EBS students better understand content lectures in English. Because the language used by business studies lecturers represents a merger of the academic, disciplinary and professional worlds (Crawford Camiciottoli 2007: 190), the ability to understand these lectures is crucially important not only for the learners' more immediate academic needs, but also for their future careers.

With this study focusing on a particularly popular area of language study (i.e. business English), I hope to have made a contribution towards bridging the gap between ESP research and ESP pedagogy. ESP is destined to play an increasingly vital role in language teaching due to the growing demand for professionals with specialized knowledge and skills. Indeed, in today's globalized workplace the possession of multiple specializations has become an important advantage for job candidates (Crawford Camiciottoli 2007: 1). For this reason, additional studies targeting corpus applications in other areas of ESP (e.g. legal English or technical English) would be extremely useful.

Notes

[1] I agree with both Gavioli (2005: 29–30) and Fuentes and Rokowski (2003: 4) that students need to be guided and supervised by the teacher when learning to work with corpora and corpus tools in the classroom.

References

Bhatia, V. J., Langton, N. and Lung, J. (2004), 'Legal discourse: opportunities and threats for corpus linguistics', in Connor, U. and Upton, T. A. (eds), *Discourse in the Professions. Perspectives from Corpus Linguistics.* Amsterdam/Philadelphia: John Benjamins, pp. 203–234.

Carter, R. A. (1998), 'Orders of reality: CANCODE, communication and culture'. *ELT Journal*, 52, (1), 43–56.

Cobb, T. (1997), 'Is there any measurable learning from hands-on concordancing?'. *System*, 25, (3), 310–315.

Crawford Camiciottoli, B. (2007), *The Language of Business Studies Lectures: A Corpus-Assisted Analysis.* Amsterdam, Philadelphia: John Benjamins.

Davies, M. (2004), 'Student use of large, annotated corpora to analyze syntactic variation', in G. Aston, S. Bernardini, and D. Stewart (eds), *Corpora and Language Learners.* Amsterdam and Philadelphia: John Benjamins, pp. 257–269.

Drew, P. and Heritage, J. (1992), 'Analysing talk at work: An introduction', in P. Drew, and J. Heritage (eds), *Talk at Work. Interaction in Institutional Settings.* Cambridge: Cambridge University Press, pp. 3–65.

Flowerdew, J. (2001), 'Concordancing as a tool in course design', in M. Ghadessy, A. Henry, and R. L. Roseberry (eds), *Small Corpus Studies and ELT. Theory and Practice.* Amsterdam and Philadelphia: John Benjamins, pp. 71–92.

Fuentes, A. C. and Rokowski, P. E. (2003), 'Using corpus resources as complementary task material in ESP'. *ESP World*, 6, (2). Retrieved July 25 2007 from: http://www.esp-world.info/articles_6/C2_.htm.

Gavioli, L. (2005), *Exploring Corpora for ESP Learning.* Amsterdam/Philadelphia: John Benjamins.

Gavioli, L. and Aston, G. (2001), 'Enriching reality: Language corpora in language pedagogy'. *ELT Journal*, 55, (3), 238–246.

Hadley, G. (2002), 'Sensing the winds of change: An introduction to data-driven learning'. *RELC Journal*, 33, (2), 99–124. Retrieved September 7 2007 from: http://www.nuis.ac.jp/~hadley/publication/windofchange/windsofchange.htm.

Johns, T. (1991), 'Should you be persuaded – two examples of data-driven learning materials'. *English Language Research Journal*, 4, 1–16. University of Birmingham.

Kelly, P. (1991), 'Lexical ignorance: The main obstacle to listening comprehension with advanced foreign language learners'. *IRAL* XXIX, 2, 135–149.

Kennedy, C. and Bolitho, R. (1984), *English for Specific Purposes. ELTS.* London: Macmillan.

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

—(2004), *Touchstone. From Corpus to Coursebook.* Cambridge: Cambridge University Press.

McCarthy, M. and Handford, M. (2004), '"Invisible to us?": A preliminary corpus-based study of spoken business English', in Connor, U. and T. A. Upton (eds), *Discourse in the Professions. Perspectives from Corpus Linguistics.* Amsterdam/ Philadelphia: John Benjamins, pp. 167–201.

Mulligan, D. and Kirkpatrick, A. (2000), 'How much do they understand? Lectures, students and comprehension'. *Higher Education Research and Development,* 19, (2), 311–335.

Nelson, M. (2005), 'Semantic associations in Business English: A corpus-based analysis'. *English for Specific Purposes,* 25, (2), 217–236.

Partington, A. (1998), *Patterns and Meanings.* Amsterdam and Philadelphia: John Benjamins.

—(2003), *The Linguistics of Political Argument. The Spin-Doctor and the Wolf-Pack at the White House.* London: Routledge.

Poncini, G. (2004), *Discursive Strategies in Multicultural Business Meetings.* Bern: Peter Lang.

Römer, U. (2004), 'Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching', in G. Aston, S. Bernardini, and D. Stewart (eds), *Corpora and Language Learners.* Amsterdam and Philadelphia: John Benjamins, pp. 151–168.

Rost, M. (1994), 'On-line summaries as representations of lecture understanding', in Flowerdew, J. (ed.), *Academic Listening. Research Perspectives.* Cambridge: Cambridge University Press, pp. 93–127.

Scott, M. (1998), *Wordsmith Tools.* Oxford: Oxford University Press.

Simpson, R., Briggs, S., Ovens, J. and Swales, J. M. (1999), *The Michigan Corpus of Academic Spoken English.* Ann Arbor: The Regents of the University of Michigan. Available at http://www.hti.umich.edu/m/micase.

Stevens, V. (1995), 'Concordancing with language learners: Why? When? Where?'. *CAELL Journal,* 6, (2), 2–10. Retrieved 25 July 2007 from: http://www.eisu.bham.ac.uk/johnstf/stevens.htm

Thompson, S. (1994), 'Frameworks and contexts: A genre-based approach to analyzing lecture introductions'. *English for Specific Purposes,* 13, (2), 171–186.

Thurstun J. and Candlin, C. (1998), 'Concordancing and the teaching of vocabulary of academic English'. *English for Specific Purposes,* 17, (3), 267–280.

Tribble, C. (1997), 'Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching', in B. Lewandowska-Tomaszczyk and J. Melia (eds), *Practical Applications in Language Corpora.* Lodz: Lodz University Press, pp. 106–117. Retrieved 19 July 2007 from: http://www.ctribble.co.uk/text/Palc.htm.

Chapter 8

Creating a Corpus of EIL Cross-Cultural Interaction in the Public Domain

Maria Georgieva and Lilyana Alexandrova Grozdanova
Sofia University St. Kliment Ohridski

8.1 Introduction

In this chapter, we discuss the creation of a corpus of cross-cultural oral communication samples from the public domain. In particular, we focus on the principles of corpus building, the selection and organization of sample material, topic and genre characteristics of the data, and issues of representability of the corpus at large. Inasmuch as the corpus consists of interchanges in contact situations, attention is also paid to the social dimensions of the communicative events of interest and the instruments ensuring cross-cultural and inter-text comparability of sample data.

This corpus is designed as part of a wider project focused on the investigation of the strategies – accommodation, compensation and social – employed by EIL speakers in domain-specific cross-cultural discourse. The aim is to establish how speakers' general disposition to reach mutual understanding in interaction is moulded to fit the demands of intercultural communicative situations under the joint influence of local and global social practices and models of communicative behaviour that make up an individual's sociocultural capital. Given that the corpus consists of samples of complete communicative encounters, backed up by ample ethnographic data on the participants in the communicative events, it will have a much broader range of applicability in research and English language teaching alike.

8.2 English as an International Language

Intercultural communication is constantly growing in scope and importance as a key feature of globalization and the ever- increasing cross-border mobility of workforce and information. From a purely elitist endeavour of relatively small circles of politicians, academics, or people involved in business or culture, international communication is gradually turning into a way of life for the mass population. Whether seeking advancement on the job market of multinational companies, or crossing geographical borders for tourism, work or study, or traversing virtual spaces, taking advantage of the opportunities offered by high communication technologies, more and more ordinary people around the globe enter into communication in languages that are not their native. Against this backdrop, knowledge of a foreign language, particularly English that is unrelentingly diffusing into all spheres of life, has grown into a most highly valued resource.

As many explorers of International English (hereafter EIL[1]) have pointed out, the causes for the selection of this particular language as an instrument for international communication are complex and varied, bearing not so much on linguistic as geo-historical, socio-political, economic and technological factors (Crystal 1997; Graddol 1997; McArthur 1998; Spolsky 2004). So, although English is deeply entrenched in our postmodern world, its world role is far from uncontested, attitudes towards it going to opposite extremes. Some scholars tend to attribute its globalization to purposefully pursued policy of intentional destruction of smaller languages (Phillipson 1992) while others perceive it as a legitimate instrument of international communication that needs to be recognized as a specific variety in its own right (Seidlhofer 2001; Jenkins 2006). The debate over global socio-linguistic situation notwithstanding, EIL is a fact of life, a product and a driver of globalization (Graddol 2006). Accordingly, the tensions triggered by its unprecedented diffusion into all spheres of our life and the ensuing variability of use caused by the ever-increasing diversity of linguistic, cultural or social background of its users deserve careful consideration as they touch on the very make up of our postmodern world.

Among the important questions with a bearing on teaching and translation practices in the countries of the so-called Expanding circle,[2] the question of the nature and status of EIL seems the most intriguing, as it challenges a basic tenet of modern EFL teaching and use – the native speaker model as a benchmark for foreign language competence. Referring to some of its general characteristics of use and function, a number of analysts have voiced the view that EIL is sufficiently different from native speaker varieties to justify claims for granting it a status of an autonomous variety (e.g. Seidlhofer 2001; Jenkins 2006, etc.). Indeed, it is widely acknowledged today that as an instrument of wider communication EIL is not connected to the culture of a specific community of native speakers. On the contrary, it tends to pick up features from the cultures of those who speak it and, given the diversity of linguistic and cultural background of its speakers, may be assumed to have a multicultural base. EIL differs from native varieties also in function, as it is utilized largely to attend to social practices associated with globalization – world business and economy, technology and communications, culture, science and education. Finally, EIL tends to establish itself alongside other languages in contexts composed of bilingual or multilingual speakers (Grozdanova 2002). As such, it is predominantly utilized for instrumental purposes, as a 'language for communication' (House 2003) or, especially in the newly established democratic states, as a symbol of modernity (Georgieva 2002). Unlike national languages, which have a strong identificatory function, the use of EIL has such more practical drivers as 'communicative efficiency and economy of language learning and use' (Seidlhofer 2001: 141).

Led by the conviction that in order to gain in credibility the claims of EIL as a specific variety have to be substantiated by concrete evidence of its manifestations in diverse domains of use, scholars have launched numerous projects aimed at a systematic description of EIL discourse practices from various perspectives. A detailed account of all existing corpora of EIL goes beyond the scope of this chapter. By way of an example, we could mention the Vienna-Oxford project created to serve as a basis for describing the most salient features of English as it is actually used as a *lingua franca* with the ultimate objective of its codification as a widely accepted and respected alternative to ENL to be employed in appropriate contexts of teaching and use (Seidlhofer 2002). The ICLE[3] corpus's main thrust is to provide a basis for contrastive studies of the written production of learners from different language backgrounds that would highlight the difficulties foreign learners have with native English. The CADIS[4] corpus is built for the more socially oriented purpose of exploring the range of 'identity-shaping strategies' linked to 'local or professional cultures, as communicated by contemporary English in various domains among native and non-native speakers' (Gotti 2006: 44). There are also numerous corpora focused on the investigation of EIL discourse patterns in different specialized domains – academic, business, legal, etc.

A common feature of most existing projects is their concern with a systemic description of EIL characteristics conducive to its ultimate codification as an autonomous variety with interaction having the auxiliary function to serve as a basis for linguistic analysis. Contrariwise, the current project focuses on the process of interaction itself aiming to throw light on the specific nature of understanding, meaning negotiation and regulation of social relations in intercultural communicative situations. In particular, we want to find out what strategies participants in intercultural communicative encounters employ to overcome or alleviate differences that would inevitably arise, given that linguistic knowledge as embedded in sociocultural knowledge tends to be moulded to a lesser or bigger extent by speakers' local social practices, cultural ideologies, moral judgments and beliefs. What patterns emerge in their mutual endeavour to bring their different stances to alignment, to co-adapt their communicative performance and construct their talk in a way that is both communicatively effective and socially respectful to all those concerned?

Public discourse, selected as the subject of our investigation, is among the discourses most strongly affected by globalization with new topics and freshly coined jargon reverberating through a wide range of social practices – social management, public relations, human and citizenship rights or education. Thus, by focusing on such a perpetually changing domain, we hope to grasp the dynamic nature of intercultural communication and the way it adapts to the growing demands global processes pose to all those involved in them. Another feature of public discourse that makes it particularly valuable as a basis for interaction analysis is the low level of highly specialized terminology, conducive to a very broad social base of the study in terms of type of participants and practices. The opportunity to draw on diverse sources of information will reasonably enhance the validity of the findings and broaden the range of their application. Hence, they could prove useful to any general English course addressed to people seeking career advancement in the public sector.

8.3 Interaction and Communicative Strategies in Corpus Design: A Problem-Oriented Approach

Interaction itself is a complex process that could be approached from various angles but the main thrust of our project is to investigate whether, and to what extent, the use of communication strategies (hereafter CS) can contribute to the enhancement of the interaction process in cross-cultural contexts. The issue is important because it links cross-cultural interaction, believed to constitute a minefield of miscommunication, with strategic behaviour that implies awareness of the need of control and conscious planning in executing messages skilfully. Unfortunately, due to the proliferation of its uses, the concept of communication strategy is notoriously difficult to define. In mainstream SLA literature CS are commonly associated with presumed competence deficiencies and are defined as 'potentially conscious plans for solving what to a participant in a communicative exchange presents itself as a problem in reaching a particular communicative goal' (Faerch and Kasper 1983: 36). Scholars working in the tradition of Communication Accommodation Theory, in turn, stress on the natural inclination of speakers to modify their speech according to some situational or personal variables and make it more like that of their interlocutors for solidarity reasons or to enhance comprehension (Giles and Powesland 1975; Boggs and Giles 1999). What these two different approaches have in common is that in either case communication strategies tend to be used in situations where the interaction process fails to meet the needs of one or all participants in the encounter, i.e. there is some kind of a real or apparent problem. On these grounds we have adopted a broader definition of communication strategy with problem-orientedness as its defining characteristics and a possibility for multiple representation by a range of strategy types – compensatory, accommodation and social. Participants' awareness of the problem may vary according to its perceived relevance to the interaction process and it is not necessary for all those concerned to agree on whether or not there is a communication problem at all. However, inasmuch as strategic behaviour is triggered by a speaker's desire to make oneself better understood, or more acceptable to the person addressed it seems reasonable to assume that it will always involve a certain level of conscious planning and expended effort.

In line with the broader interpretation of CS we also adopt a looser framework of problematic communication following in the main the typology of miscommunication processes suggested by Coupland et al. (1991, quoted in Boggs and Giles 1999: 225). Accordingly, CS identification and classification will be carried out in terms of the following problem groups:

- Problems arising from personal deficiencies in speakers' competence (resource gaps) that may be conducive to such compensatory strategies as explication, language switch or avoidance.
- Problems arising from minor misunderstandings, misalignments or information gaps that may lead to the use of interpretative strategies involving meaning negotiation or meaning adjustment.
- Problems bearing on speakers' conversational needs that may trigger various discourse management strategies, such as grounding and goal alignment, coordination, face-maintenance, etc.
- Problems stemming from differences in speakers' linguistic or socio-cultural norms that may call for various approximation strategies aimed at establishing a shared assumptive framework or common ground, mutual attuning or reconciling of divergent stances.

- Problems addressing role relations conducive to a continuum of relational strategies ranging from such aimed at reducing the social distance between interlocutors and building solidarity and rapport, to strategies of conflict resolution and power management that might signal a desire to maintain, or even increase social distance.
- Problems concerned with group or culture affiliation that may spark off various maintenance strategies aimed at identity building and self-assertion.

It is important to emphasize that contrary to expressed views that CS are just analysts' constructs of specific social-psychological processes, we consider them as inherent elements of interpersonal communication, firmly rooted in a paradigm of meaning transfer between autonomous individuals irrespective of their language background or proficiency. They may be utilized separately or in clusters, concentrated in a single interaction act or spreading across a sequence of acts and are commonly, though not always, rendered visible by participants through specific actions of 'flagging' (Wagner and Firth 1997). Inasmuch as CS use involves certain costs to interlocutors in terms of expended effort, it seems natural to believe that they would not leave a strategic move to pass unnoticed and divest themselves of the potential rewards for getting social approval. Quite the contrary, they are more likely to 'flag' an upcoming problem in discourse thereby signalling that a communicative strategy is imminent, and then, individually or conjointly, try to realign their stances and resolve the problematic situation.

The following examples serve to illustrate the type of strategies that will be analysed. The first is an example of a compensatory strategy, taken from students' performance in a simulation game. It shows how the student in the role of *travel agent* 'flags' his understanding problem by an alternative question and how in the subsequent moves his partner helps him resolve it first, by repetition of the problematic word and then through elaboration on the question put in the beginning.

Example 1: Compensatory strategy
S1 (Customer): Well, what about pollution in the city?
S2 (Travel Agent): *Oh, the population . . .? . . . Pollution?*
S1: *Pollution . . . Pollution*
S2: Aha! . . . the pollution!
S1: Yes. *Is there much pollution in the streets of your city*?

The second provides an example of an accommodation strategy employed by a proficient speaker in an intercultural context. The Bulgarian's intention to signal similarity of views with her partner, or to accommodate for solidarity reasons, was obviously conceived as problematic in terms of politeness as she 'flagged' her move both verbally by *I wanted to ask you* . . . and non-verbally by laughter presumably as a sign of modesty and uneasiness. In spite of all the precautions taken by the Bulgarian speaker however, her British interlocutor chose to opt out of conjoined rapport building and used a divergence strategy responding in a way that people would normally use in self-talk. This shows how speakers' attempts at accommodation in an intercultural situation are not always, or necessarily, unidirectional.

Example 2: Accommodation strategy

S1 (Bulgarian): *I wanted to ask you something*. I overheard . . . Yeah (*laughing*) I was eavesdropping last night and I overheard you were saying something about the tradition of theoretical presentations . . . you weren't very happy with what people were doing . . . (*laughing*)

S2 (British): Oh, yeah. I'm sorry. (laughs). *I should be careful what I'm saying* . . .

S1: Oh, no, no! Well, this seemed interesting to me, you know, and . . .

The different proficiency level of the speakers in the two encounters comes to imply further that there is no direct relationship between strategic behaviour and competence deficit as commonly assumed in mainstream SLA research. This brings into relief a second focus of our research, namely, to identify the factors – linguistic, social or situational – compelling speakers to take strategic action.

The study of strategic competence has led to identification of numerous communication strategies. While the list of strategies has been growing steadily (see Tarone 1981; Faerch and Kasper 1983; Rost and Ross 1991; Williams 1999, etc.), it became clear that speakers often make use of 'their own rules' (Cohen and Aphek 1981), which poses the dilemma which method to choose in pursuing our goal – the deductive or the inductive one. We believe that the latter will be more effective for at least two reasons: first, it investigates strategic competence in a straightforward way, i.e. the data reveals what strategies the respondents *do* use and not what they *think* they use; and second, instead of using preconceived ideas we can discover new strategies unaccounted for so far. In other words, we find the inductive approach more productive as it does not predetermine the results and can help us put together a true picture of the intercultural communication in the public domain by discovering new strategies.

8.4 Designing an EIL Corpus of Cross-Cultural Oral Communication

The first stage of our study aims to collect a corpus of semi-spontaneous conversations, which will supplement the existing comprehensive corpora of English (e.g. ICE-GB, ICLE) with data about its use in oral communication, thus serving as a source of evidence for linguists, teachers and textbook writers. What will make this product valuable is the fact that, unlike previous corpora, which over-emphasize written language, or seek comparisons between native and non-native speaker performance, it will focus on the spoken mode and provide interaction samples of competent communicators who use language as a lingua franca in intercultural encounters. The collection of intercultural conversations, duly recorded, transcribed and organized for easy reference, will enable us to move on to the second stage, when samples will be analysed and a second corpus derived – that of the communicative strategies used by the interlocutors.

Aware of the difficulties to operationalize criteria for corpus building, we shall try to achieve descriptive adequacy by organizing as many cross-cultural conversation events as feasible. Five teams of linguists with different native languages and culture, namely Bulgarian, Romanian, Spanish, Italian and British will conduct interviews in five different settings, respectively, Bulgaria, Romania, Spain, Italy and the United Kingdom. Both interviewers and interviewees will be fluent EIL speakers. The interviewees will represent various spheres of the public domain such as non-government organizations, EU-related organizations, mass media, education, welfare, etc. Thus, we hope to sample as widely as possible, in a balanced and unbiased way. Sources of data will be duly stated along with other relevant information. We believe that the corpus size and diversity, as well as the numbers of researchers and respondents with different linguistic and cultural background will guarantee maximum validity of the study.

The tool to be used is a semi-structured interview that will concern a topic of current relevance in the public domain. The interview will be constructed in such a way as to create a need for social regulation and establishing common ground on the part of the interviewees, thus provoking the use of communicative strategies. By way of an example, we present a tentative interview script on the topic of *education.* As a first step, we shall prepare a bank of snippets of information to be distributed among potential conversational partners, by means of which we hope to give more substance to the interaction, and create a basis for comparison of the produced conversations.

Topic: EDUCATION

1. Read and reflect on the situation in your country.
   FACTFILE

(*front-of-class teaching*) 'If a teacher is transmuted into a deskilled lackey facilitating learning, rather than being a highly skilled actor, artist, or craftsman, the children will lose out. From our own schooling, we remember most of all the charismatic front-of-class teacher and we remember his lessons and how excited we were to be in them. The foregrounding of learning over teaching puts an end of that romantic image; it also foretells of a move towards serried ranks of students plugged into computers, supervised by IT technician.' (*Guardian,* 15 January 2008)

(*a new right to discipline*) 'Teachers in England will be given the right to discipline unruly schoolchildren outside the school gates. The new government move will give teachers a clear legal right to restrain pupils with reasonable force and confiscate "inappropriate items" outside schools without fear of repercussions. "Children are as nice now as they ever were"', a headmaster said. 'Generally they can behave a lot better than adults do – they are idealistic and altruistic, but they are also learning and need boundaries. We put in front of children a culture of greed and get-rich-quick, and then are surprised when some of the most vulnerable ones copy it.' (*The Observer,* 5 February 2006)

(*admission practices*) Some *English schools* are ignoring rules that are supposed to ensure that all children have a fair chance of gaining admission. They are using 'covert admission practices' that discriminate against poorer families such as asking parents for their marriage certificates, insisting that children wear uniform only available from expensive shops, etc. The new rules on admissions that came into force last year are designed to stop schools using subjective admissions procedures that would discriminate against low-income families or children with disabilities or special needs. (*The Guardian,* 17 January 2008)

***

Competition to get at *Oxford and Cambridge Universities* has become so intense that a 'mini industry' has been built around it, offering advice, tips and support through the application process. Parents are paying up to £3,500 for a package of tuition, mock interviews and help with completing application forms. Sales of books on how to get into Oxbridge are rocketing. (*The Times,* 1 October 07)

(Quality assurance in education) A recently published book by a professor of English at Warwick University has sparked off a heated debate about the quality in education. The cause of disagreement seems to be the Quality Assurance Agency considered by some as a 'safeguard designed to maintain and improve academic standards' and by others as 'a cancer that gnaws at the core of knowledge, value and freedom in education' and, consequently, 'the worst thing to happen to higher education in recent times'. (*The Guardian,* 17 Jan, 08)

After reading the materials and reflecting on the situation in their own countries, the interviewers will be instructed to single out two controversial issues – one they approve of and are going to support during the interview, and one they disapprove of and are going to criticize. In the next stage – the interview proper – they are to elicit similar information from their respondents:

Interview Script

1. Elicit from your partner a brief overview of the education system in his/her country:

   a. Positive/negative aspects
   b. Current reforms
   c. People's attitude towards education, etc.

2. Express a positive stand on an educational issue, initiating a discussion. e.g. *All children should have a fair chance of gaining admission to language schools.*
3. Express a negative stand on an educational issue, initiating a discussion. e.g. *Teachers should not have a legal right to restrain pupils with force outside schools.*

4. Compare practices.
   e.g. *admission to universities; or, quality assurance programs.*
5. Find out which of the following stereotypes is relevant to your partner's nation (to be enumerated).

Various degrees of accommodation are expected to be displayed in the above interaction, depending on the respondents' communicative abilities. To keep the conversation going, they will have to resolve problems concerned with information gaps and comprehensibility, accommodation of communication patterns, overcoming inter-cultural differences, rapport building and avoiding threats to face. In short, the common topic and strategy elicitation techniques will serve as a framework for the researchers involved in the project.

Additional data are to be derived from the post-interview stage, when both interviewers and interviewees will be instructed to evaluate independently the conversation in general, the contribution of their interlocutors and their own contribution, as well as any problems or peculiarities.

In the second stage of the study, the collection of transcripts in Corpus A will be analysed by each team independently. We hope that the combination of participant observation and objective analysis will lead to the discovery and explication of the recurrent communicative strategies employed by the respondents. A coding system will be developed to account for such dimensions as interpretability, discourse management, interpersonal control, positive and negative face, assertion (Gardner and Jones 1999: 204) plus any other variables that might emerge. The strategies identified will be tagged for intercultural consideration and comparison, with a view to reaching inductive generalizations. The latter are expected to filter out idiosyncratic features and explicate recurrent communication strategies. The resulting Corpus B will help outline certain similarities and differences in the socio-cultural norms of speaking, in particular, the preferred strategies by speakers using English as an international language.

We realize that a large-scale project of this kind will consume much time and collaborative effort. What makes it worthwhile is the multifaceted application of its findings. As pointed out earlier, the text collection in Corpus A can serve as a source of linguistic evidence and information. Corpus B and the respective inventory of communication strategies, in turn, will enable applied linguists, teachers and textbook writers to work towards developing learners' strategic competence. All in all, the project will help raise people's general strategic awareness and contribute to the use of EIL in a more effective way.

Notes

[1] The distinction commonly made between EIL (English as an international language) and ELF (English as a lingua franca) is deemed irrelevant for the purposes of this paper.

[2] According to Kachru's (1985) classification of English varieties into Inner, Outer and Expanding Circles.

[3] ICLE – International Corpus of Learner English (http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm)

[4] CADIS – Corpus of Academic Discourse (http://www.unibig.it/Cerlis)

References

Boggs, C. and H. Giles (1999), '"The canary in the coalmine": The nonaccommodation cycle in the gendered workplace'. *International Journal of Applied Linguistics*, 9, (2), 223–245.

Cohen, A. and Aphek, E. (1981), 'Easifying second language learning'. *Studies in Second Language Acquisition,* 3, 221–236.

Crystal, D. [1997] (2003), *English as a Global Language.* 2nd ed. Cambridge. Cambridge University Press.

Coupland, N., Wiemann, J. M. and Giles, H. (1991), *Miscommunication and Problematic Talk.* Newbury Park: Sage.

Faerch, C. and Kasper, G. (1983), 'Plans and strategies in foreign language communication', in C. Faerch and G. Kasper (eds), *Strategies in Interlanguage Communication.* Longman, pp. 20–60.

Gardner, J. M. and Jones, E. (1999) 'Problematic communication in the workplace: Beliefs of superiors and subordinates'. *International Journal of Applied Linguistics*, 9, (2), 185–205.

Georgieva, M. (2002), 'On developing intercultural communicative competence in EFL learners', in M. Georgieva and D. Thomas (eds), *Smaller Languages in the Big World.* Plovdiv: Lettera, pp. 146–159.

Giles, H. and Powesland, P. [1975] (1997), 'Accommodation theory', in Coupland N.and A. Jaworski (eds), *Sociolinguistics: A Reader and Coursebook.* London: Macmillan Press, pp. 232–239.

Grozdanova, L. (2002), 'Cultural diversity in a unifying world–a new challenge for English textbook writers', in Georgieva, M. and Thomas, D. (eds), *Smaller Languages in the Big World.* Plovdiv: Lettera, pp. 126–145.

Gotti, M. (2006), 'Creating a corpus for the analysis of identity traits in English specialized discourse'. *The European English Messenger*, 15, (2), 44–47.

Graddol, D. (1997), *The Future of English?* London: The British Council.

—(2006), *English Next.* London: British Council.

House, J. (2003), 'English as a lingua franca: a threat to multilingualism?' *Journal of Sociolinguistics*, 7, (4), 556–578.

Jenkins, J. (2006), 'Points of view and blind spots: ELF and SLA'. *International Journal of Applied Linguistics*, 16, (2), 137–162.

Kachru, B. (1985), 'Standards, codification and sociolinguistic realism: The English language in the outer circle', in R. Quirk and H. Widdowson (eds), *English in the World: Teaching and earning the language and literatures.* Cambridge: Cambridge University Press, pp. 11–30.

McArthur, T. (1998), *The English Languages.* Cambridge: Cambridge University Press.

Phillipson, R. (1992), *Linguistic Imperialism.* Oxford: Oxford University Press.

Rost, M. and Ross, S. (1991), 'Learner use of strategies in interaction: Typology and predictability'. *Language Learning,* 41, 235–273.

Seidlhofer, B. (2001), 'Closing a conceptual gap: The case for a description of English as a lingua franca'. *International Journal of Applied Linguistics,* 11, (2), 133–158.

—(2002), 'Habeas corpus and divide et impera: "Global English" and applied linguistics', in K. Sp. Miller and Thompson, P. (eds), *Unity and Diversity of Language Use.* BAAB/Continuum, pp. 198–215.

Spolsky, B. (2004), *Language Policy: Key Topics in Sociolinguistics.* Cambridge: Cambridge University Press.

Tarone, E. (1981), 'Some thoughts on the notion of communicative strategy'. *TESOL_Quarterly,* 15, 285–295.

Wagner, J. and Firth, A. (1997), 'Communications strategies at work', in G. Kasper and E. Kellerman (eds), *Communications Strategies: Psychological and Sociolinguistic Perspectives.* London: Longman, pp. 323–344.

Williams, A. (1999), 'Communication accommodation theory and miscommunication: Issues of awareness and communication dilemmas'. *International Journal of Applied Linguistics,* 9/2, 151–165.

This page intentionally left blank

Part Three

Learner Corpora and Corpus-Informed Teaching Materials

This page intentionally left blank

Chapter 9

Spoken Learner Corpora and EFL Teaching

Sylvie De Cock
Centre for English Corpus Linguistics, Université Catholique de Louvain

9.1 Introduction

Learner corpus analysis has been a very active field of research since the emergence of computerized learner corpora in the early 1990s (Granger 1998). Computer learner corpora are systematic 'electronic collections of spoken or written texts produced by foreign or second language learners' Granger (2004: 124). As is also the case for native corpora, there are far fewer learner corpora containing spoken productions than corpora containing written productions. Although both spoken and written learner corpora are extremely variable in size, the biggest spoken learner corpora (e.g. the 2-million-word *NICT JLE*, Tono 2007) are much smaller in size than the biggest written learner corpora (e.g. the 25-million-word *Hong Kong University of Science and Technology Learner Corpus*, Pravec 2002). Not only do these two observations regarding number and size hold true for what Granger (2004) calls academic learner corpora, i.e. learner corpora compiled in educational settings, but there are as yet, at least to the author's knowledge, no spoken equivalent(s) of the big commercial corpora of learner writing such as the 10-million-word *Longman Learners' Corpus* or the 15-million-word *Cambridge Learner Corpus.* Further evidence of the predominance of written learner corpora comes from the forthcoming release of the *Louvain International Database of Spoken English Interlanguage* (henceforth *LINDSEI*) at least six years after the first release of its written counterpart, the *International Corpus of Learner of English*, in 2002 (Granger et al. 2002). The bias towards written learner corpora is hardly surprising considering that collecting and transcribing spoken data is extremely tedious and time-consuming (Granger 2004; Luzón et al. 2007). Although spoken learner corpora have lagged behind written learner corpora, they are slowly but surely catching up and finding their voice. An increasing number of spoken learner corpora are currently being compiled or have seen the light of day over the past few years.

The purpose of this chapter is twofold as it sets out to give a brief overview of both spoken corpora of learners of English and spoken learner corpus research, and to assess the contribution of this type of corpus and research to EFL teaching.

9.2 The Many Voices of Spoken Learner Corpora and Spoken Learner Corpus Research

9.2.1 Spoken Learner Corpora

Rather than providing readers with a detailed survey of spoken learner corpora, this section explores some of the key features of spoken learner corpora, namely their spoken character or 'spokenness' and a number of design criteria relating to learner and task.

9.2.1.1 The 'spokenness' of learner corpora in the spotlight

Transcribing recordings of speech has been the necessary first step for researchers embarking on the linguistic investigation of spoken discourse using corpus linguistic methods and tools. Transcriptions are, however, a subjective hand-crafted product (Leech et al. 1995: 10) and can be regarded as far removed from what they are intended to represent. Representing speech through writing fundamentally alters the very nature of spoken discourse. The change of medium involved inevitably leads to a change in perception, as is illustrated by what Stubbs calls (1983: 228) the 'estrangement effect' of transcription: because of phenomena like false starts, repetitions, hesitations and overlapping utterances, spoken discourse '*looks* odd, incoherent and broken when seen in the written medium – but it does not *sound* odd to those taking part in it' (Ibid.). In addition, transcription is 'a filtering process' (Ochs 1979: 44) and no transcription, however fine grained, could ever be complete. A selection must unavoidably be made as to the type and amount of information to be included and as to the way this information should be displayed. A distinction is often made between two extreme types of transcriptions: 'broad' and 'narrow' or 'fine-textured' transcriptions (Edwards 1995: 20). While broad transcriptions provide only little information over and above the verbatim record of what is said, narrow or fine-textured transcriptions provide considerable detail regarding aspects such as voice quality, intonation, stress and other phonetic/phonemic details of pronunciation. The level of detail included is determined both by a series of linguistic-related factors such as who the transcription is designed for, what the researcher's purposes are or whether expert knowledge is available, and by non-linguistic factors such as time, corpus size and budget.

Although spoken learner corpora all have their origins in audio and/or video recordings of spoken productions by language learners, they can be seen to exhibit various degrees of 'spokenness' depending on the transcription conventions used and on the role played by the sound recordings after the transcription process. Corpora comprising broad orthographic transcriptions (e.g. the current version of *LINDSEI*) would, e.g., occupy positions far closer to the low spokenness end of the scale than corpora containing fine-textured transcriptions with detailed prosodic and/or phonetic information (e.g. the *ISLE* corpus, Atwell et al. 2003). In the same vein, corpora that offer researchers no access whatsoever to the original sound recordings and corpora with easy and ready access to such recordings would occupy very different positions on the continuum. Recent technological developments are increasingly making it possible for spoken learner corpora to reinstate the spoken medium and, as a result, to make their voice heard towards the high spokenness end of the continuum. Developments that have helped restore and enhance the spoken character of these 'new generation' spoken learner corpora by generally enabling researchers to play back the original source of the text that they can see on their screen include the alignment of digital audio files with the transcriptions (Pérez-Paredes 2003) and the advent of multimodal corpora. The integration of video recordings with the transcriptions in multimodal corpora also gives users access to the visual non-verbal communication aspects of spoken discourse (Reder et al. 2003; Braun 2007; Luzón et al. 2007).

9.2.1.2 Spoken learner corpora: learner and task variables

As highlighted by Granger (2004) and Nesselhauf (2005), to qualify as learner corpora electronic collections of texts produced by language learners have to be collected and compiled according to strict and explicit design criteria. This section examines some of the design criteria that are specific to computerized learner corpora; i.e. those relating to learner and task variables (Granger 1998, 2002).

9.2.1.3 Learner variables

The majority of the spoken learner corpora under investigation in this chapter contain spoken productions by young adult learners of English as a foreign language (e.g. *LINDSEI*, De Cock 2004; the *Gießen-Long Beach Chaplin Corpus*, henceforth *GLBCC*, Müller 2005). Notable exceptions include the *Multimedia Adult ESL Learner Corpus* (*MAELC*, Reder et al. 2003), a corpus of learners of English as a second language and the *Evaluation of English Corpus of Norwegian School English* (henceforth *EVA*; Hasselgren 2002), a corpus of secondary school pupils. A large proportion of the subjects who have contributed data to spoken learner corpora are students in higher education. The corpora tend to be made up of data from either university students of English language, literature and/or linguistics (e.g. *LINDSEI*; the *Spoken Corpus of Chinese Learners*, henceforth *SECCL*, Wen 2006), university students of subjects other than English (e.g. the *College Learners' Spoken English Corpus*, henceforth *COLSEC*, Wen 2006), or a combination of both (e.g. *GLBCC*).

With respect to the mother tongue backgrounds of the learners in the spoken learner corpora, a major distinction can be drawn between mono-L1 corpora and multi-L1 corpora. Mono-L1 corpora contain the productions of learners that share the same mother tongue (e.g. Japanese learners in *NICT JLE*; Chinese learners in *COLSEC*). Multi-L1 corpora, by contrast, are made up of a number of distinct components that cover learners from several mother tongue backgrounds. *LINDSEI*, e.g., currently contains data from EFL learners from 12 different mother tongues (i.e. Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Norwegian, Polish, Spanish and Swedish).

Regarding proficiency level, two main types of spoken learner corpora can be identified: corpora that contain spoken productions by learners of the same proficiency level (e.g. *LINDSEI*) and corpora that are made up of productions by learners of a variety of proficiency levels (e.g. *NICT JLE*). The level of the learners in a corpus like *LINDSEI* tends to be determined on the basis of external criteria: the non-native interviewees in the corpus are considered as 'advanced' provided they are third or fourth year university students of English. Only a small proportion of the corpora that include data from learners of various proficiency levels are genuine longitudinal corpora that contain productions by the same learners over a period of time (e.g. parts of *NICT JLE* and the corpus described in Czwenar 2004). The majority (e.g. *COLSEC*; the *PAROLE* corpus; Osborne 2007) contain data collected at a single point in time from different learners and are therefore quasi-longitudinal (Granger 2004).

9.2.1.4 Task variables

The types of task that the learners contributing to spoken learner corpora are requested to carry out are extremely varied and typically include informal interviews or discussions about the learners' personal lives (e.g. travel experience, university life, films, etc.; e.g. *LINDSEI*), role-plays (e.g. *NICT JLE, EVA*), picture descriptions (e.g. *PAROLE*) and oral narratives based on picture or video prompts (e.g. *GLBCC*). While some tasks are monologic in nature, the majority tend to involve some form of interaction between either two learners (e.g. *EVA*) or a learner and an interviewer (e.g. *LINDSEI*) for example. The relationships between the participants in tasks involving interaction can be extremely diverse (compare: two fellow-students engaged in a conversation vs. a young student learner interviewed by an examiner unknown to the him/her). Factors affecting the type of relationship between the participants in interactions include, among others, the age of the participants, the power distance between them or task settings.

It is not uncommon for spoken learner corpora to be made up of more than one type of task. For example, the interviews in the *LINDSEI* corpus are rounded off with a more controlled short picture-based story-telling activity and the *LEAP* corpus (www.phonetik.uni-freiburg.de/leap/) is made up of both highly controlled reading aloud tasks and tasks involving free speech.

Among the criteria that determine the task type(s) to be included in a corpus, the following can be listed: the general purpose of the project within the framework of which the corpus is collected, the format of the exams/language tests used as a basis for the corpus, the learners' learning context or the type of research the task type enables researchers to engage in. Informal interviews are, for instance, often collected in an attempt to gather data that would be as close as possible to informal conversation from learners in EFL settings, where English is rarely used in fully authentic non-artificial communications (Granger 2002). The more controlled picture-based story-telling activity in *LINDSEI* was included to allow for targeted NS-NNS comparisons of lexis (De Cock 2004).

Learners' contributions are collected in very diverse settings. Typical task settings include data being collected within the framework of formal exam or language test situations (e.g. *COLSEC, SECCL, NICT JLE, EVA*), classroom or language laboratory activities (e.g. Kindt and Wright 2001; Pérez-Paredes 2003) or activities in which the learners take part on a voluntary basis (outside the classroom and not in exam situations; e.g. *LINDSEI, GLBCC*).

Another task variable concerns whether or not the learners were granted any preparation time before performing certain tasks (e.g. *LINDSEI*, *COLSEC*). The learners in *LINDSEI* were given a few minutes to choose a topic and gather their thoughts about it just before the interview in an attempt to make them feel at ease.

9.2.2 Spoken learner corpus research

The lion's share of the growing body of research on spoken learner corpora appears to have centred around a number of aspects of what Biber et al. (1999) call the 'grammar of speech'. Discourse markers have attracted much of the attention (e.g. He and Xu 2003; Pulcini and Furiassi 2004; Müller 2005; Buysse 2007), as have fluency and performance phenomena such as filled and unfilled pauses (Czwenar 2004; Osborne 2007; Götz 2007). Some studies have focused on other phenomena that are specific to spoken language such as intonation (Ramírez and Romero 2005) or segmental errors (Cheng 2005). Other lines of research include lexis (Miliander 2003), phraseology (taken in a wide sense; De Cock 2004), grammar (Kaneko 2004) and the organization of spoken discourse (Chen 2004). The increased availability of spoken learner corpora for research has also made it possible for linguistics to embark on systematic comparisons of learner speech and learner writing (Abe 2003; De Cock 2003; Miliander 2003; Kaneko 2004). Research investigating the development of certain linguistic phenomena across proficiency levels that make use of genuine or quasi-longitudinal spoken learner corpora has started to emerge (Czwenar 2004; Osborne 2007; Tono 2007) and will grow with the compilation of bigger and better corpora of this type. A review of publications on spoken learner corpus research also reveals that the full potential of multi-L1 spoken learner corpora has yet to be exploited.

It is noteworthy that very few studies have been carried out using part of speech (POS) tagged versions of the spoken corpora (Tono 2007). The combined presence of downright errors, high numbers of hesitation items, false starts and repeats makes learner speech a challenge for POS taggers, which have been trained on the basis of native corpora. Studies based on error-tagged versions of spoken learner corpora are also few and far between (Abe 2003). Error-tagging written learner data is notoriously complex and time-consuming and it is not difficult to see how crucial it is for existing error tagging systems developed on the basis of learner writing to be adapted to accommodate the very nature of spoken learner language. Projects of this kind are currently in the pipeline (Mukherjee 2007). Error-tagging will undoubtedly benefit a great deal from the easy and ready access to the original sound recordings the new generation spoken learner corpora give or will soon give researchers. The benefits offered by new generation spoken learner corpora are clearly not limited to error-tagging. Not only will these corpora facilitate research into all aspects of spoken learner language, they will also without doubt open up new avenues of research.

9.3 Pedagogical Applications

According to Granger (2009) 'learner corpus research has not yet fully realized its stated ambition' (. . .) in that 'it has given rise to relatively few concrete pedagogical applications.' Granger's observation is especially true for spoken learner corpus research. This should in fact come as no surprise considering that corpora of native speech, which have been around for much longer than corpora of learner speech, have only recently started to make their way into teaching (Mauranen 2004) with recent pedagogical projects based on new generation spoken native corpora like the *ELISA* (Braun 2007) or the *SACODEYL* project (Alcaraz et al., this volume). It is also clear that far more analytic work based on spoken learner corpora is required before spoken learner corpus informed teaching materials can be made available.

This section explores the contribution of learner corpora to EFL teaching. The focus is on possible applications based on spoken learner corpora in the field of materials design and in the classroom.

9.3.1 Spoken learner corpora and materials design

Learner corpus-based research has a great deal to offer specialists engaged in the design of teaching and reference materials like monolingual learners' dictionaries, grammars or textbooks as it allows for a systematic account of learners' difficulties and needs. Thanks to methods of analysis such as Contrastive Interlanguage Analysis (CIA, see Granger 1996) and computer-aided error analysis (Dagneaux et al. 1998) learner corpus-based research makes it possible to expose patterns of misuse and over- or underuse. Studies that make use of multi-L1 learner corpora such as *LINDSEI* also enable researchers to uncover which problems tend to be shared by various groups of learners and which problems tend to be shared by the members of one specific group only. While the former would require treatment in reference materials aimed at all learners regardless of their mother tongue backgrounds, the latter, if clearly shown to be transfer-related, could receive treatment in reference materials aimed at learners from one specific L1.

The production of materials that address learners' attested needs typically makes use of what Granger (2009) refers to as learner corpora for delayed pedagogical use (DPU). Learner corpora for DPU 'are not used directly as teaching/learning materials by the learners who have produced the data' (Ibid.). They are collected and used by academics and publishers in order to provide better descriptions of interlanguage and/or to create learner corpus informed teaching materials.

Granger (2009) emphasizes the need for a great deal of careful analytic work before learner corpus informed teaching materials can be created. Great care must be taken when conducting learner corpus research because of the many variables involved. For example, when contrasting learner and NS speech to bring out possible patterns of overuse or underuse, it is essential to compare data from the same task type collected in the same setting as these variables have been shown to have a significant impact on the type of language that is used (De Cock 2002, 2003; Müller 2005; Luzón et al. 2007). Collecting fully comparable corpora containing data from native speakers performing the same task(s) as the learners in identical settings is one way of making sure researchers do not draw hasty and erroneous conclusions based on comparisons between apples and oranges. Fully comparable native speaker corpora have, e.g., been collected within the framework of the *GLBCC* and the *LINDSEI* project (the *Louvain Corpus of Native English Conversation,* De Cock 2004).

An illustration of the type of findings from spoken learner corpus research that would be particularly relevant when designing reference materials is learners' attested underuse of sentential relative clauses in informal contexts (De Cock 2003, 2007). As the following example from native speaker speech illustrates, sentential relative clauses can be seen to play an important role in informal native speech because they tend to have evaluative function displaying speakers' affective involvement with and attitudes to the events and experiences they are relating (see also Tao and McCarthy 2001):

> <B> they all come and visit me cos they think it's great having a student life so close to <X> so a lot of them travel up at weekends and that [*which is quite nice* <B> (*Louvain Corpus of Native English Conversation*)

Sentential relative clauses would in fact be a particularly good candidate for inclusion in a contextualized discourse-oriented grammar of speech. As well as providing learners with a wider range of ways of expressing attitudinal stance in interactions, giving more prominence to sentential relative clauses could also help learners cope with the pressures of online processing in unplanned spoken discourse. The use of sentential relative clauses is indeed consistent with the 'clause chaining style' or clause 'add-on strategy' that has been shown to be particularly well-suited to the constraints of real-time planning (Biber et al. 1999).

In addition to helping researchers identify potential candidates for inclusion and in-depth treatment in reference materials, spoken learner corpora can also serve as a testbed to assess whether existing teaching and/or reference materials are providing learners with the help they should provide. Possible starting points for investigating the content of teaching materials can for instance be found in De Cock (2002, 2004, 2007) and in Götz (2007).

De Cock (2002, 2004) has shown that advanced EFL learners tend to overuse and misuse the sequence *(yes/yeah) of course* in a way that may well make them sound rather over-emphatic and even impolite: the learners are reported to use the sequence to answer a request for information or to respond to an opinion expressed by another speaker. An analysis of the treatment of *of course* in recent editions of monolingual learners' dictionaries reveals that two of the major learners' dictionaries, namely the *Longman Dictionary of Contemporary English* (LDOCE 2001, LDOCE 2005 – Summers 2005) and the *Oxford Advanced Learners' Dictionary* (OALD 2000, OALD 2005 – Wehmeier 2005) actually address learners' inappropriate use of *(yes) of course* in such contexts in usage notes.

Several recent studies investigating contracted forms in advanced learner speech (De Cock 2007; Götz 2007) have highlighted learners' underuse and inappropriate use of these forms in informal speech. A possible follow-up to these studies would be to use the *Corpus of Textbook Material* (*TeMa*, Meunier and Gouverneur 2007), which contains over 700,000 words of upper-intermediate/advanced textbook material, to examine the extent to which contracted forms are included in listening comprehension activities and whether or not contracted forms are the focus of discussions and/or exercises in textbooks.

9.3.2 Spoken learner corpora in the classroom

Another possible contribution of learner corpora to ELT is as part of data-driven learning (DDL) activities in the classroom (Johns 1991). DDL activities involving learner corpus data are presented as particularly useful when attempting to raise learners' awareness of their own fossilized errors or persistent overuse of certain words or phrases (Granger and Tribble 1998; Granger 2002; Nesselhauf 2004). As pointed out by Granger (2002: 26), the use of learner corpus data in the classroom is however ‘a highly controversial issue’. Exposing learners to erroneous data from learner corpora can be regarded as highly dangerous in that it may well reinforce erroneous usage. Nesselhauf (2004) and Granger and Tribble (1998) believe this danger can be addressed by ensuring that the learners are presented with ample positive evidence and that the activity is followed by consolidating exercises.

Mukherjee and Rohrbach (2006) advocate the use of ‘local learner corpora’, also called ‘home-grown learner corpora’ (Kindt and Wright 2001) or ‘learner corpora for immediate pedagogical use’ (Granger 2009) when preparing DDL activities for the classroom. These corpora are ‘collected by teachers as part of their normal pedagogical classroom activities’ (Granger 2009) and the learners who produce the data are also the users of the data. Not only can the use of these corpora help increase learners’ motivation as they are working on their own productions, but the activities will also be relevant to the needs of the learners. That said, learners would probably also find activities based on learner corpora for DPU that contain data from learners who have similar profiles to theirs rather motivating. Mukherjee and Rohrbach (2006) argue that the use of local corpora has the added bonus of involving more teachers in corpus-based activities. This will however only be possible if teachers are given access to (learner) corpus training through in-service or teacher training (cf. Mauranen 2004).

As emphasized by Mukherjee (2009: 213), ‘it is neither desirable nor useful to establish a rigid dichotomy between good and correct usage in native data on the one hand and incorrect usage in learner output on the other.’ In other words, the part played by learner corpora needs not be limited to providing the teachers with negative evidence only. Using learner corpora as a source of positive evidence can lead to increased motivation as the focus is also on what the learners have already mastered and get right.

The use of spoken learner corpora in the classroom is still very much in its infancy. This section therefore mainly focuses on how these corpora could be integrated into the teaching of spoken English. Although learner corpus informed classroom activities can be created to focus on any number of linguistic phenomena such as lexico-grammatical patterns (e.g. verb or noun complementation), collocations (see Mukherjee 2009) or pairs of commonly confused words, I confine myself to examples of possible concrete applications that concentrate on aspects that are specific to spoken discourse.

Mukherjee (2009) illustrates how data from spoken learner corpora can be integrated into a two-step activity designed to improve advanced learners' spoken fluency. The activity centres around discourse markers as research has revealed (1) that there is a strong correlation between learners' discourse-markers competence and their overall fluency (Hasselgren 2002), and (2) that advanced learners tend to underuse them in informal speech (Müller 2005). The focus of the activity presented here is more specifically on the discourse marker *you know*. The aim of the first step consists in making learners aware of the natural use of the discourse marker *you know* in spontaneous spoken discourse. To this end they are presented with concordances of *you know* and are instructed to identify the typical uses of the discourse marker (e.g. 'can be used if you want to indicate that the next words are perhaps not very precise'). Mukherjee suggests following this first step by exercises that would help learners automatize the use of *you know*. Learners could be asked to explain how new and unfamiliar games or computer programmes work 'by using less precise vocabulary, which they should indicate by *you know* as an approximator'. It is unfortunately not clear whether this activity has been tested in a classroom situation. One concern might be that this activity may well lead to learners' overuse of *you know* without adequate supervision from the teacher.

Pérez-Paredes (2003) reports on learner corpus informed language laboratory activities (based on local corpora) in which the learners work on the basis of digital sound files collected within the framework of networked-based language teaching. These activities require teachers' prior digital bookmarking (a facility offered by some digital players) of points in the flow of discourse to highlight segmental pronunciation problems or faulty discourse organization. The learners are instructed to review the bookmarked passages and to relate them to the contents of the language course they are taking.

Annotating spoken learner corpora for a number of typical language functions and notions used in informal speech (Coccetta 2008) or for communication strategies could also provide teachers and learners with a useful starting point when concentrating on spoken English in the classroom.

9.4 Looking Ahead

These are truly exciting times for spoken learner corpora. Although a great deal of collection and analytic work is still needed before they can realize their full potential for ELT, all the ingredients are there to help them make their voice heard louder in the not too distant future: technological developments that will make it easier to collect and transcribe spoken data, new generation spoken corpora, increased activity in the field, and increased focus on speech in ELT.

References

Abe, M. (2003), ‘A corpus-based contrastive analysis of spoken and written learner corpora: The case of Japanese-speaking learners of English’, in Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds), *Proceedings of the Corpus Linguistics 2003 Conference (CL 2003). Technical Papers 16.* Lancaster University: University Centre for Computer Corpus Research on Language, pp. 1–9.

Alcaraz, J. M. Pérez-Paredes, P., Mercader, A. and Tornero, E. (this volume), ‘A generic tool for annotating TEI-compliant corpora’, in Campoy, M. C. Bellés-Fortuno B. and Gea-Valor, M. L. (eds) (2009), *Corpus-Based Approaches to English Language Teaching.* Series: Corpus and Discourse. London/New York: Continuum.

Atwell, E., Howarth, P. and Souter, C. (2003), ‘The ISLE Corpus: Italian and German spoken learners English’. *ICAME Journal,* 27, 5–18.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999), *Longman Grammar of Spoken and Written English.* London: Longman.

Braun, S. (2007), ‘Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora’. *ReCALL,* 19, (3): 307–328.

Buysse, L. (2007), ‘Discourse marker *so* in the English of Flemish university students’. *Belgian Journal of English Language and Literatures (BELL),* New Series 5, 79–95.

Chen, X. (2004), ‘Personal referential strategies in Chinese EFL learners’ oral narratives, in He, A., Wang, L., Xu, M. and Chen, X. (eds), *Application of Corpora to Foreign Language Education: Theory and Practice.* Guangzhou: Guangdong Higher Education Press, pp. 351–365.

Cheng, Chunmei (2005), ‘Segmental errors in Chinese learners’ oral English’, in He, A., He, G., Chen, X., Wang, L., Xu, M. and Cheng, C. (eds), *Research and Application of Corpus Linguistics.* Changchun: Northeast Normal University Press, pp. 117–120.

Coccetta, F. (2008), ‘Multimodal functional-notional concordancing’ in Frankenberg-Garcia, A. et al. (eds), *8th Teaching and Language Corpora Conference.* Lisboa: Associação de Estudos de Investigação Científica do ISLA-Lisboa, pp. 72–79.

Czwenar, I. (2004), ‘Oral proficiency of Polish EFL students. Corpus-based analysis’, in Lewandowska-Tomaszczyk, B. (ed.), *Practical Applications in Language and Computers (PALC 2003).* Frankfurt: Peter Lang, pp. 391–399.

Dagneaux, E., Denness, S. and Granger, S. (1998), ‘Computer-aided Error Analysis’. *System,* 26, (2), 163–174.

De Cock, S. (2002), ‘Pragmatic prefabs in learners’ dictionaries’, in Braasch, A. and Povlsen, C. (eds), *Proceedings of the Tenth EURALEX International Congress, EURALEX*

2002, (vol II). Copenhagen, Denmark, August 13–17, 2002. Copenhagen: Center for Sprogteknologi, pp. 471–781.

—(2003), 'Recurrent sequences of words in native speaker and advanced learner spoken and written English'. Unpublished Ph.D. thesis, Université catholique de Louvain.

—(2004), 'Preferred sequences of words in NS and NNS speech'. *Belgian Journal of English Language and Literatures (BELL), New Series*, 2, 225–246.

—(2007), 'Routinized building blocks in native speaker and learner speech: Clausal sequences in the spotlight', in Campoy, M. C. and Luzón, M. J. (eds), *Spoken Corpora in Applied Linguistics*. Bern: Peter Lang, pp. 217–233.

Edwards, J. (1995), 'Principles and alternative systems in the transcription, coding and mark-up of spoken discourse', in Leech G., Meyers, G. and Thomas, J. (eds), *Spoken English on Computer. Transcription, Mark-up and Application*. London: Longman, pp. 19–34.

Götz, S. (2007), 'Performanzphänomene in gesprochenem Lernerenglisch: eine korpusbasierte Pilotstudie' [Performance phenomena in spoken learner English: A corpus-based pilot study]. *Zeitschrift für Fremdsprachenforschung*, 18, (1), 67–84.

Granger, S. (1996), 'From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora', in Aijmer, K., Altenberg, B. and Johansson, M. (eds), *Languages in Contrast. Text-Based cross-linguistic studies. Lund Studies in English 88*. Lund: Lund University Press, pp. 37–51.

—(1998), 'The computerized learner corpus: A versatile new source of data for SLA research', in Granger, S. (ed.), *Learner English on Computer*. London and New York: Addison Wesley Longman, pp. 3–18.

—(2002), 'A Bird's-eye view of computer learner corpus research', in Granger, S., Hung, J. and Petch-Tyson, S. (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Language Learning and Language Teaching 6. Amsterdam and Philadelphia: Benjamins, pp. 3–33.

—(2004), 'Computer learner corpus research: Current status and future prospects', in Connor, U. and Upton, T. (eds), *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam and Atlanta: Rodopi, pp. 123–145.

—(2009), 'The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation', in Aijmer, K. (ed.), *Corpora and Language Teaching*. Amsterdam: John Benjamins. 13–32.

Granger, S. and Tribble, C. (1998), 'Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning', in Granger, S. (ed.), *Learner English on Computer*. Addison Wesley Longman: London and New York, pp. 199–209.

Granger, S., Dagneaux, E. and Meunier, F. (2002), *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Hasselgren, A. (2002), 'Testing learner fluency: the role of "small –words"', in Granger, S., Hung, J. and Petch-Tyson, S. (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Language Learning and Language Teaching 6. Amsterdam and Philadelphia: Benjamins, pp. 143–174.

He, A. and Xu, M. (2003), 'Small words in Chinese EFL learners' Spoken English'. *Foreign Language Teaching and Research*, 35, (6), 446–453.

Johns, T. (1991), 'Should you be persuaded – two examples of data-driven learning materials'. *English Language Research Journal*, 4, 1–16.

Kaneko, T. (2004), 'The use of past tense forms by Japanese learners of English', in Nakamura, J., Inoue, N. and Tabata, T. (eds), *English Corpora under Japanese Eyes*. Amsterdam: Rodopi, pp. 215–228.

Kindt, D. and Wright, M. (2001), *Integrating Language Learning and Teaching with the Construction of Computer Learner Corpora*. Retrieved May 22, 2003, from http://www.nufs.ac.jp/~kindt/media/corpora.pdf

Leech, G., Meyers, G. and Thomas, J. (1995), 'Editors' general introduction', in Leech, G., Meyers, G. and Thomas, J. (eds), *Spoken English on Computer. Transcription, Mark-up and Application*. London: Longman, pp.1–11.

Luzón, M. J., Campoy, M. C., Sánchez, M. M. and Salazar, P. (2007), 'Spoken corpora: New perspectives in oral language use and teaching', in Campoy, M. C. and Luzón, M. J (eds), *Spoken Corpora in Applied Linguistics* . Bern: Peter Lang, pp. 3–30.

Mauranen, A. (2004), 'Spoken corpus for an ordinary learner', in Sinclair J. (ed.), *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: Benjamins, pp. 89–105.

Meunier, F. and Gouverneur, C. (2007), 'The treatment of phraseology in ELT textbooks', in Hidalgo, E., Quereda, L. and Santana, J. (eds), *Corpora in the Foreign Language Classroom. Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC6)*, University of Granada , Spain , 4–7 July, 2004 . [Language and Computers Series 61]. Amsterdam/New York: Rodopi, 119–139.

Miliander, J. (2003), *We Get the Answer We Deserve. A Study of Vocabulary in a Spoken and Written Corpus of Advanced Learner English*. Karlstad University: Karlstad University Studies 2003: 16.

Mukherjee, J. (2007), 'Exploring and annotating a spoken English learner corpus: A work-in-progress report', in Volk-Birke, S. and Lippert, J. (eds), *Anglistentag 2006 Halle: Proceedings*. Trier: WVT, pp. 365–375.

—(2009), 'The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications', in Aijmer, K. (ed.), *Corpora and Language Teaching*. Amsterdam: John Benjamins. 203–230.

Mukherjee, J. and Rohrbach, J. (2006), 'Rethinking Applied Corpus Linguistics from a Language-Pedagogical Perspective: New Departures in Learner Corpus Research', in Ketteman, B. and Marko, G. (eds), *Planing, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop*. Frankfurt am Main: Peter Lang. 205–232.

Granger, S. and Tribble, C. (1998), 'Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning', in Granger, S. (ed.), *Learner English on Computer*. London and New York: Addison Wesley Longman, pp. 199–209.

Müller, S. (2005), *Discourse Markers in Native and Non-native English Discourse*. Amsterdam/Philadelphia: John Benjamins.

Nesselhauf, N. (2004), 'Learner Corpora and their potential in language teaching', in Sinclair, J. (ed.), *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: Benjamins, pp. 125–152.

—(2005), *Collocations in a Learner Corpus*. Amsterdam: Benjamins.

Ochs, E. (1979), 'Transcription as theory', in Ochs, E. and Schieffelin, B. (eds), *Developmental Pragmatics*. New York, San Francisco and London: Academic Press, pp. 43–72.

Osborne, J. (2007), 'Investigating L2 Fluency through oral learner corpora', in Campoy, M. C. and Luzón, M. J. (eds), *Spoken Corpora in Applied Linguistics*. Bern: Peter Lang, pp. 181–197.

Pérez-Paredes, P. (2003), 'Integrating networked learner oral corpora into foreign language instruction', in Granger, S. and Petch-Tyson, S. (eds), *Extending the Scope of Corpus-based Research: New Applications, New Challenges*. Amsterdam and Atlanta: Rodopi, pp. 249–261.

Pravec, N. (2002), 'Survey of learner corpora'. *ICAME Journal,* 26, 81–114.

Pulcini, V. and Furiassi, C. (2004), 'Spoken interaction and discourse markers in a corpus of learner English', in Partington, A., Morley, J. and Haarman, L. (eds), *Corpora and Discourse*. Bern: Peter Lang, pp. 107–123.

Ramírez, M. D. and Romero J. (2005), 'The pragmatic function of intonation in L2 discourse: English tag questions used by Spanish speakers'. *Intercultural Pragmatics*, 2, 151–168.

Reder, S., Harris, K. and Setzler, K. (2003), 'The multimedia adult ESL learner corpus'. *TESOL Quarterly*, 37, (3), 546–557.

Stubbs, M. (1983), *Discourse Analysis. The Sociololinguistic Analysis of Natural Language*. Oxford: Basil Blackwell.

Summers, D. (ed.) (2005), *Longman Dictionary of Contemporary English*. Pearson Education Limited: Harlow.

Tao, H. and McCarthy, M. J. (2001), 'Understanding non-restrictive *which*-clauses in spoken English, which is not an easy thing', *Language Sciences*, 23, 651–677.

Tono, Y. (2007), 'The roles of oral L2 learner corpora in language teaching: The case of the NICT JLE corpus', in Campoy, M. C. and Luzón, M. J. (eds), *Spoken Corpora in Applied Linguistics*. Bern: Peter Lang, pp. 163–179.

Wehmeier, S. (ed) (2005), *Oxford Advanced Learner's Dictionary*. Oxford University Press, Oxford.

Wen, Q. (2006), *Chinese learner corpora and second language research*. Paper presented at the 2006 International Symposium of Computer-Assisted Language Learning, June 2–4, 2006, Beijing.– unpublished but available from http://call2006.fltrp.com/PPT/Keynote/Wen%20Qiufang.ppt

Chapter 10

Designing and Exploiting a Small Online English-Spanish Parallel Textual Database for Language Teaching Purposes

Julia Lavid, Jorge Arús Hita and Juan Rafael Zamorano-Mansilla
Universidad Complutense de Madrid

10.1 Introduction

Twenty years ago Leech and Candlin advocated classroom access to 'language databases, lexicographic and grammatical corpora, oriented towards learners' interlanguages and displayed in terms that learners can understand' (1986: xvi). Today many dictionaries, several grammars and a growing number of EFL courses and teaching materials proclaim themselves to be 'corpus-based'. Indeed, the availability of large corpora and the enormous potential they offer for empirical linguistic research has meant a revolution for linguistic studies in the information age.

However, in spite of the growing number of studies advocating the use of corpora in language teaching (Burnard and McEnery 2000; Sinclair 2004; *inter alia*), relatively few attempts have been made to use corpora directly in the classroom. One of the reasons is the confusion over the distinction between what is 'scientifically interesting' and what is 'pedagogically useful' (Kennedy 1992: 364–367). The developers of large corpora were not concerned with language pedagogy when compiling corpora, and this fact has consequences for the learning context. While a good number of relevant studies have acknowledged the potential of large corpora as useful tools for classroom activities (see Bernardini 2000; Lavid 2007b; *inter alia*), many of those experiences have also reported teachers' and learners' problems when working with large corpora.

In this chapter, we report on a current effort at creating and exploiting a small bi-directional English-Spanish textual database for a variety of linguistic tasks in a mixed learning scenario at Universidad Complutense de Madrid (UCM).[1]

The chapter is organized as follows: section 10.2 presents some preliminary background issues to the work reported in the rest of the chapter. Section 10.3 focuses on design criteria used in the compilation of the database, and explains some issues concerned with its online access and management. Section 10.4 outlines some examples of exploitation activities and section 10.5 provides a summary and some concluding remarks.

10.2 Background Issues

The students which formed the basis for the development of the bilingual textual database are registered in the area of English language and linguistics, applied and contrastive linguistics (English-Spanish) and translation at the UCM. While all of them are computer-literate and frequent internet users, none of them – except those attending previous courses taught by the authors of this chapter – has had previous contact with computer corpora or corpus analysis tools. Moreover, their exposure to authentic language materials in the subjects mentioned above is limited and very much dependent on the preferences and expertise of the instructor. Therefore, the creation of an online textual database which can be consulted from the Virtual Campus was felt as a much needed tool to extend and enrich their learning environment.

The needs and backgrounds of such a heterogeneous group of students are varied. Those students registered in the studies programme of English Philology take courses on English Syntax, Semantics, Pragmatics and Discourse Analysis as compulsory subjects, while Contrastive Linguistics (English-Spanish) is an optional course in this programme. All of them are native speakers of Spanish and have an advanced level of proficiency in English. Those students registered in the Master on Translation take different courses on translation theory and practice, and Contrastive Linguistics (English-Spanish) is a compulsory course. These students have different linguistic and educational backgrounds and, except for those with a degree in Linguistics or Philology, only have a basic knowledge of linguistic theory.

Even though the needs of these two groups might not be exactly the same, it is clear that an initial mixture of everyday genres could be a starting point for the compilation of the bilingual database. On the basis of this general principle, a bi-directional (English-Spanish) textual database is being compiled using a series of design criteria, as described in the following section.

10.3 Designing and Compiling the Textual Database

A series of design criteria have been considered in the compilation of the bilingual database. The first criterion is its bi-directional, also called ‘reciprocal’, character. Thus, the database consists of original English texts and their translations into Spanish, and of original Spanish texts and their translations into English, as illustrated in Figure 10.1.

This design allows to meet the needs of a variety of learners and carry out linguistic comparisons on a number of different dimensions. One can use the database to compare original texts in both languages, or original and translated texts in both languages or in the same language, or compare translated texts in the two languages, to reveal general features of translations. It must be pointed out that even though a conscious effort is made to achieve a balance with respect to the direction of the translations, in many genres one direction (English to Spanish) tends to dominate over the other. Thus, it is common to find translations into English of Spanish academic article abstracts, while the reverse is not the case.

Another design criterion is genre variation. Initially we have included a variety of non-specialized genres, both monologic and dialogic, which may be of interest to a variety of learners at our University. Among the monologic texts, the database includes short stories, editorials, tourist descriptions, academic article abstracts and scientific essays. The dialogic variety currently includes interviews and parliamentary debates. As explained before, the compilation is limited by the availability of comparable genres which have been translated in both directions. Whereas comparable translations from English into Spanish are varied and numerous, those in the opposite direction are limited in type and size.

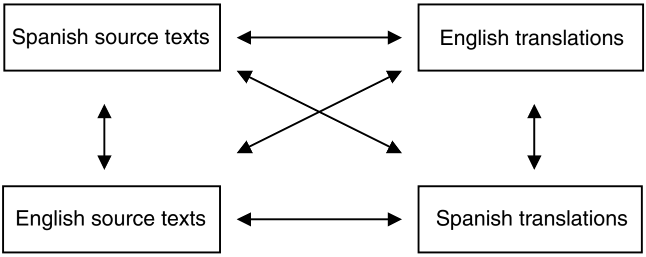
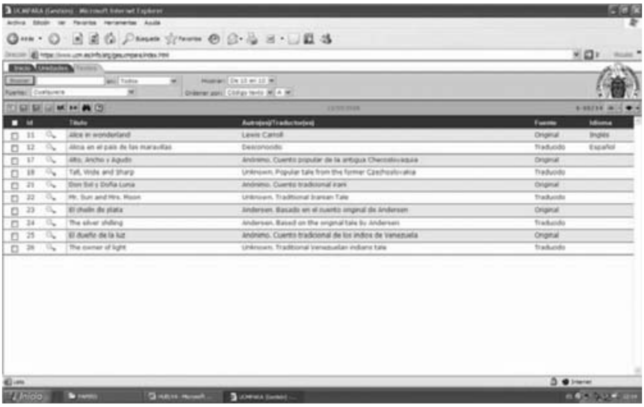


FIGURE 10.1 Architecture of the bilingual database

As to the current size of the database, we follow Douglas Biber (1990, 1993) who shows that smaller corpora are capable of covering all linguistic features of a given register. His calculations, i.e. ten texts per register/genre with a length of 1,000 words, serve as an orientation for the size of our current core corpus. However, the database will keep growing as more texts in both directions of translation become available, but also when the need arises to work with specific corpora from a given genre. For example, suppose that we are interested in analysing and comparing the linguistic features of letters of application in both languages. It would be possible to compile such a genre-specific corpus from different sources and then add it to the database. Search facilities in the current interface design will allow users to download and search only the (sub)corpus the user is interested in. Likewise, it is also possible for users to extract specific subsets of texts from the database to compile specific (sub)corpora. For example, it is possible to extract a small corpus of editorials, or of academic article abstracts. Figure 10.2 below presents a partial view of the online database, more particularly, the one corresponding to the subcorpus of short stories.

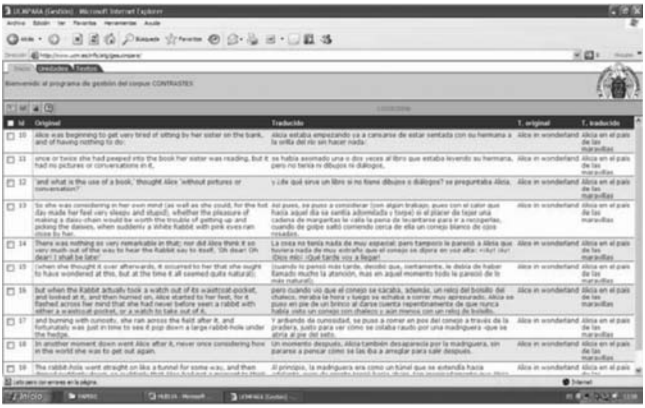
As shown in Figure 10.2, the first column displays the title of the text, the second column shows the name of the author(s) and of the translator(s), when available, the third column shows the direction of the translation (original or translated text), and the fourth column shows the language in which the text is written.

Texts stored in the database are aligned to allow column display on the web-based computer application, as shown in Figure 10.3.



Id	Título	Autor(es)/Traductor(es)	Fuente	Idioma
11	Alice in wonderland	Lewis Carroll	Original	Inglés
12	Alicia en el país de las maravillas	Desconocido	Traducido	Español
17	Alto, Ancho y Agudo	Anónimo. Cuento popular de la antigua Checoslovaquia	Original	
18	Tall, Wide and Sharp	Unknown. Popular tale from the former Czechoslovakia	Traducido	
21	Don Sol y Doña Luna	Anónimo. Cuento tradicional iraní	Original	
22	Mr. Sun and Mrs. Moon	Unknown. Traditional Iranian Tale	Traducido	
23	El chelín de plata	Andersen. Basado en el cuento original de Andersen	Original	
24	The silver shilling	Andersen. Based on the original tale by Andersen	Traducido	
25	El dueño de la luz	Anónimo. Cuento tradicional de los indios de Venezuela	Original	
26	The owner of light	Unknown. Traditional Venezuelan indians tale	Traducido	

Figure 10.2 Partial view of the text database


Bienvenido al programa de gestión del corpus CONTRASTES
Inicio
Unidades
Textos
Listo pero con errores en la página
Internet
Inicio

■	Id	Original	Traducido	T. original	T. traducido
□	10	Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:	Alicia estaba empezando ya a cansarse de estar sentada con su hermana a la orilla del río sin hacer nada:	Alice in wonderland	Alicia en el país de las maravillas
□	11	once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it,	se había asomado una o dos veces al libro que estaba leyendo su hermana, pero no tenía ni dibujos ni diálogos.	Alice in wonderland	Alicia en el país de las maravillas
□	12	'and what is the use of a book,' thought Alice 'without pictures or conversation?'	y ¿de qué sirve un libro si no tiene dibujos o diálogos? se preguntaba Alicia.	Alice in wonderland	Alicia en el país de las maravillas
□	13	So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.	Así pues, se puso a considerar (con algún trabajo, pues con el calor que hacía aquel día se sentía adormilada y torpe) si el placer de tejer una cadena de margaritas le valía la pena de levantarse para ir a recogerlas, cuando de golpe saltó corriendo cerca de ella un conejo blanco de ojos rosados.	Alice in wonderland	Alicia en el país de las maravillas
□	14	There was nothing so very remarkable in that; nor did Alice think it so very much out of the way to hear the Rabbit say to itself, 'Oh dear! Oh dear! I shall be late!'	La cosa no tenía nada de muy especial; pero tampoco le pareció a Alicia que hubiera nada de muy extraño que el conejo se dijera en voz alta: «¡Ay! ¡Ay! ¡Dios mío! ¡Qué tarde voy a llegar!	Alice in wonderland	Alicia en el país de las maravillas
□	15	(when she thought it over afterwards, it occurred to her that she ought to have wondered at this, but at the time it all seemed quite natural);	(cuando lo pensó más tarde, decidió que, ciertamente, le debía de haber llamado mucho la atención, mas en aquel momento todo le pareció de lo más natural);	Alice in wonderland	Alicia en el país de las maravillas
□	16	but when the Rabbit actually took a watch out of its waistcoat-pocket, and looked at it, and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it,	pero cuando vio que el conejo se sacaba, además, un reloj del bolsillo del chaleco, miraba la hora y luego se echaba a correr muy apresurado, Alicia se puso en pie de un brinco al darse cuenta repentinamente de que nunca había visto un conejo con chaleco y aún menos con un reloj de bolsillo.	Alice in wonderland	Alicia en el país de las maravillas
□	17	and burning with curiosity, she ran across the field after it, and fortunately was just in time to see it pop down a large rabbit-hole under the hedge.	Y ardiendo de curiosidad, se puso a correr en pos del conejo a través de la pradera, justo para ver cómo se colaba raudo por una madriguera -que se abría al pie del seto.	Alice in wonderland	Alicia en el país de las maravillas
□	18	In another moment down went Alice after it, never once considering how in the world she was to get out again.	Un momento después, Alicia también desaparecía por la madriguera, sin pararse a pensar cómo se las iba a arreglar para salir después.	Alice in wonderland	Alicia en el país de las maravillas
□	19	The rabbit-hole went straight on like a tunnel for some way, and then	Al principio, la madriguera era como un túnel que se extendía hacia	Alice in wonderland	Alicia en el país

FIGURE 10.3 Web display of translation units

The current web-based design allows for different types of searches once the user has accessed the online database which is password protected.[2]

10.4 Exploiting the Textual Database

A number of exploitation activities have been implemented in the context of several of the courses mentioned above. Some of them focused on the analysis of the individual texts in the corpus and on 'whole-corpus reading' (Henry and Roseberry 2001), thus becoming the source for teaching materials. Other activities were based on direct access to the database by students. In the following subsections we describe some examples of these activities.

10.4.1 Some examples of exploitation activities

Example 1: Aspectual distinctions in English and Spanish

Purpose of the activity

This activity was planned in the context of the Contrastive Linguistics course and was designed to contrast aspectual distinctions about past state of affairs in English and Spanish. The purpose of the activity was to make students apply theoretical notions learnt in class about the different types of *Aktionsart,* how they interact with various grammatical aspects in a past-time environment and its influence on the selection of tenses in English and Spanish.

For this activity one single text was chosen: a fragment of over 1,000 from Robert Graves' *I, Claudius* and its translation into Spanish. The main reason for selecting this text was that it contained a good number of Past tenses with different translations into Spanish, which made it particularly suitable for the purposes of the activity. The activity consisted of several phases:

Phase 1

The students received a questionnaire that they should apply to each verb form in the Past tense in the text. The questionnaire consisted of the following questions:

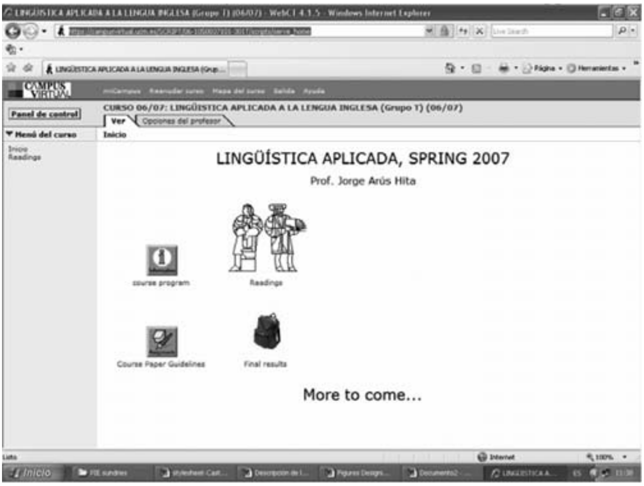
1. Which tense would you use to translate the verb?
2. What is the *Akstionsart* of the state of affairs? How do you know?
3. What is the grammatical aspect of the verb? How do you know?
4. Is the subject definite or indefinite?
5. Is the complement definite or indefinite?
6. Are there time circumstantials? If so, of what type?
7. Is there a match between your translation and the one provided by the corpus? If not, which one is more accurate? Why?

The questions were aimed at helping the students identifying the factors potentially responsible for the translator's choice, such as the stativity and telicity of the situation, the presence of definite or indefinite subjects or direct objects, pragmatics factors or knowledge of the world. The comparison between the translation proposed by the students and that provided by the corpus also raised their awareness about the semantic contrast between tenses in Spanish.

Phase 2

After filling in the questionnaire, the students shared the results, discussed the discrepancies and were asked to write a final report with the conclusions that could be drawn about the factors responsible for tense selection in English and Spanish.

The next two activities involved direct access to the database by students. Figure 10.4 below shows a partial view of the Applied Linguistics site on the UCM Virtual Campus, with a direct link to the activity that is described next.



LINGÜÍSTICA APLICADA A LA LENGUA INGLESA (Grupo T) (06/07) - WebCT 4.1.5 - Windows Internet Explorer

CURSO 06/07: LINGÜÍSTICA APLICADA A LA LENGUA INGLESA (Grupo T) (06/07)

LINGÜÍSTICA APLICADA, SPRING 2007

Prof. Jorge Arús Hita

More to come...

FIGURE 10.4 Partial view of the online Applied Linguistics course

Example 2: Practice in contrastive discourse analysis

Purpose of the activity

This activity, involving 28 students, was carried out in the context of the Applied Linguistics course as part of the curriculum of the degree in English Studies. It allowed practice in a number of applied areas in linguistics (e.g. corpus linguistics, contrastive linguistics, translation and discourse analysis), as will be explained in the description of the activity. This activity was compulsory as it represented the two credits corresponding to the Academic Activities of the course. The activity consisted of different phases:

Phase 1

Students were asked to choose one original text from the corpus and carry out a contrastive analysis with its corresponding translation in terms of either of the following: context analysis, topicality and thematic progression, rhetorical structures analysis or genre analysis.

Phase 2

Regarding context analysis, students had to trace how the contextual variables of field, tenor and mode were linguistically realized in English and Spanish. For topicality and thematic progression, their task was to compare different patterns of topical thematization for the same clauses in each language, e.g. whether both languages thematized the same or different experiential constituents. Those choosing rhetorical structure analysis would have to compare the rhetorical devices employed by each language to arrange the clause constituents in terms of the Theme/Rheme and Given/New structures. In the case of genre analysis – the most challenging choice – students were supposed to compare both texts in terms of their register configuration, schematic structure and realizational patterns.

Phase 3

They were to write a 3,000 word paper which would be submitted to their instructor via email by a given date. Before the paper submission, the students had to give a 15-minute oral presentation where they explained their most relevant findings.

Example 3: Testing students on semantic analysis through the use of new technologies

Purpose of the activity

This activity, involving 30 students of a course on English Semantics, represented an innovative type of take-home final exam for the subject. Using the date officially assigned to the final exam as deadline for its submission, the exam was a take-home in the sense that students did not have to sit in a classroom to take it; it was innovative because it involved the use of new technologies for its completion. The activity consisted of several phases.

Phase 1

Students were instructed in this case to work exclusively with the English texts from the corpus, following the instruction given in the exam for the English Semantics course.

Phase 2

The exams were submitted, as an attachment, to the instructor by the due date.

Phase 3

The instructor corrected them on the computer, using blue and red to highlight slight and big mistakes, respectively.

Phase 4

After the correction, students had their exams returned to them, again via email, with the corresponding grade.

10.4.2 Evaluation of the activities

For the evaluation of the tasks, students were requested to fill in an anonymous questionnaire. The feedback on the activities was highly satisfactory in general. The immense majority acknowledged that the activities had been quite stimulating and expressed the wish that this kind of practice be extended to other subjects. Students were particularly enthusiastic about the exam activity. Among the reported advantages of this type of examination, perhaps the most outstanding is the lowering of anxiety as opposed to traditional exams and the 'cleanliness' of the method, environmentally speaking, since no paper whatsoever was used in the whole process. On the negative side, a few students found that their limited computer-skills negatively affected their performance.

Regarding activities 2 and 3, i.e. those involving direct access to the database by students, a good deal of students admitted that these turned out to be harder than they had originally expected. The students' tasks were in principle facilitated by the fact that they 'simply' had to apply the theory and analyses explained in class to the texts they chose. However, linguistic analysis based on exposure to whole texts from the database turned out to be harder than the use of selected extracts by the instructor to illustrate the points under consideration. For many of them this was their first exposure to linguistic analysis of authentic language data. The difficulty in applying theory to real corpus data also fostered discussion among students and emphasized the sense that there was not a single correct answer or analysis, as every theoretical framework has problematic cases which cannot be handled without debate.

10.5 Summary and Concluding Remarks

The main motivation for the work reported in this chapter is the distinction between what is 'scientifically interesting' from what is 'pedagogically useful' when using corpora in the context of language learning. In an attempt to create a pedagogically useful resource at our university, we have compiled a web-based bilingual resource which may be useful for a variety of learners of English linguistics, Contrastive Linguistics and Translation studies. The main criteria for building such a resource are its availability through the Virtual Campus and its opportunistic character, including an initial mixture of genres which might be interesting for a variety of learners' needs.

Current exploitation activities focus either on the analysis of the individual texts in the corpus and on 'whole-corpus reading', while others are based on direct access to the database by students. Future work will concentrate on extending the range of genres and number of texts of the database, and on the design of specific exploitation activities tailored to the students' needs.

Notes

[1] The work reported in this paper is part of a larger project on "Corpus Linguistics and Contrastive Online Learning (English-Spanish)" (930175), led by Dr Julia Lavid as principal investigator and financed by the Programa de Creación y Consolidación de Grupos de Investigación Universidad Complutense-Comunidad de Madrid. We gratefully acknowledge the support provided by these authorities.

[2] For a more detailed description of the tasks involved to manage the database efficiently through the UCM Virtual Campus, see Lavid (2007a).

References

Bernardini, S. (2000), 'Systematising serendipity: Proposals for concordancing large corpora with language learners', In Burnard, L. and McEnery, T. (eds), *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt: Peter Lang Verlag, pp. 225–234.

Biber, D. (1990), 'Methodological issues regarding corpus-based analyses of linguistic variation'. *Literary and Linguistic Computing,* 5, (3), 257–269.

—(1993), 'Representativeness in Corpus Design'. *Literary and Linguistic Computing,* 8, (4), 243–257.

Burnard, L. and McEnery, T. (eds), (2000), *Rethinking language pedagogy from a Corpus Perspective.* Frankfurt: Peter Lang Verlag.

Henry, A. and Roseberry, R. (2001), 'Using a small corpus to obtain data for teaching a genre', in Ghadessy, M. Henry, A. and Roseberry, R. (eds), *Small Corpus Studies and ELT: Theory and Practice.* Amsterdam: John Benjamins, 93–133.

Kennedy, G. (1992), 'Preferred ways of putting things with implications for language Teaching', in Svartvik, J. (ed.), *Directions in Corpus Linguistics.* Berlin: Mouton de Gruyter, 335–373.

Lavid, J. (2007a), 'Contrastes: An online English-Spanish database for contrastive and translation learning'. *Practical Applications in Language and Computers 2007.* Frankfurt am Main: Peter Lang Publishing Group, pp. 431–443.

—(2007b), 'Contrastive patterns of mental transitivity in English and Spanish: A student-centred corpus-based study', in Quereda, L., Hidalgo, E. and Santana, J. (eds), *Corpora in the Foreign Language Classroom.* Amsterdam: Rodopi, 237–252.

Leech, G. and Candlin C. N. (1986), 'Introduction', in Leech, G. and Candlin, C. N. (eds), *Computers in English Language Teaching and Research.* London: Longman, xi–xvii.

Sinclair, J. McH. (ed.) (2004), *How to Use Corpora in Language Teaching.* Amsterdam: Benjamins.

Chapter 11

L2 Spanish Acquisition of English Phrasal Verbs: A Cognitive Linguistic Analysis of L1 Influence

Rafael Alejo González
University of Extremadura

11.1 Introduction

Almost any teacher with experience in the area of English as a Second Language (ESL) will point out Phrasal Verbs (PVs) as a source of difficulty for their students. To bear witness to this fact we only have to look at the shelves of an ESL specialized library. The amount of material published on PVs, both in the form of dictionaries and workbooks, is phenomenal.

However, not all teachers would put PVs at the top of their list of difficulties, and equally not all students experience the same level of difficulty in tackling these verbs. That is, PVs are acknowledged as a source of trouble but they may affect students in different ways, the L1 of the students being of the most likely explanation.

The scant research on the subject, which we shall see later on, has corroborated these impressions and has shown that Swedish and Dutch students experience a lesser degree of difficulty, while students whose L1 is a Romance language (e.g. Italian), may have more trouble with these verbs. In other words, language transfer may well be the main factor accounting for the problems experienced by students (Odlin 1989). Despite being ignored by some researchers in Second Language Acquisition (cf. Dulay and Burt 1974 or Krashen 1981, as cited in Jarvis and Pavlenko 2008: 100), it seems that transfer, which has now been purified and adapted to the new findings in the field, has become respectable again.

Identifying, in a general way, the origin of the problem, however, may not be enough to help students. L1 transfer merely points to the dissimilarity between the source and the target language, in this case between those languages which contain two- or three-word verbs and those languages which do not, but it does not identify the nature of the problems students face when acquiring PVs and to what extent some verbs may be more difficult to learn than others. Furthermore, a lexical approach in which general reference to opacity in meaning is made does not help either. Difficulty and opacity may be considered as equivalent terms and do not provide an explanation.

In this chapter, an attempt is made to identify the specific source of difficulty that PVs pose for learners. Thus, the key element of the problem will be located within one of the words that constitute the PV: the particle. Particles are highly frequent, non-salient polysemous words and as such they are, like other words with the same features, very difficult to learn. They apparently provide redundant information and are difficult to notice by learners whose first languages have not trained them to pay attention to this specific linguistic element. In short, as established by the Associative-Cognitive Creed (Ellis 2007: 84), they constitute typical elements that are 'blocked' by the learners' first language experience.

The organization of the chapter is as follows. After giving a broad linguistic definition of PVs, so that the limits of the phenomenon studied become clear, I make a brief summary of the main findings in the literature on the subject and identify how progress can be made from there. Then, I introduce the methodology used in the study and I finally present the results and the discussion of the analysis carried out.

11.2 Definition of PV

The concept of Phrasal Verb has traditionally been used in language teaching to refer to those verbs that are made up of two (e.g. *look up*) or three words (e.g. *look forward to*). The tendency has mainly been to consider them as vocabulary items that needed to be learned as a whole because their meaning was sometimes difficult to be inferred from their constituent parts. No specific linguistic criterion is adopted and the defining trait in this context is related to their formulaic and opaque nature, something which makes them closer to idioms.

Within the linguistics literature, however, a different terminology is adopted. Thus, Quirk et al. (1985) talk about 'Multi-word Verbs' to refer to the same group of verbs that language teaching texts and dictionaries refer to as 'Phrasal Verbs' and, what is more important, this reference grammar uses the term 'Phrasal Verb' to refer to a subgroup of 'Multi-word Verbs'. The other two groups would be 'Prepositional Verbs' (PRVs) and 'Phrasal Prepositional Verbs' (PPVs).

The reason for this classification is mainly syntactic. Even though apparently similar, PVs are used in constructions which the other groups do not licence. Thus, most PVs allow for a change in the position of the direct object complement, whereas PRVs and PPVs do not. This makes 2b unacceptable in the following examples:

1a *I'll have to look up his records at least*
1b *I'll have to look them up at least*
2a *I'm looking for a biography on Lincoln*
2b **I'm looking it for*

There are other tests and, even though it is true that not all of them work equally well, they can serve to make this distinction appropriate from a grammatical point of view (see O'Dowd 1998; Cappelle 2005 for a full account). I will therefore adhere to this restrictive definition of PVs, since I consider that some of their syntactic and phonological peculiarities may have an influence on their acquisition.

11.3 Cognitive Linguistics Approach to PVs

The acceptance of the linguistic terminology explained in the previous section does not imply that I understand PVs in a similar way. Although I believe that the exploration of the syntactic properties of PVs is appropriate, they do not account for the way meaning is created within these linguistic units. The total rejection of compositionality of meaning, apparently in accord with the idiomatic meaning of PVs, does not really fit a view of language where syntactic behaviour is related to meaning. In other words, their emphasis on grammar produces a separation from lexis that I do not think is appropriate.

This means that I subscribe to a Cognitive Linguistics approach to the analysis of PVs. As a consequence, I will propose that it is possible to posit some sort of compositionality in the interpretation of the meaning of PVs and that, based on this compositionality, the key to the understanding of these verbs lies in the meaning of particles.

According to the cognitive linguistics literature (Lindner 1981; Tyler and Evans 2003), particles are linguistic elements whose basic meaning can be traced back to their proto-typical use as spatio-temporal adverbs. From this basic meaning, sometimes resulting from metaphor and sometimes from what Gibbs (1997) calls experiential correlation, a radial web of senses are derived.

Thus, the cognitive framework used for the interpretation of spatial scenes can also be used, and as a consequence concepts such as Figure/Trajector or Ground/Landmark can serve to establish the meaning of the particle even when used in a PV construction. In this way, in a sentence like (3), the verb *walk out* will establish 'staff' as the trajector and 'bank' as the landmark of an action where the meaning of the verb is metaphorically derived from the spatial interpretation: 'the staff of most banks are not in their jobs today, i.e. they are on strike'.

(3) Staff at most banks walked out today

Cognitive Linguistics, then, provides a more adequate framework to study the acquisition of PVs since the concepts this school uses are more likely to reflect learners' intuitions of their meaning.

11.4 L2 Research on PVs

Research on the acquisition of PVs by L2 learners has mainly adopted a vocabulary approach. It is mainly concerned with the factors and circumstances that explain why L2 learners avoid using PVs (Dagut and Laufer 1985; Hulstijn and Marchena 1989; Laufer and Eliasson 1993; Sjöholm 1995; Liao and Fukuya 2004).

Although the amount of research on PVs can be considered to be scarce, some preliminary conclusions have already been drawn:

a. A preliminary finding established by the literature (Ishii and Sohmiya 2006; Siyanova and Schmitt 2007) is the clear distinction in the use of PVs between native and non-native speakers.
b. There are language distance effects, which will explain why L1 Dutch (Hulstijn and Marchena 1989) or Swedish (Sjöholm 1995) learners show less avoidance than L1 Hebrew learners (Dagut and Laufer 1985). Dutch and Swedish, in so far as they are Germanic languages, share certain similarities with English with respect to the use of particles.
c. A developmental sequence from avoidance in the first stages of acquisition to non-avoidance in the later stages has also been identified, although individual variability has been found with regard to proficiency. Advanced students show less avoidance than students at other levels (Liao and Fukuya 2004). However, this remains a controversial matter since a more recent study (Siyanova and Schmitt 2007) has found no difference in avoidance between proficiency levels.

d. The context of acquisition makes no difference (Siyanova and Schmitt 2007). Thus, learners of English as a Foreign Language and as a Second Language have difficulty in acquiring PVs. This, of course, may be another way of stating that proficiency levels have no determining influence on the final outcome.

e. Idiomaticity has been demonstrated to play a role in avoidance. Thus more opaque PVs will be susceptible to higher avoidance by L2 learners (Dagut and Laufer 1985; Liao and Fukuya 2004; however see Ishii and Sohmiya 2006, for different findings).

f. Avoidance is related to task effects (Liao and Fukuya 2004). Thus, more controlled tasks, like multiple choice tests, will produce fewer instances of avoidance.

These conclusions, however, have mainly been obtained in experimental conditions and one should consider whether a different methodology might produce different results or whether at least these could be qualified. Experimental tasks overlook learners' word choices in more normal, extended and unguided language use.

Thus, key factors such as frequency effects have not been taken into account and avoidance, which by definition should considered as a gradable concept, has been expressed in dual terms (yes/no, all/nothing). Finally, the selection of PVs used in the tasks of all the studies has not been justified and therefore it may well be that the verbs chosen are not representative of the phenomenon of PVs as such.

What is more important, if language distance or language typology determines L1 transfer, the concept should be used for something more than classifying L1s into different groups. Language typology (Talmy 2000) has been established on solid cognitive grounds and the explanation that it provides for other areas, such as motion events, may prove of great heuristic value, if only because it also deals with words belonging to the same parts of speech: verbs of motion and particles or prepositions expressing path.

11.5 Goals of the Study

In this context, the present chapter aims to study the acquisition of PVs by L1 Spanish learners of English and to confirm that language transfer or cross-linguistic influence may be posited as one of the driving forces behind the avoidance of this group of verbs. To support this hypothesis I have chosen to use natural rather than experimental data since, as has already been pointed out in the previous section, this approach may be more suitable to the task or at least offer a different perspective. This methodological decision also made it necessary to use natural data from learners with a different L1 background (in our case Swedish learners) and from native speakers; they both served as a benchmark against which to interpret the results obtained with the Spanish speakers.

The use of natural data also made it possible to broaden the scope of our study, and it allowed us to focus on other factors that may have an influence on the acquisition of PVs by Spanish learners. It is important here to note that the new developments in the field of transfer, or Cross-Linguistic Influence, as Jarvis and Pavlenko (2008) like to call it, allow for the consideration of simultaneous intervening factors.

As there is a considerable number of PVs, this study has only focused those containing the particle 'out', on the grounds that the conclusions drawn from it could be useful to explain some of the patterns of PV avoidance by L2 students.

11.6 Methodology

a. Corpora

As we have seen, the study has been conducted using natural data from existing corpora of learner language. In this case, both the Spanish and Swedish sections of the ICLE were used. The ICLE is a non-tagged corpus of short essays written by 3rd and 4th year university students on non-academic, non-technical controversial topics (e.g.: 'Television is the opium of the masses'). The average length of the essays is about 500 words, with the Spanish subcorpus containing about 200,000 words.

To provide an element of contrast and comparison of results, the university and school essay sections from the BNC were also used. Like the ICLE, they mainly consist of argumentative essays written in this case by native speakers of British English.

The most important aspect to be considered was the comparability of the corpora in terms of subject matter and size. The subject matter of all the corpora used includes an array of topics ranging from literature to current affairs with a degree of specificity which may be regarded as intermediate. With regard to the size of the corpora used (see Table 11.1), the match could be considered more than appropriate to be able to reach some preliminary conclusions.

Table 11.1 Corpora

Corpus	Tokens	Types	TTR
Spanish section of ICLE (SPICLE)	200,926	12.161	6
Swedish section of ICLE (SWICLE)	198,675	11.434	6
Written School and University Essays from the BNC	202,247	14.366	7

b. Procedure

Once the corpora were selected, the procedure carried out was the following:

1. All the concordances of the particle 'out' found in the three corpora were extracted irrespective of whether the particle was located on its own or accompanied by the preposition 'of'. This was done using the concordancer *WSmith Tools*.
2. The concordances containing the particle 'out' were then tagged using the following labels:

   a. *PV status*. The categories used here were

      i. VPC (Verb Particle Construction), i.e. Phrasal Verbs.
      ii. V+P (Verb and Particle), constructions where the verb and the particle, though constructed together, do not constitute a syntactic unit, and basically describe motion events.
      iii. Non-VP, instances where the particle 'out' is used outside the scope of direct influence by the verb.

   b. *PV syntax*. This was encoded in two related fields: The first field indicated whether the verb was used transitively or intransitively, and the second expressed the position of the complement (before or after the particle) when the verb was used transitively.
   c. *Errors*. The existence of a deviation from native usage was also recorded and then classified into different categories: lexical, syntactic, collocational or orthographic.
   d. *Particle meaning*. Since the meaning of the particle was considered essential, all the instances were coded using the different meanings proposed by Tyler and Evans (2003: 203–216) for the particle 'out' (see Table 11.2).

3. The resulting database was included in a spreadsheet and the results were analysed for statistical significance using a binomial test.

Table 11.2 Meanings of 'out'

OUT

1. PROTO-SCENE: *Exterior to a Landmark*
2. LOCATION CLUSTER
   a. Not In Situ Sense: *Amy is out sick for the day*
   b. No More Sense: *We're out of business*
   c. Completion Sense ('completely'): *The ground has now thawed out*
3. THE VANTAGE POINT IS INTERIOR CLUSTER
   a. Exclusion Sense: *They voted out the unpopular member*
   b. Lack of Visibility Sense: *He switched out the light; He crossed out the typo*
4. THE VANTAGE POINT IS EXTERIOR CLUSTER
   a. Visibility Sense: *The sun is out*
   b. Knowing sense: *The secret is out; We figured out the problem*
5. THE SEGMENTATION CLUSTER
   Distribution Sense: *I'm always having to fork out on my old car*
6. REFLEXIVITY: *Spread out the butter*

OUT OF

7. MATERIAL SOURCE: *The chair is made out of wood*
8. THE CAUSE SENSE: *John sacrificed himself out of love*

11.7 Results and Discussion

The analysis presented here mainly focuses on what is considered to be the leading factor explaining the acquisition of PVs by L2 learners: avoidance. But PV avoidance is not a simple phenomenon and I will attempt to portray some of its complexity by relating it to other factors and by looking at it from different perspectives. Besides, I will also deal with 'frequency effects', which, although apparently contradicting the importance of L1 transfer, can also be interpreted in ways that are consistent with our findings on PV avoidance.

a. PV Avoidance

In a corpus study like the present one, it is not possible to measure avoidance behaviour by resorting to 'think aloud' protocols or similar tasks used in L2 studies, to access the intention or awareness of the learner in producing or, as in this case, in avoiding a specific structure. As a consequence avoidance will be defined, in this study, in comparison to habitual behaviour by native speakers in a similar context. That is, L2 speakers will be considered to avoid using a PV when they use a significantly smaller number of instances of the particle 'out' than native speakers.

In using this definition of avoidance, I share the view expressed by Liao and Fukuya (2004), who do not think it necessary to ascertain the learner's previous knowledge of a PV. In contexts of real language use, it would require an enormous methodological effort to adopt the criterion that learners can only avoid what they already know. More importantly, when this previous knowledge is measured, as some of the studies have done (see, for example, the case of Siyanova and Schmidt 2007), some sort of experimental task is necessary, thus failing to take advantage of the richer information derived from natural contexts. In a way, my definition of avoidance is very similar to one given, in corpus linguistics studies, for the term underuse (see, for example, Cobb, 2003), which I think is more appropriate. I keep the former term because it is the one used in the bibliography dealing with PVs.

As Table 11.3 shows, the total number of 'out'-PVs used by both L1 Swedish and L1 Spanish learners is significantly smaller than that habitually used by L1 English speakers. This means that avoidance of 'out'-PVs can be said to affect all the second language learners studied, although, as we can see in the same table, Swedish learners have a much lower level of avoidance than Spanish students.

These data are confirmed by the results analysed in Table 11.4, where the data for 'out'-PV types are recorded. As one would expect, avoidance is not only related to stylistic choices but to the size of the vocabulary that students show.

These results confirm for PVs what has been hypothesized for formulaic sequences in general (Wray 2002), i.e. that non-native speakers are less likely to use PVs than native speakers. In this sense, the findings are not new

Table 11.3 Out-PV tokens

VPC Tokens	Number	%	Avoidance
BNC	283	100,00	–
SWICLE	194	69,26	30,74
SPICLE	127	44,88	55,12

Table 11.4 Out-PV types

VPC Types	Number	%
BNC	107	100,00
SWICLE	60	56,07
SPICLE	35	32,71

and bear out, for the formal written context analysed here – essay writing – the ones obtained by Siyanova and Schmitt (2007) for more informal spoken contexts. But, at the same time, they contradict the hypothesis these authors put forward at the end of their article ('the notion that learners tend to avoid multi-word verbs in spoken colloquial, but perhaps not in written contexts' Siyanova and Schmitt 2007:133) when they compared their results with previous studies on PVs and attempted to explain why avoidance was not found in some of them (Hulstijn and Marchena 1989; Liao and Fukuya 2004). In my opinion, avoidance was found to be statistically non-significant in the latter studies not because of the written context but because of thee types of experimental tasks used, such as multiple-choice tests, which tap less into natural language use and the skills typical of online processing.

On the other hand, the fact that Spanish L1 learners are less likely to use PVs than are Swedish L1 learners can be explained – as Laufer and Eliasson (1993) and Sjöholm (1995) do for Hebrew and Finnish vs. Swedish – on the grounds that Spanish lacks this category of verbs while Swedish does not. In other words, the present study shows, once again, that the L1 of the learner is highly influential and that the distance between the L1 and the L2 can explain a great number of the problems learners may have with this construction.

b. PV Avoidance as a particular case of particle avoidance

The importance of PV avoidance can be more clearly perceived if considered in the larger context of particle use. Thus, as Figure 11.1 shows and Figure 11.2 corroborates, 'out'-PV avoidance is by far the most important phenomenon if compared to avoidance in similar constructions. In other

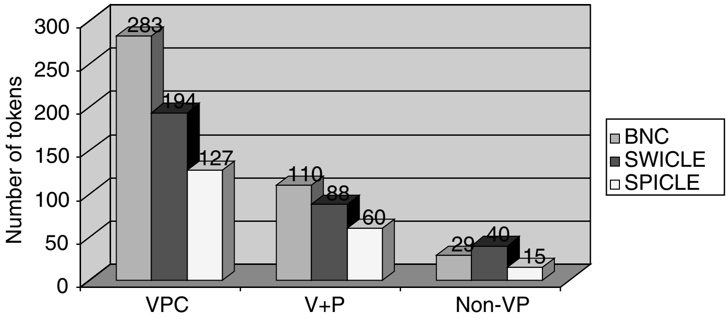


FIGURE 11.1 Number of tokens in the different constructions

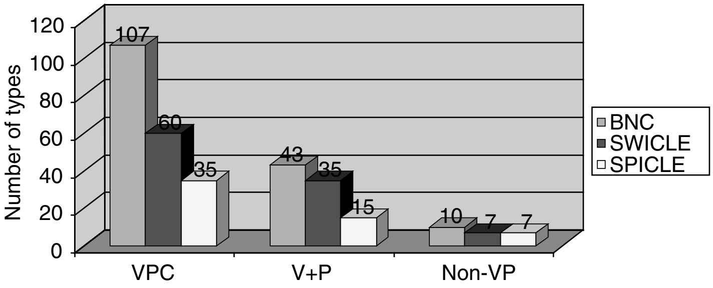


Figure 11.2 Number of types in the different constructions

words, the particle 'out' is avoided to a much a greater extent when used in PV constructions (VPCs) than when it is used in other constructions of the English language to describe motion events (V+P) or in other adverbial or prepositional uses that fall outside the immediate scope of influence of the verb. This is not surprising given the idiomatic or non-transparent meaning of PVs in comparison with other constructions.

However, in spite of these differences, the avoidance of 'out'-PVs should be understood as a special case within a general trend of avoidance of the particle 'out', as L2 learners, especially Spanish speakers, also tend to avoid the use of the particle in the remaining constructions types (V+Ps and Non-VPs). Indeed, these data suggest that PV avoidance could be related with lower frequency with which path morphemes and manner verbs occur in the narratives of L1 Spanish speakers while describing motion events in an L2 (see Cadierno 2004). In other words, avoidance of PVs could be seen as further support of the 'thinking for speaking hypothesis' (Slobin 1996, 1997, 2000, 2003), since the particles used PVs, as defended by Cognitive Linguistics, can be considered as metaphoric extensions from their spatio-temporal meanings (Tyler and Evans 2003).

c. Specific areas of avoidance within PV use

Further insight into PV avoidance may gained if, instead of looking at PVs as a homogenous phenomenon, we consider their semantic and syntactic variation. I analyse this variation using a cognitive-linguistic framework, which is particularly useful for the semantic analysis of PVs. Thus, as indicated in the methodology, I use Tyler and Evans (2003) to identify the meaning of 'out' in PVs. This will allows a more objective and fine-grained semantic analysis, less dependent on judgment than the labels 'opaque' and 'transparent', used so far in the bibliography, to classify PVs from a semantic point of view.

Table 11.5 Number of errors in L2 use

Deviations	No. of errors	% over total
SWICLE	15	7.73
SPICLE	49	38.58

Table 11.6 presents the range of meanings associated with the particle 'out' in English and the frequency with which each meaning is found in the three corpora studied (BNC, SWICLE and SPICLE). If we pay attention to significant results, we can see that the relative frequency of certain meanings does not really correspond to typical native speaker behaviour. Thus, in the case of L1 Spanish speakers, we find an overproduction of PVs expressing 'visibility/ knowing' and of those expressing 'completion', which is also noticeable, as we can see (cf. Table 11.5 above), in the number of deviant PVs L1 Spanish students use to convey these meanings (e.g. 'From this point on everything tries to clear out'). On the other hand, PVs expressing 'location'/'motion' and 'distribution' are much less frequently used by L1 Spanish speakers. For their part, L1 Swedish speakers show a less marked preference for 'visibility' and 'knowing' meanings than Spanish speakers, but still use more PVs expressing those meanings than native speakers. L1 Swedish speakers, however, make less use of PVs expressing 'completion' and 'invisibility'.

These results indicate that avoidance may be affected by the meaning of the particle. This would imply that the L1 of the learner may have an influence in the avoidance of certain meanings of the particle, perhaps those that are less transferable from their L1. On the other hand, the overuse of the 'visibility/knowing' meaning may be the result of a more complex picture where L1 transfer and frequency effects have a combined influence. As we will see later, frequency is a factor that should also be taken into account.

For its part, the syntax of both L1Spanish and L1 Swedish learners (see Table 11.7) also shows areas of cross-linguistic influence. While the former group seems reluctant to insert a Nominal Phrase (NP+out) before the particle (*Try to help sort things out*) and concomitantly overproduce the symmetrical structure (out+NP), the latter also underproduce the NP insertion (NP+out). Finally, both groups of learners underproduce sentences where the particle is stranded.

The similarity in behaviour of both groups of learners indicates that the L1 would be less influential and that factors at play here could be related to intrinsic difficulty of the syntactic construction for L2 learners in general. As suggested by the cognitive-linguistic bibliography (see Dirven 2001), NP insertion before the particle, which is usually called Construction 2, is used for

Table 11.6 Numbers and percentages of meanings for Out-PVs

	VPC #			VPC %		
MEANING2	**BNC**	**SPICLE**	**SWICLE**	**BNC %**	**SPICLE %**	**SWICLE %**
VISIBILITY/KNOWING	131	75	103	46.29	59.06*	53.09*
COMPLETION	54	28	31	19.08	22.05	15.98
LOC/MOV	36	4	23	12.72	3.15*	11.86
EXCLUSION	18	8	17	6.36	6.30	8.76
INVISIBILITY	14	8	3	4.95	6.30	1.55**
DISTRIBUTION	14	2	7	4.95	1.57*	3.61
BEYOND	7		2	2.47	0.00**	1.03
NOT IN SITU	2	2	5	0.71	1.57**	2.58**
REFLEXIVITY	7		1	2.47	0.00	0.52
NO MORE			2	0.00	0.00	1.03
Grand Total	283	127	194	100.00	100.00	100.00

Notes Percentages whose differences with the BNC were found to be *significant $p < 0.05$; **marginally significant = $p < 0.10$

Table 11.7 Numbers and percentages for structures used with transitive Out-PVs

PARTICLE PLACEMENT	**BNC**	**SPICLE**	**SWICLE**	**BNC (%)**	**SPICLE (%)**	**SWICLE (%)**
NP + OUT	16	3	1	8,04	3,03**	0,80*
OUT + NP	62	44	44	31,16	44,44*	35,20
OUT + CLAUSE	53	25	40	26,63	25,25	32,00
PR + OUT	12	6	10	6,03	6,06	8,00
NON-FOLL.OBJ	56	21	30	28,14	21,21	24,00*
Total	199	99	125	100,0	100,00	100,00

Notes Percentages whose differences with the BNC were found to be *significant $p < 0.05$; **marginally significant = $p < 0.10$

discourse cohesion purposes and involves a greater degree of automaticity in language use since direct objects placed in mid-position require a lesser degree of awareness.

d. Errors in the use of PVs

Avoidance only reflects part of the problems students find when using PVs. Language transfer may also occur in the form of deviation from native speaker usage. Again there is a difference between L2 learners and native speakers, but as Table 11.5 shows Swedish learners would seem to have significantly fewer problems than Spanish speakers.

Here we can see some examples of errors made by Spanish learners:

- ‘And the last main point **to jut out** but not the less important is the role psychiatrics play inside jail.’ (SPICLE, lexical)

- 'It **points out** the idea of the family as the pillars of Victorian Society.' (SPICLE, lexical: selection of the particle)
- 'we should **stand out**, the incipient attention to the social problems of the pre-capitalist society.' (SPICLE, lexical and syntactic)
- 'From this point on everything tries to **clear out**.' (SPICLE, lexical)
- 'Nowadays the number of people who don't **carry out** Military Service has increased.' (SPICLE, collocation)

The mostly lexical nature of the errors suggests that learners overgeneralize the use of the particle 'out' to form PVs that are non-existent in the English language. Language transfer may not be the only factor at play, however, as will be seen in the following subsection.

e. Frequency effects

The corpus analysis carried out in this article is especially helpful to identify factors that are much less evident using an experimental methodology. This is the case of frequency effects, which, following Ellis (2002), I understand as both the ease of processing and the learning outcomes derived from the frequency with which some linguistic elements are found in the input. Thus, since 'fluent language users tend to produce the most probable utterance for a given meaning on the basis of the frequencies and recencies of utterance representations' (Ellis 2002: 162), this article assumes that those PVs used with a high frequency in the corpora analysed are most probably the result of implicit intralingual learning factors, dependent on the input speakers have been exposed to, rather than the result of cross-linguistic influence.

This assumption is supported by the fact that the most frequent PVs in all three corpora – all with a frequency higher than 5 – are, not surprisingly, the same: *point out, carry out, find out* and *turn out.* It seems as if the specific essay writing task used to compile the corpora activates these specific PVs as they are very frequent in argumentative text-types. Figure 11.3 offers further confirmation of how L2 learners use of PVs parallels that of native speakers.

The importance of frequency effects is further emphasized if we consider not only the particular verbs but the meanings expressed through the particle. As Figure 11.4 shows, L2 students are very aware of the prototypical meanings that are expressed through the particle. This is also the case of syntactic constructions (see Figure 11.5).

Finally, frequency effects can also be seen in the syntactic patterns used with transitive verbs. Learners are not only aware of the frequency with which certain meanings are used but they also pay attention to the most prototypical syntactic structure in which PVs appear.

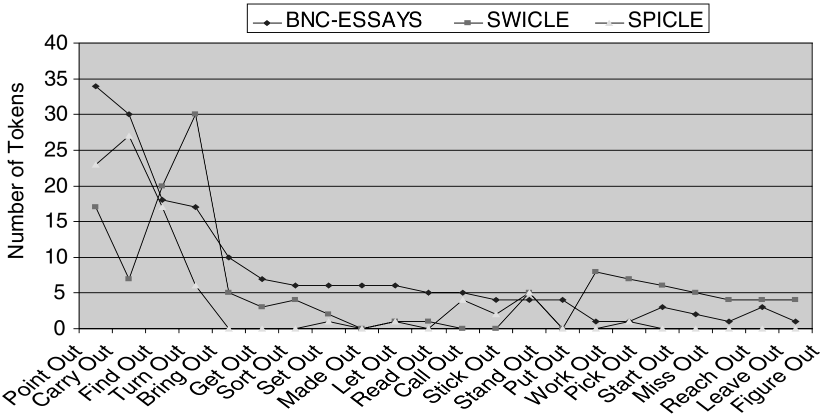


FIGURE 11.3 Comparison of most frequent Out-PVs

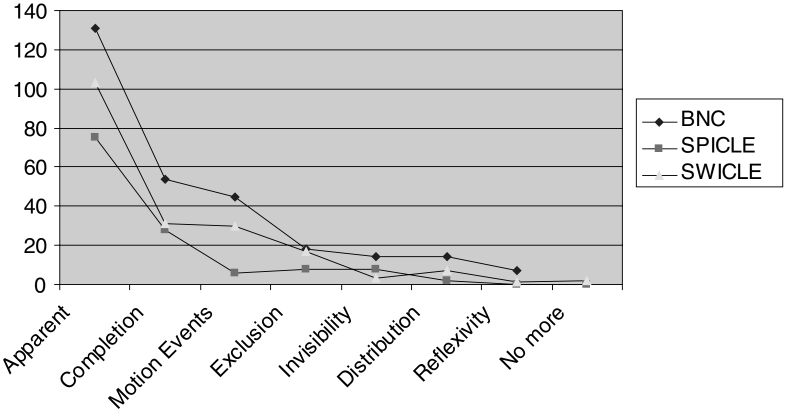


FIGURE 11.4 Most frequent meanings of Out-PVs

It is by looking at these frequency effects that we can now explain some of the phenomena of overuse that were detected along with avoidance. As Cobb (2003) states, overuse is the other side of avoidance and is closely linked to it.

11.8 Conclusion

The analysis of OUT-PVs shows that learners' L1may have an influence on underproduction of this group of verbs. This finding would seem to be in

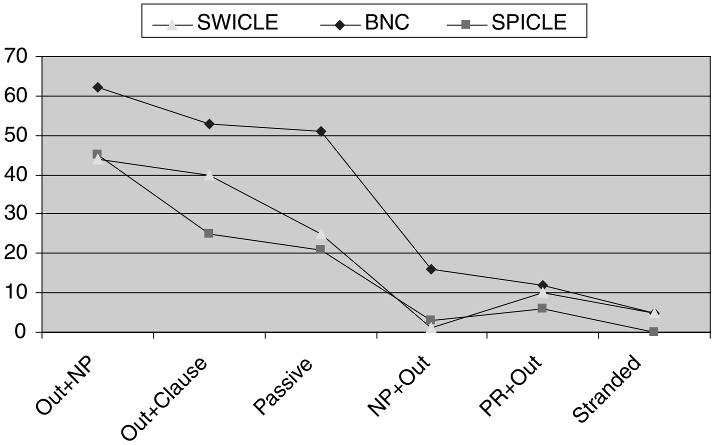


Figure 11.5 Most frequent structures used with transitive Out PVs

accordance with the analyses which, based on Slobin's 'thinking for speaking' hypothesis, posit that what Talmy (2000) calls Verb-framed languages (Romance languages in general) are less likely to express the path of motion events.

The data shown here for Spanish speakers reveal that the use or avoidance of PVs will reflect this tendency even more markedly than motion events. The level of avoidance detected in this study is a reliable indication that L1 Spanish learners of English underproduce 'out'-PVs to a much greater degree than the speakers of Swedish.

However, avoidance of PV use seems to be compatible with awareness on the part of learners of the frequency and prototypicality of the different PVs, their most frequent meanings and the structures in which they are used.

Future research should establish whether the patterns described in this study will also hold for the entire phenomenon of PVs.

References

Cadierno, T. (2004), 'Expressing motion events in a second language: A cognitive typological perspective', in Achard, M. and Niemeirer, S. (eds), *Cognitive Linguistics, Second Language Acquisition and Foreign Language Teaching*. Berlin: Mouton de Gruyter, pp. 13–49.

Cappelle, B. (2005), *Particle Patterns in English. A Comprehensive Coverage*. Ph.D. Dissertation at the Faculteit Letteren, Katholieke Universiteit Leuven.

Cobb, T. (2003), 'Analyzing late interlanguage with learner corpora: Quebec replication of three European studies'. *Canadian Modern Language Review*, 59, 393–423.

Dagut, M. and Laufer, B. (1985), 'Avoidance of Phrasal Verbs: A Case for Contrastive Analysis'. *Studies in Second Language Acquisition,* 7, 73–79.

Dirven, R. (2001), 'English phrasal verbs: Theory and didactic applicaton', in Pütz, M., Niemeier, S. and Dirven, R. (ed.), *Applied Cognitive Linguistics.* 2 Vols. Berlin. Mouton de Gruyter, pp. 3–27.

Dulay, H. and Burt, M. (1974), 'Natural sequences in child second language acquisition'. *Language Learning*, 24, 37–53.

Ellis, N. (2007), 'The associative-cognitive CREED', in Van Patten, B. and Williams, J. (eds), *Theories in Second Language Acquisition. An Introduction*, Mahwah, New Jersey/London: Lawrence Erlbaum Associates, pp. 77–98.

Ellis, N. C. (2002), 'Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition'. *Studies in Second Language Acquisition*, 24, 143–188.

Hulstijn, J. H. and Marchena, E. (1989), 'Avoidance: Grammatical or Semantic causes?' *Studies in Second Language Acquisition*, 11, 241–255.

Ishii, Y. and Sohmiya, K. (2006), 'On the semantic structure of English. Spatial particles involving metaphors', in A. Yohitomi (ed.), *Readings in a Second Language Pedagogy and Second Language Acquisition: In Japanese Context.* Amsterdam: John Benjamins, pp. 381–402.

Jarvis, S. and Pavlenko, A. (2008), *Crosslingistic Influence in Language and Cognition.* New York/London: Routledge.

Krashen, S. (1981), *Second Language Acquisition and Second Language Learning.* Oxford: Pergamon.

Laufer, B. and Eliasson, S. (1993), 'What causes avoidance in L2 learning: L1-L2 difference, L1-L2 similarity, or L2 complexity?', *Studies in Second Language Acquisition*, 15, 35–48.

Lee, D. (2001), *Cognitive Linguistics: An Introduction.* Victoria, Australia: Oxford University Press.

Liao, Y. and Fukuya, Y. J. (2004), 'Avoidance of phrasal verbs: The case of Chinese learners of English'. *Language Learning*, 54, (2), 193–226.

Lindner, S. (1981), *A Lexico-Semantic Analysis of Verb-Particle Constructions with up and Out.* San Diego: University of California.

Odlin, T. (1989), *Language Transfer.* Cambridge: Cambridge University Press.

O'Dowd, E. M. (1998), *Preposition and Particles in English. A Discourse-Functional Account.* New York/Oxford: Oxford University Press.

Quirk, R., Greenbaum, S., Leech, G.N. and Svartvik, J. (1985), *A Comprehensive Grammar of English.* London: Longman.

Siyanova, A. and Schmitt, N. (2007), 'Native and non-native use of multi-word vs. one-word verbs'. *IRAL*, 45, 119–139.

Sjöholm, K. (1995), *The Influence of Crosslinguistic, Semantic, and Input Factors on the Acquisition of Phrasal Verbs.* Abo: Abo Akademi University Press.

Slobin, D. I. (1996), 'Two ways to travel: Verbs of motion in English and Spanish', in Shibatani, M. and Thompson, S. A. (eds), *Grammatical Constructions: Their Form and Meaning.* Oxford: Clarendon Press, pp. 195–220.

—(1997), 'Mind, code, and text', in Bybee, J., Haiman, J. and Thompson, S. A. (eds), *Essays on Language Function and Language Type: Dedicated to T. Givón.* Amsterdam/Philadelphia: John Benjamins, pp. 437–467.

—(2000), 'Verbalized events: A dynamic approach to linguistic relativity and determinism', in Niemeier, S. and Dirven, R. (eds), *Evidence for Linguistic Relativity*. Amsterdam/Philadelphia: John Benjamins, pp. 107–138.

—(2003), 'Language and thought online: Cognitive consequences of linguistic relativity', in Gentner, D. and Goldin-Meadow, S. (eds), *Language in Mind: Advances in the Study of Language and Thought*. Cambridge, MA: MIT Press, pp. 157–192.

Talmy, L. (1985), 'Lexicalization patterns: Semantic structure in lexical forms', in Shopen, T. (ed.), *Language Typology and Syntactic Description*. Cambridge: Cambridge University Press.

—(2000), *Toward a Cognitive Semantics, Vol. 1: Concepts Structuring Systems*. Cambridge, Massachusetts, MA: MIT Press.

Tyler, A. and Evans, V. (2003), *The Semantics of English Prepositions*. Cambridge: Cambridge University Press.

Wray, A. (2002), *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Chapter 12

Analysing EFL Learner Output in the MiLC Project: An Error *It's, but Which Tag?

Mª Ángeles Andreu Andrés, Aurora Astor Guardiola,
María Boquera Matarredona, Penny MacDonald,
Begoña Montero Fleta and Carmen Pérez Sabater
*Grupo DIAAL, Departamento de Lingüística Aplicada,
Universidad Politécnica de Valencia, Valencia*[1]

12.1 Introduction

The development of Computer Learner Corpora (CLC) in the early 1990s marked a new direction in the field of corpus linguistics and its relation to foreign language learning research and pedagogy. According to Granger (2003), CLC are electronic collections of authentic foreign or second language data. Undoubtedly, the International Corpus of Learner English (ICLE), founded and coordinated by Sylvianne Granger of the Université Catholique de Louvain in Belgium (Granger 1993, 1998), is the most cited in the literature. It was based on a large collection of essays written by French-speaking undergraduates of English Language and Literature. The original project was later expanded to include texts produced by language learners from a variety of different L1 backgrounds, including French, German, Dutch, Spanish, Swedish, Finnish, Polish, Czech, Bulgarian, Russian, Italian, Hebrew, Japanese and Chinese. In general the sub-corpora in the ICLE are of approximately 200,000 words per native language, and are therefore much smaller than native speaker (NS) corpora in general.

Our corpus, which is in its early stages, will be a multilingual learner corpus involving the written work of students learning English, Spanish, French and German as a foreign language, and also Catalan, as a first, second or foreign language. With a student population of approximately 30,000, and a total of 1,750 credits on the curriculum assigned to the Department of Applied Linguistics, we expect to achieve a reasonable sized corpus. The degree courses on offer include Architecture, Fine Arts, Civil Engineering, Agronomy, Applied Computer Science, Industrial Engineering, Geodesy, etc. The students are required to read a large amount of scientific and technical texts, and to produce written texts themselves which may be of a specific nature and related to their mainstream subjects, or may involve general language output. To our knowledge, the only other multilingual learner corpus to date, COMET (Tagnin 2001), is being developed at the University of Sao Paulo, Brazil, and has been created for teaching and translation studies.

In the research reported here, we aimed to carry out an exhaustive analysis of linguistic errors, trying to define each category in detail and thus avoiding cases of overlap, as far as possible. With this in mind, our research question was the following: is it possible for the group of researchers involved in the MiLC project to carry out a computer-aided error analysis of student output and coincide with the detection, classification and correction of the errors? And if not, why?

12.2 Materials and Methodology

An analysis of our students' linguistic errors was carried out. It involved the following stages:

- Looking in detail at the different errors detected and how they are tagged by the researchers involved.
- Commenting on similarities and differences in tagging.
- Analyzing nuances/interferences that may affect this tagging.

The topic was presented to the students as shown in Appendix I. They were asked to write a short text on the subject of 'Immigration'. The variables were controlled by asking the students to fill in a form providing information concerning their sex, mother tongue, years at university, degree course, etc.

Our research work was concerned with detecting, classifying and correcting the errors in our interlanguage (IL) corpus using the Université Catholique de Louvain (UCL) Error Editor.[2] This tagging method, developed by Dagneaux et al. (1996), uses codes to classify the deviant forms according to their surface linguistic description.

To practise using the error tagging method, we took one text and analysed it, first individually, and then the group met to discuss the findings. All the tags in the manual were used, as well as two tags of our own related to punctuation and code-switching. When comparing the results of the analysis, it was found that there were notable differences regarding the classification of the errors and the suggested corrections. The analysis was carried out by 7 researchers, all of them experienced university teachers of ESP: 3 working individually and 4 in pairs, coded as Researcher 1 = R1, R2, R3, R4 (two researchers working together), R5 (two researchers working together) and R6. The results presented in this study show the error analysis before any consensus was attempted to be reached on the classification.

Although there are a total of forty error categories, the use of only five tags was decided in order to quantify results, differences and coincidences more easily:

- **GA** (article *the* and *a/an*).
- **GP** (pronouns).
- **LS** (lexical single).
- **GVN** (noun-verb concordance).
- **LSF** (false friends).

The original text before tagging can be seen in Appendix II.

12.3 Results and Discussion

The quantitative results are shown in Table 12.1.

As can be seen, the most notable differences concerning the error tags used involve article errors (GA), and lexical errors (LS). We discuss the results below in the order they appear in the table.

12.3.1 GA (Article errors)

Following Quirk et al. (1972), the way we use articles with nouns having generic reference varies according to the type of noun. More specifically, article usage varies depending on whether the noun is countable or non-countable.

Table 12.1 Quantitative results of error tagging

Researcher groups	GA	GP	GVN	LS	LSF
R1	9	4	1	6	0
R2	3	4	0	5	0
R3	9	3	1	0	1
R4	7	3	1	11	0
R5	5	5	1	1	0
R6	10	2	0	0	2

In generic reference, countable nouns need an article in the singular but not in the plural. However, as Swan and Smith (1987: 83) point out, in Spanish the definite article goes with mass nouns and plural countable nouns when used with a general meaning, whereas in English this is not the case. Also there are certain contexts in English (i.e. with single countable nouns) where articles are needed, and are not required in Spanish e.g. **My sister is teacher* (*Mi hermana es profesora*). Article errors are therefore quite frequent among Spanish learners of English.

In the first use of the word 'immigrants', the learner has made no errors as he/she copied directly from the instructions that were given as an introduction to the topic.[3] However the next mention of the noun 'immigrants' prompts a correction in exactly half of the researchers. Surprisingly, some evaluators did not consider this use of the article to be an error. In the third case, four of the six researchers classed this as an article error. It must be noted, however, that the researchers who decided not to tag the use of the article as erroneous were consistent with this view of the specific reference being made throughout.

R1

(LS)Since for some years, in Spain and a lot of countries more, the number of immigrants is increasing.

(GA) The 0 immigrants are a problem but, also, they are a benefit for the city. (LCLC) In one hand $on the one hand$, (GA) the 0 immigrants emigrate

R2

Since some years (WM) ago, in Spain and a lot (WR)of 0 (WO) countries more $more countries$, the number of immigrants is increasing.

The immigrants are a problem (PW) but, 0 (WO) also, they are $they are also$ a benefit for the city. (LCLC) In one hand $On the one hand$, the immigrants emigrate

R3

(SU) Since some years $?$, in Spain and a lot of countries (GADJN) more 0, the number of immigrants is increasing.

(GA) The 0 (GWC) immigrants $immigration$ are a problem but (PW), also (PW), they are a benefit for the city. (LCLC) In one hand $on one hand$, (GA) the 0 immigrants emigrate

R4

(LS) Since in (LS) some $the past few$ years, in Spain and (S) a lot of countries more $many more/other countries$, the number of immigrants (GVT) is $has been$ increasing.

The immigrants are a problem but (PW), $0$ (GADVO) (FS) also (PW), $0$ they are $are also$ a benefit (WR) for the city$0$. (LS) In $On$ one hand, (GA) the $0$ immigrants emigrate

R5

(GA) Since some years $for some years$, in Spain and (WM) $in$ (GADJCS) a lot of countries more $many more countries$, the number of immigrants (GVT) is increasing $has been increasing$.

The immigrants are a problem but(PW), (WO) also, they are (LS) a benefit $beneficial$ for (GA) the $0$ city (FM)(GNN) $cities$. (LCLC) In one hand $On the one hand$, (S) the immigrants emigrate

R6

(LCC) Since some years $for$, in Spain and (WM) a lot of countries $also in a lot of countries$ more (SU), the number of immigrants is increasing. (GA) The immigrants $immigrants$ are a problem but, (WO) also, they are $they are also$ a benefit for the city. (LCLC) In one hand $On the one hand$, (GA) the immigrants $0$ emigrate because

In attempting to understand the reasoning behind these differences, we referred once again to Quirk et al. (1972). The use of the definite article 'the' depends on the concept of shared knowledge, encompassing the reference to the 'immediate situation' (this may be linguistic and/or extralinguistic), and also to the 'larger situation' involving general knowledge. The term 'immigrant' has already been mentioned in the general introductory sense in the first sentence, and therefore the definite article that follows when the next reference is made to 'the immigrants' may be considered anaphoric as 'the term anaphoric reference is used where the uniqueness of reference of some phrase *the X* is supplied by information given earlier in the discourse' (Quirk et al. 1972: 267). It may also be the case of the 'larger situation' whereby the definite article is used to refer to cases where the mutual understanding is derived from the extralinguistic situation (Ibid.). As the phenomenon of immigration has become a prominent issue in the media, it may be that when referring to 'the immigrants' the evaluators who did not mark the use of the definite article as incorrect understand that the writer is writing about the immigrants we all see and hear about every day, and in this sense they are specific.

Thus the disparity in the results is due to the fact that the researchers did not agree upon the nature of one particular noun as being of generic or specific usage, and also the frequency of use of the noun 'immigrants' created a further imbalance in the results.

12.3.2 GP (Pronouns)

Spanish L1 learners of English have a particularly high incidence of this type of error (MacDonald 2004). Most errors in this category involve incorrect choice of personal pronouns and possessive adjectives. This may be due to the fact that subject pronouns are mostly unnecessary in Spanish since the verb inflection indicates person and number. Errors are frequently made by learners in their elementary practice of English in the correct use of possessive pronouns both attributive and predicative. In the analysis of the corpus under study, the subcategory GP – the use of errors involving not only all categories of pronouns but also reference problems – seemed to be easy to identify by the researchers, and there was a high rate of agreement among them. However, only one of the researchers (R5) included reference when analysing this subcategory: 'A city (GP) as Valencia $such as Valencia$', 'Problems (GP) as $such as$ the decrease'. As regards the tagging of both 'like' and 'as', there was a certain amount of disagreement, for instance, R2 thought it should be an LS (single lexical error).

R1

(GP) this $these$ people
(GP) his $their$ miserable life
(GP) other a country

R2

(GP) they $themselves$
(GP) this $these$ people
(GP) other $another$
(GP) his $their$ miserable (FM) life $lives$

R3

(GP) his $their$ miserable
for (GP) they $them$ and immigrants (GP) that 0 come

R4

(GP) It $Immigration$
(GP) This fact $which$ (not correctly tagged)

R5

for (GP) they $them$ and for their families
(GP) this $these people$
(GP) his miserable life $their

A city (GP) as Valencia $such as Valencia
Problems (GP) as $such as$ the decrease

R6
(GP) this $these$ people
(GP) his $their$ miserable life

12.3.3 GVN (noun-verb concordance)

According to Quirk et al. (1972: 359), concord can be broadly defined as the relationship between two grammatical elements such that if one of them contains a particular feature (e.g. plurality) then the other also has to have that feature. The most important type of concord in English is concord of number between subject and verb. In the present error study, GVN, errors of concord between a subject and its verb were identified by four correctors; one other (R2) was chosen to mark the whole of the first clause of the sentence as a style error (although including a change in subject verb concordance), while R6 had not noticed the error, possibly because of high reading speed or lack of attention.

R1, R3, R4, R5.
This fact (GVN) cause $causes$

R2
(S) This fact cause a lot of people begin to suspicious about them. $This fact makes a lot of people begin to be suspicious of them$

12.3.4 LS (lexical single)

First it should be noted that this category of error, like the definite article, also showed a great disparity in its classification. The differences range from zero tags in the case of R3, to the highest number of instances, 11, detected by R4. We shall look in greater detail at the cases presented here and the possible reasons for the differences in tagging by the participants in the project.

12.3.5 * Since some years

R1
(LS)Since for some years

R2

Since some years (WM) ago

R3

(SU) Since some years $?$

R4

(LS) Since in (LS) some $the past few$ years

R5

(GA) Since some years $for some years$,

R6

(LCC) Since some years for

In this case, although 3 of the 5 researchers proposed the same correction for this error, the tagging of the error does not coincide. One researcher was not sure how to handle it at all (R3), and the other two proposed completely different alternatives. In retrospective, R5 realized that it could not be tagged as GA.

12.3.6 *In one hand

The connectors that involve more than one word are, on the whole, more difficult to acquire as the learner has to memorize a longer term whose different parts are completely arbitrary i.e. *Por otra parte* (Sp.), On the other hand. These are what Nattinger and de Carrico (1992) describe as 'strings of specific lexical items which allow no paradigmatic or syntagmatic substitution' (1992: 36):

R1

(LCLC) In one hand $on the one hand$,
(LCLC) In the other hand $on the other hand$,

R2

(LCLC) In one hand $On the one hand$
(LCLC) In the other hand $On the other hand$

R3

(LCLC) In one hand $on one hand$,
(LCLC) In the other hand $on the other hand$,

R4

(LS) In On one hand,
(LS) In 0 the other hand

R5
(LCLC) In one hand $On the one hand$,
(LCLC) In the other hand $On the other hand $,

R6
(LCLC) In one hand $On the one hand$,
(LCLC) In the other hand, $On the other hand $,

In two cases, the group with the highest rate of tags in the LS category wrongly tagged the expression *'In one hand', not realizing that, following the norms suggested in the manual, this error is a 'complex logical connector', and should have been tagged (LCLC).

12.3.7 LSF (false friends)

There was a certain amount of doubt concerning the identification and classification of this category. When we carried out the first consultations, there was initially confusion caused by a tendency to classify as false friends those terms that could be attributed to direct transfer from the L1 to the L2, but which technically speaking were not false friends. Moss (1992: 142) makes a distinction between false cognates and false friends as defined below:

A. False cognates are those words that are similar in appearance but do not descend from a common ancestor, e.g. Spanish *pie* not cognate with English *pie*, or Spanish *pipa* not cognate with English *pip*.
B. False friends groups together those words that have similar ancestors but whose meanings (or some of their meanings) have diverged over time, e.g. Spanish *éxito* and English *exit* or Spanish *remover* and English *remove*. At times there may only be partial semantic identity, as Odlin (1989: 79) explains, e.g. Spanish *suceder*, and English *succeed*.

When using either false cognates or false friends, learners presume that there is both formal and semantic similarity between the familiar L1 form and the TL, and negative transfer results. Transfer can also be caused due to what the learner takes to be semantic equivalence between words, i.e. **He bit himself in the language* (in Spanish the word *lengua* means both *tongue* and *language*).

In the following case, both R2 and R6 considered deviant the expression used by the learner, *contract of employment*, and a more 'Englishy' sort of expression was proposed. These two cases do not coincide with their tagging of the 'error'. With the other researchers there was agreement on the non-marking of this expression as an error.

R2
(LP) contract of employment $work contract$.

R6
(LSF) contract of employment $working contract$

Another interesting case is that of the word 'subsist' as used by the learner in **they look for other way to subsist.*

R1, R2, R3
to subsist

R4
(LS) subsist $survive$.

R5
(R) subsist $survive$.

R6
(LSF) subsist $survive$.

Three groups coincided in classifying 'subsist' as an error. In one case, R5 considered its use to be related to a problem of register. R4 and R6 classified this word as an error, one group as a false friend and the other as a lexical single error. It is thought to be an error as there is a word with the same form that exists in Spanish, and these researchers most likely thought it did not actually exist in English. However, it does, but it would not normally be used in this context and by an intermediate learner of English; therefore, it may well be an error of register as indicated by R5.

12.4 Conclusions

Although there was agreement regarding detection of the errors made by the students and in their correction, there was also great disparity in the classification of these errors, this being most apparent with the article errors and lexical errors.

The first stage of detection of errors is achieved, logically, by comparing what was said or written with what the researcher thinks the learner meant to say or write. Corder (1981) explains it as follows: 'We identify errors by comparing original utterances with what I shall call reconstructed utterances, i.e. correct utterances having the meaning intended by the learner' (1981: 37). This idea of intended meaning has been the subject of great debate in the literature referring to the analysis of errors. How can we be sure of what the learner meant to say? Having analysed in our research work a 40,000-word corpus of learner language, we can confirm that both the linguistic context and the topic sequence were decisive in helping to detect and classify those errors which could be classed as problematic, and it must be noted that only a very small percentage proved to be of this nature. In the cases in our corpus where non-native-like language was used, and it was difficult to pinpoint the exact nature of the error, the tags (S) for Style and its subcategories (SI = style incomplete, and SU = style unclear) were used.

Another important factor related to the detection and classification of errors concerns the judge's knowledge of the learner's L1. The more familiar she/he is with the nuances of the language and culture, the more likely a correct interpretation of meaning can be achieved.

It is also necessary to point out that the researchers do not always have a clear vision of the error typification so that a simplified, easy-to-use and clear version of the tags list is desirable. It may be useful to give more examples in the error tagging manual of the different errors in order to clarify the doubts the evaluators have concerning classification.

We may also add that, as the corrected versions of the errors do coincide on many occasions, we understand that the researchers probably need more practice with the actual error-tagging process.

During the process of our research, we observed differences among raters regarding both tagging and corrections. Future studies will organize the evaluators into NS and NNS groups in order to study in greater depth the variables involved in error correction and the similarities and differences to be found between these two groups.

Notes

[1] We gratefully acknowledge the support given to the DIAAL research group (Project UPV PPI-06-04) by the Research Incentive Programme at the Universidad Politécnica de Valencia, Spain, for the opportunity to carry out the research work described in this chapter.

[2] An MS Windows programme which does not carry out an automatic analysis of errors but helps to simplify the classification and tagging of the IL by the analyst.

It was developed by researchers at the Centre for English Corpus Linguistics (CECL) at the Université Catholique de Louvain in Belgium.

[3] It was noted that this strategy was used on many occasions throughout the learners' texts.

References

Corder, S. P. (1981), *Error Analysis and Interlanguage.* Oxford: Oxford University Press.

Dagneaux, E., Denness, S., Granger, S. and Meunier, F. (1996), *Error Tagging Manual Version 1.1.* Louvain-la-Neuve, Belgium: Centre for English Corpus Linguistics, Université Catholique de Louvain.

Granger, S. (1993), 'International corpus of learner English', in Arts, J., De Haan, P. and Oostdijk, N. (eds), *English Language Corpora: Design, Analysis and Exploitation.* Amsterdam: Rodopi, pp. 57–71.

Granger, S. (ed) (1998), *Learner English on Computer.* London: Longman.

—(2003). 'The Corpus approach: A common way Forward for Contrastive Linguistics and Translation Studies?' in Granger, S., Lerot, J. and Petch-Tyson, S. (eds), *Corpus-Based Approaches to Corpus Linguistics and Translation Studies.* Amsterdam and Atlanta: Rodopi, pp. 17–29.

MacDonald, P. (2004), *An Analysis of Interlanguage Errors in Synchronous/Asynchronous Intercultural Communication Exchanges.* València: Universitat de València.

Moss, G. (1992), 'Cognate recognition: Its importance in the teaching of ESP reading courses to Spanish speakers'. *English for Specific Purposes,* 11, 141–158.

Nattinger, J. R. and De Carrico, J. S. (1992), *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Odlin, T. (1989), *Language Transfer.* Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G. and Startvik, J. (1972), *A Grammar of Contemporary English.* Harlow: Longman.

Swan, M. and Smith, B. (eds) (1987), *Learner English: A Teacher's Guide to Interference and Other Problems.* Cambridge: Cambridge University Press.

Tagnin, S. E. (2001), *COMET – A Multilingual Corpus for Teaching and Translation. PALC '01.* Frankfurt am Mein: Peter Lang, pp. 535–540.

Appendix I

You have 30 minutes to write an essay on the topic of immigration. Essay length: 15–25 lines (approximately 150–250 words). Make sure you write your name, your School/Faculty, the degree course and/or specialism, and your language teacher's name.

Immigration

In a world where nearly every day we hear reference made to the phenomenon of globalization and the free movement of capital, many people ask why the inhabitants of this world do not also have the fundamental right to emigrate and move around freely in order to improve the quality of their lives. On the other hand, many others think that immigration (both legal and illegal) is a problem; that the immigrants take our jobs; that crime increases; that our socio-cultural and religious traditions are threatened . . .

Discuss this issue, taking into account both sides of the phenomenon as presented here.

Appendix II

Since some years, in Spain and a lot of countries more, the number of immigrants is increasing. The immigrants are a problem but, also, they are a benefit for the city.

In the one hand, the immigrants emigrate because they want a better life, for they and for their families. It is one of the most important problems. Some people in Spain think the immigrants come to our country to get our jobs, but this people don't understand that the 90 per cent of all the immigrants that come to other country is for change his miserable life. An advantage is the cultural wealth. A city as Valencia, e.g., now, it has an enormous cultural diversity, Colombians, Argentineans, Romans or Africans are only a few examples of the big cultural diversity caused by the increase of the immigrants.

In the other hand, the immigrants that arrive to a new country and not find a job, they look for other way to subsist. They begin with little thefts, and later, they steal in house or stores. This fact cause a lot of people begin to suspicious about them.

I think the immigrants are a solution for some problems, as the decrease of the birthrate, and they contribute with an important improvement of the culture. Although, the immigrant's flow must be controlled and all of the immigrants that come to Spain or other country, they must have contract of employment.

Chapter 13

Focus on Errors: Learner Corpora as Pedagogical Tools

Amaya Mendikoetxea, Susana Murcia Bielsa and Paul Rollinson
Universidad Autónoma de Madrid

13.1 Introduction

As it has often been pointed out, there are two ways in which corpora can be exploited in language teaching (Aston 2000): (i) identifying features to be taught and developing materials to teach them, and (ii) as resources for autonomous language learning. The INTELeNG project[1] belongs to the first of these. In constructing a database of errors extracted out of a corpus, our purpose is to identify problematic areas, to evaluate their level of difficulty and to develop relevant learning materials. This chapter offers a general description of the project, its motivation, objectives and its current state of development. Our purpose is to show that a small learner corpus can help (i) develop pedagogical materials which are appropriate for particular learners, and (ii) improve curriculum design (through the selection and sequencing of grammatical phenomena). We will first outline some of the ideas that motivate the approach adopted (sections 13.2 to 13.4), and will then describe the project: the learners (subsection 13.5.1), the corpus (subsection 13.5.2), the database (subsection 13.5.3) and the pedagogical materials (subsection 13.5.4).

13.2 Conceptual Motivation for a Database of Errors for Pedagogical Purposes

The conceptual motivation for the INTELeNG project is to be found in the ideas behind contrastive analysis (CA) and error analysis (EA), and in the resurgent interest in these areas as a consequence of the availability of both learner and native corpora. CA was the favoured paradigm for second and foreign language teaching and learning in the 1950s and 1960s

(Lado 1957). The general idea behind this approach was that difficulties in learning were associated with differences in structure between the mother tongue (MT) and the target language (TL). The way it proceeded was by describing and comparing features of MT and TL, in order to formulate predictions about the areas that could cause interference and error. By the early 70s, CA was discredited partly as a result of its association with structuralism and behaviourism, but also because its predictions were thought to be unreliable.

E. A. (Corder 1967) was the paradigm to replace CA. It proceeded by describing the learner's Interlanguage (IL) and the TL and then comparing the two in search for mismatches, without referring to the MT. It went out of fashion in the 1980s, with the advent of the communicative approach, but some contend that EA became, and is still, a much more widespread practice than it is given credit for. James (1998: 19), for instance, suggests that 'EA has never been abandoned, but has rather lain in the doldrums perhaps awaiting the signal to ply the main' (see James 1998 for the historical and theoretical background for EA).

We will define *error* in a very loose way as an 'unsuccessful bit of language' (James 1998: 1). In the literature errors are to be distinguished from mistakes. While errors are taken to be 'overt manifestations of learners' systems' (Brown 1987: 171), mistakes are performance deviances which are self-correctible. In a study of the type we are conducting, in which there is no direct access to the author of the text, the distinction between the two is blurred. This is why we will use the very general definition given above, though we are aware of the fact that it needs much refinement.

A fully contrastive approach is adopted which brings CA and EA together, and involves MT vs. TL and IL vs. TL comparisons. It is also useful to compare the learners' MT (Spanish, in this case) and their IL, to see how the features of the MT are present in the IL, as research shows that a significant number of errors are interlingual (due to MT interference). On the other hand, the motivation for the focus on error approach is the idea that learning proceeds by learners comparing both their IL and TL ('something that learners are naturally inclined to do but often need teacher guidance in doing effectively', quoting James 1998: 258), as well as by comparing their MT and their TL. Our hypothesis is that by having a good understanding of learners' difficulties, teachers and teaching materials can help students become better error analysts and contrastive analysts by fostering language awareness, with the ultimate purpose of promoting proficiency. It is in connection with this that a learner corpus can be used as a powerful pedagogical tool.

13.3 The role of learner corpora in relation to EA and CA

There is an increasing interest in learner corpora both as a pedagogical tool (to assist IL development), as well as a research tool (for the analysis of the features of IL). Learner corpora have been used mostly to provide information on learners' common errors, as in the *Longman Dictionary of Common Errors* (based on the Longman Learner Corpus, Heaton and Turton 2001) and the more recent *Macmillan English Dictionary for Advanced Learners* (2007), developed in collaboration with the Centre for English Corpus Linguistics (CECL) at the Université Catholique de Louvain, and based on ICLE (International Corpus of Learner English, Granger et al. 2002a) (see Nesselhauf 2004 for a state of the art account of the use of learner corpus for teaching purposes). But though it can be used for pedagogical purposes, ICLE (and earlier corpora like COALA, Pienemann 1992) is mainly designed for research purposes together with many other academic and even commercial corpora (for an overview of learner corpora from a research perspective see, among others, Granger 2002, 2004; Barlow 2005; Myles 2005).[2]

As Granger (2002: 11–12) points out, linguistic exploitation of learner corpora may involve one of the following two methodological approaches: (i) Contrastive Interlanguage Analysis (Granger 1996; Gilquin 2001), involving quantitative and qualitative comparisons between (a) native and non-native data or (b) different varieties of non-native data, from learners with different mother tongue (see, among many others, Lozano and Mendikoetxea 2008), and (ii) Computer-aided error analysis, focusing on errors in IL and using computer tools to retrieve them. The use of computer tools has allowed researchers to handle vast corpora, giving a new dimension to both traditional EA and CA. But there is also a growing number of researchers and practitioners who are collecting their own smaller corpora to cater for the needs of a particular group of learners (see some of the papers in Hidalgo et al. 2007). INTELeNG is an example of the latter. By creating our own corpus and identifying and classifying errors in our database, we are able to design pedagogical materials which are more 'locally' oriented for learners of a particular MT in a particular context.

Learner corpora may also be used for classroom methodology (exploiting the corpus in class for inductive learning) and as the basis for curriculum design (selection and sequencing of grammatical phenomena), an area on which learner corpora have had, so far, very little or no impact (see, for instance, Aston 2000; Meunier 2002; Hunston 2002). An advantage of using learner data for these purposes is that it is 'authentic' data, though as Granger (2002: 8) says, the notion of authenticity is somewhat problematic with reference to learner data. Under Sinclair's (1996) definition of 'authentic' data is that gathered in real communication situations, with people going about their normal business (vs. experimental data). In that sense, learner data is very rarely authentic, and learner corpora are often 'experimental'. Granger (2002) defines learner data as 'authentic' when it is data resulting from 'authentic' classroom activity, as in the case of the written essays collected for our corpus.

It must be said, though, that while we strongly advocate the use of learner corpora for teaching and learning activities, the combination of both learner and native corpora is essential in our view: learner corpora show the gap between the learners' IL and their TL, but the features of the TL are fully present in the native corpora. Our pedagogical materials combine both.

13.4 On grammar and formal instruction

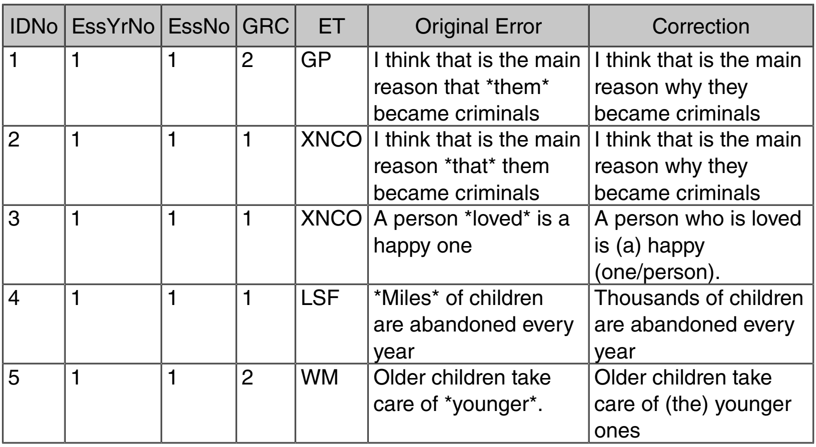
Before we describe the INTELENG project in some detail, it is necessary to say a few words about the role of grammar instruction in language learning. Though the value of EFL grammar teaching has been much debated, there seems to be a growing consensus that grammar must play a larger role in the classroom than that granted by the communicative approaches of the 1980s (see, for instance, Hawkins and Towell 1996). A large number of empirical studies have shown that drawing students' attention to form (explicit teaching, corrective feedback and so on) gives better results than implicit learning (see, for instance, Hulstijn and Hulstijn 1984; Rutherford 1987; Fotos 1993 and Robinson 1996) and can help avoid fossilization and pidginization (Harley 1993).

There remains, however, considerable disagreement about how grammar should be taught, as illustrated by the focus-on-form vs. focus-on forms debate. Without entering that debate, we agree with an increasing number of researchers and practitioners in the view that some sort of formal instruction is required for raising learner consciousness of grammatical structures in the TL in order to promote advanced level of TL attainment (Ellis 1990; Fotos 1993). According to Ellis (1995), attention to form facilitates 'noticing'; learners can accelerate their acquisition by noticing the gap between their own TL forms ('no noticing, no acquisition' (Ellis 1995: 98)). The kind of grammar teaching we believe is appropriate for our goals and methodology is based on the idea of *consciousness-raising* (Rutherford 1987) and *language awareness* (James and Garrett 1991).[3] Being aware of the forms one uses in the MT can help monitor transfer into the TL (Joyce and Burns

1999: chapter 3), while consciousness-raising may assist the language learning process by providing data through which the learners can test their hypotheses in order to confirm them or disprove them, or to formulate new hypotheses.

Presenting the students with the errors they, or other co-learners, make and asking them to correct them is an activity appropriate for raising grammatical consciousness in an attempt to bridge the gap between what is known and the forms of the TL. A database of errors like that of the INTELeNG project (see Figure 13.1 below) lends itself naturally to that kind of activity. Though researchers have emphasized the value of error correction (see, for instance, Ellis 1995 and Rutherford 1987), there are, however, those who point out the risks of exposing learners to printed errors produced by themselves or their peers. In our experience, as long as the errors are appropriately signalled, no such confusion arises. Students seem to find error correction exercises motivating and fun, and they can get quite skilled at them, though whether they can then transfer that knowledge to their output is a different matter.

Before concluding this section, it has to be emphasized that not all learners have the same need for grammatical instruction. The type of instruction we have specified here is particularly suited to the type of learner for whom the INTELeNG project has been designed (see section 13.5.1 below).



IDNo	EssYrNo	EssNo	GRC	ET	Original Error	Correction
1	1	1	2	GP	I think that is the main reason that *them* became criminals	I think that is the main reason why they became criminals
2	1	1	1	XNCO	I think that is the main reason *that* them became criminals	I think that is the main reason why they became criminals
3	1	1	1	XNCO	A person *loved* is a happy one	A person who is loved is (a) happy (one/person).
4	1	1	1	LSF	*Miles* of children are abandoned every year	Thousands of children are abandoned every year
5	1	1	2	WM	Older children take care of *younger*.	Older children take care of (the) younger ones

Figure 13.1 Sample entries in the INTELeNG database

13.5 The INTELeNG Project

13.5.1 Our learners and their context

Our learners are first year university students of English Philology at the Universidad Autónoma de Madrid (Spain) – a four year degree with courses in English language, literature and linguistics, in which English is the medium of instruction and evaluation. Thus, they belong to a group with very specific needs – the non-native student in a continental English department. These learners are atypical regarding the length of the period of formal instruction and the level of excellence they are expected to achieve (see Mair 2002). While in other contexts error only tends to matter as long as it impedes communication, correctness and accuracy are highly valued in an academic context such as the one we are describing. English, as well as the TL, is also an object of study. Explicit reference to grammar rules and concepts is expected to play a significant role in the language class, and in the third and fourth year of their degree, students have specific courses in English grammar and syntax.

13.5.2 The corpus

The essays from which the errors have been extracted are examination papers of the Academic Writing Component of the 'English Language I' course, which is divided into 5 components: Listening, Reading, Academic Writing, Vocabulary and Grammar. In the Academic Writing component, students learn to write short academic essays, with emphasis placed on proper organization according to academic conventions, and on the presentation, development, support and analysis of arguments. In the examination, students are given two hours to write an essay of a minimum of 500 words on current topics (e.g. domestic violence, homosexual marriages).

13.5.3 The database

Our database consists so far of a collection of all the grammar and lexis errors found in around 80 essays. The data collection was done manually, by marking the errors on the exam papers and then typing them into a Microsoft Access database specifically created for our purposes. It includes around 2,000 errors, and contains the following information (see Figure 13.1):

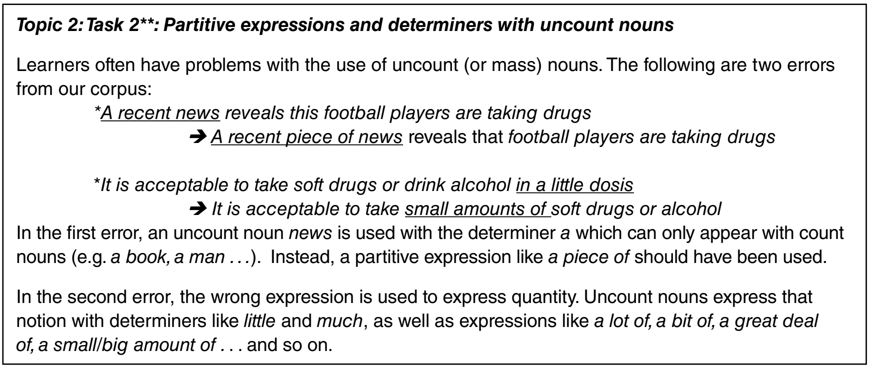
- Error identification number (IDNo)
- Essay year (EssYrNo)[4]
- Essay number (EssNo), in order to enable us to refer back to the original if necessary (for example, to provide a wider context to

understand the error, to check whether the wrong spelling is an error or a typo, etc.)

- Grammatical category (GRC) to which the error belongs. Here we classified errors linguistically, according to a taxonomy which includes 11 grammatical categories. Ten of these categories correspond to the grammar topics covered in each of the ten 2-hour sessions of the Grammar Component within the course, and the eleventh category contains errors that do not fit into any of the other ten categories (e.g. word order, subject missing, style problems).[5]
- Error Type (ET): a further linguistic classification of errors using the error tags employed in the ICLE project (see Dagneux et al. 1998).[6]
- Original Error, which includes the segments of text containing errors;[7]
- Correction, which gives the corrected text.

13.5.4 The pedagogical materials

The Database has been used as a source for elaborating pedagogical materials for the Grammar component of the 'English Language I' course (a 20 hour component spread over 5 weeks and covering the grammar topics 1–10 given in footnote 5). We have elaborated tasks for each of the 11 grammatical categories in the database. Tasks are graded with one, two or three stars depending on their degree of difficulty, which we have calculated on the basis of our experience. This is to cater for different learners as we have mixed-ability classes. Tasks are divided into two main parts: (1) the errors and comments, and (2) the exercises. The starting point is always an error made by our students: taken from the INTELeNG database. Underlining is used in Figure 13.2 for an incorrect word or phrase that has to be changed



Topic 2: Task 2**: Partitive expressions and determiners with uncount nouns

Learners often have problems with the use of uncount (or mass) nouns. The following are two errors from our corpus:

**A recent news reveals this football players are taking drugs*

➔ *A recent piece of news* reveals that *football players are taking drugs*

**It is acceptable to take soft drugs or drink alcohol in a little dosis*

➔ *It is acceptable to take small amounts of soft drugs or alcohol*

In the first error, an uncount noun *news* is used with the determiner *a* which can only appear with count nouns (e.g. *a book, a man . . .*). Instead, a partitive expression like *a piece of* should have been used.

In the second error, the wrong expression is used to express quantity. Uncount nouns express that notion with determiners like *little* and *much*, as well as expressions like *a lot of, a bit of, a great deal of, a small/big amount of . . .* and so on.

FIGURE 13.2 Task Sample 1: error and comments

or replaced by another. Each error is then followed by the corrected version. The error and correction are then followed by comments about the possible sources of the error, brief grammatical explanation, explanation of the error, comments on usage, etc., as appropriate.

In providing the comments, our goal is to assist interlanguage development, to help learners distinguish what is correct or appropriate use of the TL structures and what is incorrect or inappropriate use. As Osborne (2000) points out, providing grammatical explanations to language learners involves some assumption about: (a) the role that explanation may play in language learning, (b) how technical explanations should be, and (c) theoretical background.

The comments in Figure 13.2 assume knowledge of the distinction between count and uncount nouns. In a context like ours, 'external' knowledge of the type provided by grammatical explanation is expected by learners and can assist them in organizing their knowledge about the grammar of the TL. Since these are students who have had a fairly thorough grounding in descriptive grammar of a traditional type (in English and Spanish) and are familiar with the terms used in most pedagogical grammars, some technical terms are included in the comments (noun phrases, partitive expressions, degree adverbs and so on). The theoretical background is 'neutral' but coherent – using terminology and ideas typical of traditional descriptive grammar that the students are familiar with.

Another example is provided in Figure 13.3, where we have provided reference to the possible source of the error – putatively intralingual. This is followed by a usage comment. Overuse of 'very' and lack of finesse in using intensifiers (or downtoners) is an error that has often been pointed out in the bibliography (see, for instance, Lorenz 1998; Flowerdew 2000).

Topic 4- Task 1. Degree adverbs: use of very***

A typical error of learners is to use *very* in contexts where other degree adverbs or other expressions should be used.

**This has been very commented on*
≡ This has been much commented on
**This point is very related to the previous one*
≡ This point is closely related to the previous one

These errors may be influenced by the fact that the Spanish adverb *muy* is more widely used than its English equivalent *very.* Thus, if we translate these sentences into Spanish we would use *muy* with *comentado (commented)* or *relacionado (related)*. In English, instead, the adverb *much* and degree adverbs like those below should be used:

highly	strongly	hugely	fantastically	intensely

FIGURE 13.3 Task Sample 2: error and comments

Exercise 1: Collocations with degree adverbs

The following adjectives, verbs and participles are often used with the adverbs like *highly, strongly* and so on. On the basis of these examples, try to identify the meaning that the adverb adds to the verbal/adjectival elements it appears with. Can you observe any patterns: i.e. semantic types of verbal or adjectival elements which appear with particular adverbs?

closely	related, linked, spaced, surrounded, interlinked, followed
highly	respected, recommended, regarded, respected, specialized, talented, critical, possible, productive, protective, educational, intelligent, significant
strongly	advised, condemned, criticized, denied, favoured, preferred, object, supported, competitive
greatly	admired, affected, attracted, desired, enhanced, helped, pleased, surprised, perturbed, encouraged
hugely	amused, impressed, pleased, underrated, popular, complex, successful, satisfying, expensive
largely	disregarded, ignored, limited, undetected mysterious, responsible
intensely	committed, concentrated, focused, involved, dramatic, emotional, irritating, moving, personal, proud, religious, (un)popular
extensively	damaged, explored, pursued, read

Figure 13.4 Sample of discovery learning activity

No general rule or explanation can be given in the comments since the reasons that guide the choice of a specific adverb for a particular adjective or vice versa are hard to pinpoint (apart from collocability) (see Partington 1998; Bernardini 2002).

The errors and the comments are followed by a variety of exercises of both a deductive and inductive type involving common exercises such as error correction, cloze, rewriting and so on as well as data-driven (Johns 1990) and discovery learning type of activities (McEnery and Wilson 1997) (based on concordances from native corpora). An example of the latter is the exercise in Figure 13.4, from a task on the use of degree adverbs, where learners are presented with a variety of adjectives typically found with certain intensifiers (on the basis of examples extracted from the Cobuild Concordance Sampler at http://www.collins.co.uk/Corpus/CorpusSearch.aspx) and are asked to identify collocation patterns.

Translation exercises are used as part of a contrastive approach which seeks to emphasize those grammatical points in which the grammars of English and Spanish show degrees of discrepancy. We make extensive use of translation exercises like that in Figure 13.5 (which was prepared for the task in Figure 13.2), and the final test for the course contains an exercise in which they have to translate six sentences into English.[8]

While the translation exercises emphasize the differences between the learner's MT and the TL, error-based exercises, which we have included for other topics, emphasize the differences between the learners' IL and

Exercise: Partitive expressions

Translate the following sentence into English using the noun in parentheses. Use partitive expressions when appropriate.

<u>*Example:*</u> *Me dió un buen consejo* (advice)
➔ *He gave me* ***some*** *good advice*
He gave me ***a*** *good* ***piece of*** *advice*

1. Tuvimos un **tiempo** buenísimo en vacaciones el año pasado. (weather)
2. Los trabajadores causaron **daños** en el **material** y tuvieron que pagarlos. (damage, equipment)
3. Compramos un **mueble** muy bonito para nuestro piso nuevo. (furniture)
4. Se hizo mucho **daño** a las perspectivas de paz. (harm)
5. Nos compraron una **cubertería** muy bonita como regalo de bodas. (cutlery)

Figure 13.5 Sample of translation exercise

Exercise: Some errors in the use of passive sentences

Some of the following sentences contain one or more errors. Identify the errors and correct them:

1. To sum up, in this essay it was discussed some possible solutions
2. Linguistics is taught in some universities in Britain and Spain.
3. His articles are referred to by relevant academics
4. Also can be said that for a single parent to bring up a child it's a very hard work
5. This bench has been sat on all day
6. John is resembled by his brother
7. The prize was given me by the jury
8. This bed was slept in by Queen Victoria
9. John is rumoured to be a millionaire
10. It was believed the letter to be a forgery

Figure 13.6 Sample of error identification and correction exercise

the TL. A database of errors like the one we have compiled lends itself naturally to this type of exercise, which some researchers have pointed out is most effective when learners are asked to correct typical errors of learners of the group they belong to (with the same L1) (see Rutherford 1987; Ellis 1995). An example is given in Figure 13.6, from a task on passive structures, in which learners are asked to identify the sentences containing errors and correct them.

13.6 Conclusions

We hope to have shown how a small learner corpus (and its associated database) can assist the development of pedagogical materials specifically suited for a particular learner group. Learner corpora may be used not only for the elaboration of teaching materials, but also as part of classroom methodology (exploiting the corpus in class for inductive learning) and as the basis for curriculum design (selection and sequencing of grammatical phenomena). Regarding classroom methodology, discovery learning type of activities of the type developed, for example, by Bernardini (2000, 2002) in which 'learners browse corpora much in the same way as they would explore an unknown land' (Bernardini 2002: 166) do not fit well within a 20-hour grammar course with first year undergraduates, but we realize that we must make an effort to encourage this type of exploratory learning at perhaps a smaller scale. As for curriculum design, learner corpora have so far had little or no impact (see Aston 2000; Meunier 2002). Perhaps as more data is gathered for the INTELeNG project we will be able to identify clear areas of difficulty and influence the way topics are presented in the course. The objective of all this is to gain a better understanding of learner difficulties and how to help learners overcome them by designing materials which focus on their errors.

Notes

[1] INTELeNG stands for 'Innovación tecnológica para la enseñanza/aprendizaje de las lenguas'. This project has been partly funded by the Universidad Autónoma de Madrid and is part of a general program designed to encourage innovation in teaching practices. Funding by the Spanish Ministry of Education and Science (research grant HUM2005–01728) is also gratefully acknowledged.

[2] Granger (2002) classifies computer learner corpora into 'academic' and 'commercial'. The latter includes the Longman Learner Corpus (10 million words), mentioned above, and the Cambridge Learner Corpus (16 million words). Academic corpora are smaller in size, the largest being The Hong-Kong University of Science and Technology Learner Corpus (25 million words). ICLE contains 2.5 million words. Together with the INTELeNG corpus described here, we are currently involved in the compilation of two learner corpora (WriCLE- Written Corpus of Learner English and CEDEL2-Corpus Escrito del Español como L2) under a research project which seeks to identify word order patterns in L2 English (L1 Spanish) and L2 Spanish (L1 English) (see http://www.uam.es/proyectosinv/woslac/). It is worth emphasizing the commercial corpora can be used for academic purposes and likewise academic corpora can be used for commercial purposes.

[3] Though these two approaches are often taken to be synonymous, James (1998: 260) proposes a way of distinguishing the two concepts. For this author, language awareness is a 'learned ability to analyse one's own repertoire – be they in the L1 or in that part of the TL that one has learned so far', while consciousness-raising

refers to 'getting explicit insight into what one does not yet know implicitly of the TL in order to identify the discrepancy between one's present state of knowledge and the target knowledge'.

[4] So far, our database includes only first year essays, but we intend to include essays from other years in the future.

[5] Grammatical categories: (1) Nouns and Noun Phrases; (2) Pronouns and Determiners; (3) Adjectives and Adjective Phrases; (4) Adverbs and Prepositions; (5) Verbs and Verb Phrases (I): Tenses and Modals; (6) Verbs and Verb Phrases (II): Phrasal Complements of Verbs; (7) Complex Sentences; (8) Passive and Causative Constructions; (9) Reported Speech; (10) Conditionals, and (11) Other Problems.

[6] In Figure 13.1, **GP** (error 1) stands for **G**rammar, **P**ronoun (meaning here wrong choice of pronoun); **XNCO** (error 2) stands for Le**X**ico-Grammar, **N**ouns, **Co**mplementation: there is an error in the complement of a noun; **LSF** (error 4) stands for **L**exical **S**ingle, **F**alse Friends: in this case, the error is due to MT interference, since Spanish *miles* means *thousands* in English; and **WM** (error 5) stands for **W**ord **M**issing.

[7] Where one segment of text contained more than one error, the segment was repeated as many times as errors it contained, and each of the errors was classified independently.

[8] Though translation into TL is a much more common exercise for learners, we have also used in some tasks exercises involving translation into MT, in an effort to promote awareness of the differences between the MT and the TL.

References

Aston, G. (2000), 'Corpora and language teaching', in Burnard, L. and McEnery, T. (eds), *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt am Main: Peter Lang, pp. 7–18.

Barlow, M. (2005), 'Computer-based analysis of learner language', in Ellis, R. and Barkhuizen, G. (eds), *Analysing Learner Language.* Oxford: Oxford University Press, pp. 335–354.

Bernardini, S. (2000), 'Systematizing serendipity: Proposals for concordancing large corpora with language teachers', in Burnard, L. and McEnery, T. (eds), *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt am Main: Peter Lang, pp. 225–234.

—(2002), 'Exploring new directions for discovery learning', in Kettemann, B. and Marko, G. (eds), *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora.* Amsterdam: Rodopi, pp. 165–182.

Brown, H. (1987), *Principles of Language Learning and Teaching.* Englewood Cliffs: Prentice Hall.

Corder, S. P. (1967), 'The significance of learner's errors'. *International Review of Applied Linguistics*, 5, (4), 161–170. Reprinted in Corder, S. P. (1981), *Error Analysis and Interlanguage.* Oxford: OUP, pp. 14–25.

Dagneux, E., Denness, S. and Granger, S. (1998), 'Computer-aided error analysis. system'. *An International Journal of Educational Technology and Applied Linguistics*, 26, (2), 163–174.

Ellis, R. (1990), *Instructed Second Language Acquisition: Learning in the Classroom.* Oxford: Blackwell.

—(1995), 'Interpretation tasks for grammar teaching'. *TESOL Quarterly*, 29, (1), 87–105.

Flowerdew, L. (2000), 'Investigating referential and pragmatic errors in a learner corpus', in Burnard, L. and McEnery, T. (eds), *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt am Main: Peter Lang, pp. 145–154.

Fotos, S. (1993), 'Consciousness raising and noticing through focus on form: Grammar task performance versus formal instruction'. *Applied Linguistics*, 14, (49), 385–407.

Gilquin, G. (2001), 'The Integrated Contrastive Model. Spicing up your data'. *Languages in Contrast*, 3, (1), 95–123.

Granger, S. (1996), 'From CA to CIA and back. An integrated approach to computerized bilingual and learner corpora', in Aijmer, K. B., Altenberg, B. and Johansson, M. (eds), *Languages in Contrast. Papers from a Symposium on Text-Based Cross-Linguistic Studies.* Lund: Lund University Press, pp. 37–51.

—(2002), 'A bird's eye view of learner corpus research', in Granger, S., Hung, J. and Petch-Tyson, S. (eds) (2002b), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Amsterdam: John Benjamins, pp. 3–36.

—Granger, S. (2004), 'Computer learner corpus research: Current status and future prospects', in Connor, U. and T. A. Upton (eds), *Applied Corpus Linguistics: A Multidimensional Perspective.* Amsterdam and New York: Rodopi, pp. 123–145.

Granger, S., Dagneux, E. and Meunier, F. (eds) (2002a), *International Corpus of Learner English.* Louvain: Presses Universitaires de Louvain.

Harley, B. (1993), 'Instructional strategies and SLA in early French immersion'. *Studies in Second Language Acquisition*, 15, (2), 245–259.

Hawkins, R. and Towell, R. (1996), 'Why teach grammar', in Engels, D. and Myles, F. (eds), *Teaching Grammar: Perspectives in Higher Education.* London: CILT, pp. 195–211.

Heaton, J. B. and Turton, N. D. (eds) (2001), *Longman Dictionary of Common Errors.* (2nd ed). London: Longman.

Hidalgo, E., Quereda, L. and Santana, J. (2007), *Corpora in the Foreign Language Classroom: Selected papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6). University of Granada, Spain, 4–7 July, (2004).* Amsterdam: Rodopi.

Hulstijn, J. and Hulstijn, W. (1984), 'Grammatical errors as a function of processing constraints and explicit knowledge'. *Language Learning*, 35, 23–43.

Hunston, S. (2002), *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

James, C. (1998), *Errors in Language Learning and Use: Exploring Error Analysis.* Harlow: Longman.

James, C. and Garret, P. (eds) (1991), *Language Awareness in the Classroom.* London: Longman.

Johns, T. (1990), 'From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning'. *CALL Austria,* 10, 14–34. Revised version in Odlin, T. (1994), *Perspectives on Pedagogical Grammar.* Cambridge: Cambridge University Press, pp. 293–313.

Joyce, H. and Burns, A. (1999), *Focus on Grammar.* Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Lado, R. (1957), *Linguistics across Cultures.* Ann Arbor, MI: University of Michigan Press.

Lorenz, G. (1998), 'Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification', in S. Granger (ed.), *Learner English on Computer.* London: Longman, pp. 53–66.

Lozano, C. and Mendikoetxea, A. (2008), 'Postverbal subjects at the interfaces in English and Italian learners of English: A corpus study', in B. Díaz, G. Gilquin and S. Papp (eds), *Linking up Contrastive and Learner Corpus Research.* Amsterdam: Rodopi, pp. 85–125.

Macmillan (2007), *Macmillan English Dictionary for Advanced Learners.* (2nd ed). Oxford: Macmillan, pp. 85–125.

Mair, C. (2002), 'Empowering non-native speakers: The hidden surplus value of corpora in continental English departments', in B. Kettemann and G. Marko (eds), *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora.* Amsterdam: Rodopi, pp. 119–130.

McEnery, T. and Wilson, A. (1997), 'Multimedia corpora', in B. Lewandowska-Tomasczczyk and P. J. Melia, (eds), *Palc '97 Practical Applications in Language Corpora.* Lódz: Lódz University Press, pp. 24–33.

Meunier, F. (2002), 'The pedagogical value of native and learner corpora in EFL grammar teaching', in S. Granger, J. Hung, and S. Petch-Tyson (eds), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Amsterdam: John Benjamins, pp. 119–141.

Myles, F. (2005), 'Review article. Interlanguage corpora and second language acquisition research'. *Second Language Research,* 21, (4), 373–391.

Nesselhauf, N. (2004), 'Learner corpora and their potential for language teaching', in J. Sinclair (ed.), *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins, pp. 125–156.

Osborne, J. (2000), 'What can students learn from a corpus? Building bridges between data and explanation', in L. Burnard and T. McEnery (eds), *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt am Main: Peter Lang, pp. 165–172.

Partington, A. (1998), *Patterns and Meanings: Using Corpora for English Language Research and Teaching.* Amsterdam: John Benjamins.

Pienemann, N. (1992), 'COALA – A computational system for interlanguage analysis'. *Second Language Research,* 8, 59–92.

Robinson, P. (1996), *Consciousness, Rules and Instructed Second Language.* New York: Peter Lang.

Rutherford, W. (1987), *Second Language Grammar: Learning and Teaching.* London: Longman.

Sinclair, J. (1996), *EAGLES. Preliminary Recommendations on Corpus Typology.* Pisa: Consiglio Nazionale delle Ricerche, Istituto di Linguistica Computazionale. Retrieved from: http://www.ilc.pi.it/EAGLES96/corpustyp/corpustyp.html

Chapter 14

The Monolingual Learners' Dictionary as a Productive Tool: The Contribution of Learner Corpora[1]

Sylvie De Cock and Magali Paquot
Centre for English Corpus Linguistics, Université catholique de Louvain

14.1 Introduction

This chapter reports on the outcomes of a large-scale collaborative corpus-based project between the Centre for English Corpus Linguistics and Macmillan Education. The main aim of this project was to produce learner corpus informed materials for inclusion in the second edition of the *Macmillan English Dictionary for Advanced Learners* (Rundell 2007; hereafter *MED2*).

Monolingual learners' dictionaries (MLDs) 'are dictionaries that are specially designed to cater for the needs of foreign language learners and provide all the information in the learner's target language' (De Cock and Granger 2004: 72). As highlighted by Rundell (1999), MLDs have reaped enormous benefits from the increased use of computer corpora in dictionary making. Corpus-related improvements (e.g. enhanced descriptions of word meanings and phraseological patterns) have so far mainly involved the use of corpora of native speaker speech and writing. The use of learner corpora (i.e. see De Cock, this volume) when compiling MLDs is comparatively relatively recent: the first learner corpus informed dictionary, the *Longman Language Activator*, was published in 1993 (Summers 1993). *The Longman Dictionary of Contemporary English* (Summers 1995), the *Longman Essential Activator* (Summers1997) and the *Cambridge Advanced Learners' Dictionary* (Gillard 2003) followed suit.

Findings from learner corpus-based research, which shed light on learners' difficulties and deficiencies, can be used by lexicographers to anticipate learners' errors (Rundell 1999). De Cock and Granger (2004) identify two types of learner corpus-based information that can help enhance MLDs: information relating to learners' misuse of target language words or phrases and information relating to learners' overuse and underuse of target language words or phrases. While the extraction of downright errors from learner corpora has greatly been facilitated by error annotation (De Cock and Granger 2004), cases of overuse and underuse can be uncovered by contrasting the frequency counts of words and phrases in a learner corpus and in a comparable native speaker corpus.

Information from learner corpora has overwhelmingly been included in the form of explicit 'warning' or 'common learner error' notes in MLDs. An investigation of these error notes in two learner corpus informed advanced MLDs (De Cock and Granger 2004) revealed that there was still room for improvement. It showed that, although some of the notes in the MLDs addressed corpus-attested frequently recurring advanced learner errors, a number of them were misguided as they focused on errors that were clearly inappropriate to the learners' advanced level of proficiency or on errors that were either rare or non-existent in a corpus of advanced learner writing like the *International Corpus of Learner English*. The study also highlighted a distinct lack of overlap between the errors targeted in the two dictionaries, which points to the problematic selection of the learner errors that should be focused on.

14.2 The *Macmillan English Dictionary for Advanced Learner*s (Second Edition, *MED2*)

14.2.1 Objective of the project

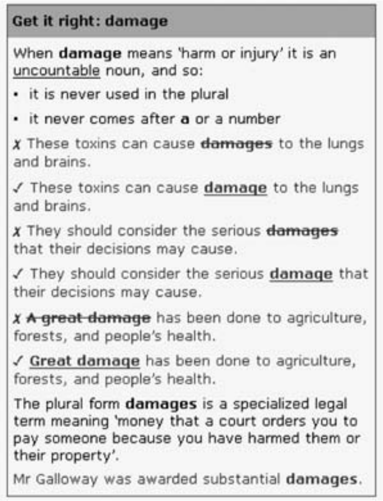
The main objective of the collaborative corpus-based project between the Centre for English Corpus Linguistics and Macmillan Education reported on in this chapter was to produce learner corpus informed materials that would help advanced learners cope with attested areas of difficulty in their writing. Three learner corpus-based components were developed with this objective in mind: 100 'Get it right' (GIR) boxes, six Grammar Sections and 12 EAP Writing Sections.

The three components were compiled on the basis of detailed analysis of (1) learner corpus data from the 3.5-million-word *International Corpus of Learner English* (*ICLE*), which contains essay writing by EFL learners from 16 different mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana, Turkish) and (2) native speaker data from a 15-million-word corpus of academic writing.

14.2.2 'Get it right' boxes

GIR boxes are located at individual dictionary entries. They are designed to alert learners to typical erroneous usage associated with a particular word or phrase. *MED2* is aimed at all advanced learners regardless of their mother tongue backgrounds. It was therefore decided that, to make the shortlist of flagged words or phrases, the errors associated with these items had to be shown to be both frequent and widespread (i.e. attested in data from learners from at least five different mother tongue backgrounds) in the learner corpus. Possible candidates for the shortlist were identified on the basis of, on the one hand, careful scrutiny of the 680,000-word error-tagged component of *ICLE* available at the time and, on the other, systematic comparisons of word frequency lists from the advanced learner corpus and the native speaker corpus used. Once identified, the possible candidates were subjected to rigorous analysis in the whole 3.5-word *ICLE* corpus. The errors in the GIR boxes cover a number of categories including 'countability', 'register', 'verb patterns' or 'spelling'.

As Figures 14.1, 14.2, 14.3 and 14.4 illustrate, GIR boxes typically contain authentic erroneous learner examples clearly marked as incorrect



Get it right: damage

When **damage** means 'harm or injury' it is an uncountable noun, and so:

- it is never used in the plural
- it never comes after **a** or a number

✗ These toxins can cause ~~damages~~ to the lungs and brains.

✓ These toxins can cause **damage** to the lungs and brains.

✗ They should consider the serious ~~damages~~ that their decisions may cause.

✓ They should consider the serious **damage** that their decisions may cause.

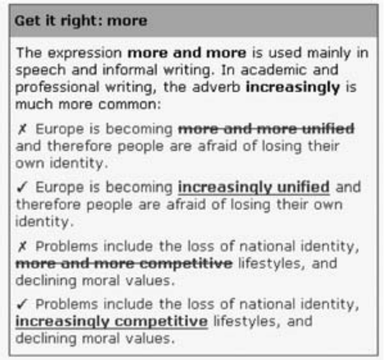
✗ ~~A great damage~~ has been done to agriculture, forests, and people's health.

✓ **Great damage** has been done to agriculture, forests, and people's health.

The plural form **damages** is a specialized legal term meaning 'money that a court orders you to pay someone because you have harmed them or their property'.

Mr Galloway was awarded substantial **damages**.

Figure 14.1 GIR box at 'damage' in *MED2*



Get it right: more

The expression **more and more** is used mainly in speech and informal writing. In academic and professional writing, the adverb **increasingly** is much more common:

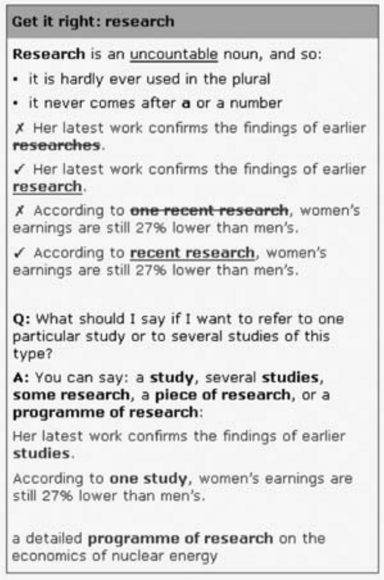
✗ Europe is becoming ~~**more and more unified**~~ and therefore people are afraid of losing their own identity.

✓ Europe is becoming **increasingly unified** and therefore people are afraid of losing their own identity.

✗ Problems include the loss of national identity, ~~**more and more competitive**~~ lifestyles, and declining moral values.

✓ Problems include the loss of national identity, **increasingly competitive** lifestyles, and declining moral values.

Figure 14.2 GIR box at ‘more’ in *MED2*



Get it right: research

Research is an uncountable noun, and so:

- it is hardly ever used in the plural
- it never comes after **a** or a number

✗ Her latest work confirms the findings of earlier ~~**researches**~~.

✓ Her latest work confirms the findings of earlier **research**.

✗ According to ~~**one recent research**~~, women’s earnings are still 27% lower than men’s.

✓ According to **recent research**, women’s earnings are still 27% lower than men’s.

Q: What should I say if I want to refer to one particular study or to several studies of this type?

A: You can say: a **study**, several **studies**, **some research**, a **piece of research**, or a **programme of research**:

Her latest work confirms the findings of earlier **studies**.

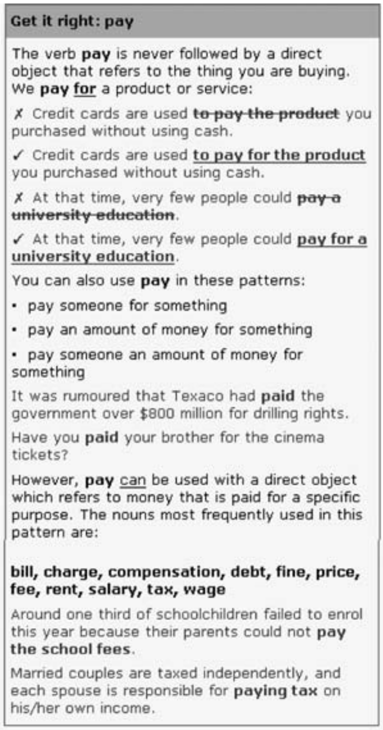
According to **one study**, women’s earnings are still 27% lower than men’s.

a detailed **programme of research** on the economics of nuclear energy

Figure 14.3 GIR box at ‘research’ in *MED2*

(the examples are preceded by a cross and the erroneous items have been crossed out), clear explanations of the source of the problem and clear and practical advice on how the error can be corrected and avoided.

It is worth noting that emphasis was also laid on prominent eye-catching presentation (with the use of colour, bold type and underlining) to ensure that the information in GIR boxes is not overlooked by learners.



Get it right: pay

The verb **pay** is never followed by a direct object that refers to the thing you are buying. We **pay for** a product or service:

✗ Credit cards are used ~~**to pay the product**~~ you purchased without using cash.

✓ Credit cards are used **to pay for the product** you purchased without using cash.

✗ At that time, very few people could ~~**pay a university education**~~.

✓ At that time, very few people could **pay for a university education**.

You can also use **pay** in these patterns:

- pay someone for something
- pay an amount of money for something
- pay someone an amount of money for something

It was rumoured that Texaco had **paid** the government over $800 million for drilling rights.

Have you **paid** your brother for the cinema tickets?

However, **pay** can be used with a direct object which refers to money that is paid for a specific purpose. The nouns most frequently used in this pattern are:

bill, charge, compensation, debt, fine, price, fee, rent, salary, tax, wage

Around one third of schoolchildren failed to enrol this year because their parents could not **pay the school fees**.

Married couples are taxed independently, and each spouse is responsible for **paying tax** on his/her own income.

FIGURE 14.4 GIR box at 'pay' in *MED2*

14.2.3 Grammar sections

Six so-called 'Grammar Sections' are included as part of the 'Improve your Writing Skills' component (De Cock et al. 2007) located in the 'Study Section' in the middle of the dictionary. The sections address key areas of difficulty relating to either word grammar (namely Articles, Complementation: Patterns used with nouns, verbs and adjectives, Countable and Uncountable Nouns, and Quantifiers) or specific aspects of written discourse (i.e. Punctuation and Spelling). The focus on word grammar and punctuation and spelling was deemed particularly appropriate in a dictionary (words, words, words!) that seeks to help learners with their writing.

Research based on the error-tagged component of *ICLE* (cf. above) served as the starting point for the selection and content of the sections. The larger unannotated *ICLE* corpus and the native speaker corpus (cf. above) were also used in the compilation of the sections.

The Grammar Sections are typically structured around the following key elements:

1. an introductory presentation and description of the phenomenon in the spotlight
2. learners' problems and errors relating to it (with authentic learner examples)
3. clear explanations of why the authentic leaner examples included are problematic
4. suggested corrections of the authentic leaner examples
5. general advice on how to avoid making such errors

Frequency data is also provided where relevant to raise learners' awareness of inappropriately overused items in academic writing, as illustrated by Figures 14.5 and 14.6 from the sections Punctuation and Quantifiers respectively.

The final paragraph of each section is entitled 'Advice on avoiding errors'. In addition to providing learners with general advice on how to avoid making the errors discussed in the sections, the paragraph also systematically highlights how the dictionary can help them deal with these problem areas. For example, the section Complementation is rounded off with clear explanations and illustrations of where and how information relating to complementation can be found at dictionary entries (cf. Figure 14.7). The sections also include systematic cross-references to related 'Get it right' boxes and to related Grammar Sections and/or Writing Sections.

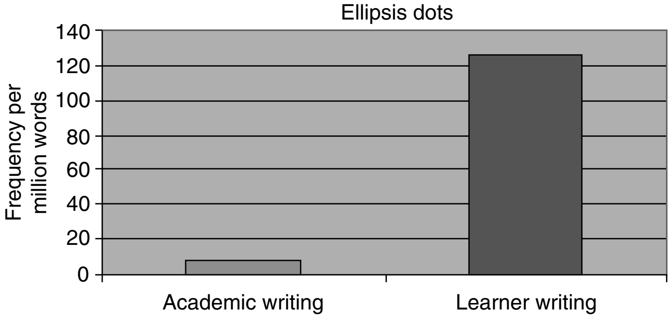
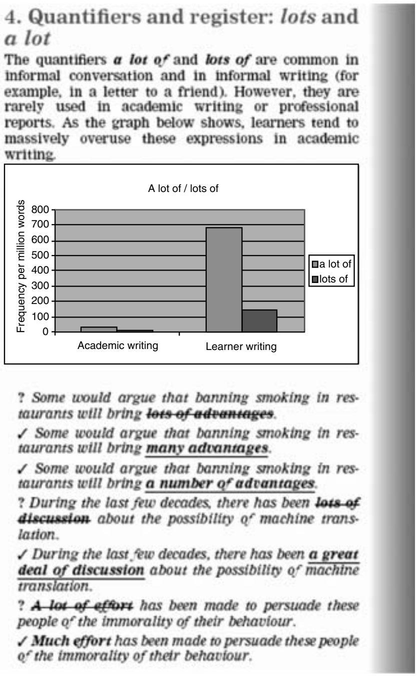


FIGURE 14.5 Frequency of ellipsis dots in learner and native speaker writing (section about Punctuation, *MED2*)



4. Quantifiers and register: *lots* and *a lot*

The quantifiers ***a lot of*** and ***lots of*** are common in informal conversation and in informal writing (for example, in a letter to a friend). However, they are rarely used in academic writing or professional reports. As the graph below shows, learners tend to massively overuse these expressions in academic writing.

A lot of / lots of
Frequency per million words
800
700
600
500
400
300
200
100
0
Academic writing
Learner writing
■a lot of
■lots of

? *Some would argue that banning smoking in restaurants will bring* ~~***lots of advantages***~~.

✓ *Some would argue that banning smoking in restaurants will bring* ***many advantages***.

✓ *Some would argue that banning smoking in restaurants will bring* ***a number of advantages***.

? *During the last few decades, there has been* ~~***lots of discussion***~~ *about the possibility of machine translation.*

✓ *During the last few decades, there has been* ***a great deal of discussion*** *about the possibility of machine translation.*

? ~~***A lot of effort***~~ *has been made to persuade these people of the immorality of their behaviour.*

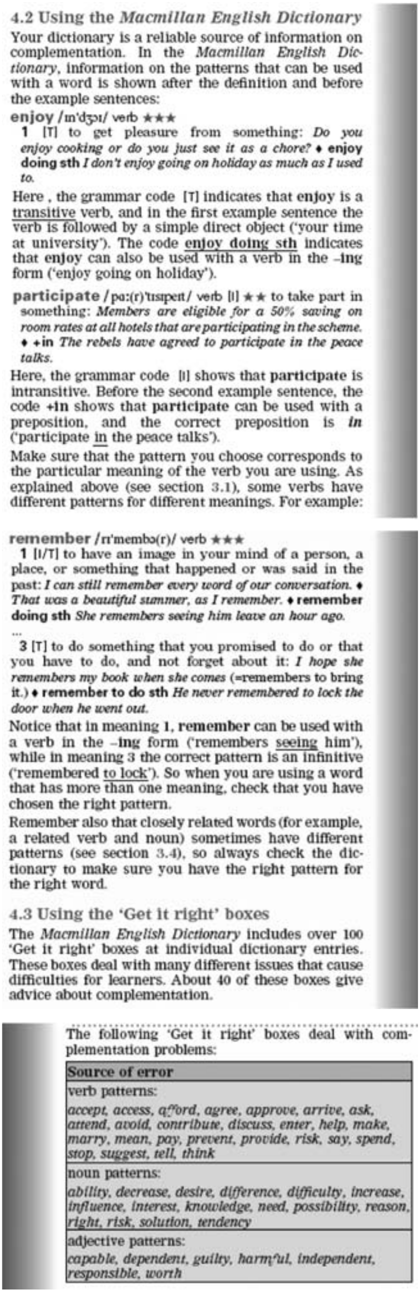
✓ ***Much effort*** *has been made to persuade these people of the immorality of their behaviour.*

FIGURE 14.6 Frequency of the quantifiers *a lot of* and *lots of* in learner and native speaker writing (section about Quantifiers, *MED2*)

These 'Advice on avoiding errors' paragraphs can thus also be seen to provide learners with some form of dictionary training, which has been shown to be extremely valuable if learners are to make the most of the riches contained in MLDs.

14.2.4 EAP writing sections

Beside the Grammar Sections described above, the 'Improve your Writing Skills' component in the middle of the dictionary also includes twelve 'Writing Sections'. These Writing Sections focus on rhetorical or organization functions that are particularly prominent in academic writing: (1) Adding information, (2) Comparing and Contrasting: Describing similarities and differences, (3) Exemplification: Introducing examples, (4) Expressing Cause and Effect, (5) Expressing Personal Opinions, (6) Expressing Possibility and Certainty, (7) Introducing a Concession, (8) Introducing



4.2 Using the *Macmillan English Dictionary*

Your dictionary is a reliable source of information on complementation. In the *Macmillan English Dictionary*, information on the patterns that can be used with a word is shown after the definition and before the example sentences:

enjoy /ɪn'dʒɔɪ/ verb ★★★
1 [T] to get pleasure from something: *Do you enjoy cooking or do you just see it as a chore?* ♦ **enjoy doing sth** *I don't enjoy going on holiday as much as I used to.*

Here , the grammar code [T] indicates that **enjoy** is a transitive verb, and in the first example sentence the verb is followed by a simple direct object ('your time at university'). The code **enjoy doing sth** indicates that **enjoy** can also be used with a verb in the **-ing** form ('enjoy going on holiday').

participate /pɑː(r)'tɪsɪpeɪt/ verb [I] ★★ to take part in something: *Members are eligible for a 50% saving on room rates at all hotels that are participating in the scheme.* ♦ **+in** *The rebels have agreed to participate in the peace talks.*

Here, the grammar code [I] shows that **participate** is intransitive. Before the second example sentence, the code **+in** shows that **participate** can be used with a preposition, and the correct preposition is ***in*** ('participate in the peace talks').

Make sure that the pattern you choose corresponds to the particular meaning of the verb you are using. As explained above (see section 3.1), some verbs have different patterns for different meanings. For example:

remember /rɪ'membə(r)/ verb ★★★
1 [I/T] to have an image in your mind of a person, a place, or something that happened or was said in the past: *I can still remember every word of our conversation.* ♦ *That was a beautiful summer, as I remember.* ♦ **remember doing sth** *She remembers seeing him leave an hour ago.*
...
3 [T] to do something that you promised to do or that you have to do, and not forget about it: *I hope she remembers my book when she comes* (=remembers to bring it.) ♦ **remember to do sth** *He never remembered to lock the door when he went out.*

Notice that in meaning **1**, **remember** can be used with a verb in the **-ing** form ('remembers seeing him'), while in meaning 3 the correct pattern is an infinitive ('remembered to lock'). So when you are using a word that has more than one meaning, check that you have chosen the right pattern.

Remember also that closely related words (for example, a related verb and noun) sometimes have different patterns (see section 3.4), so always check the dictionary to make sure you have the right pattern for the right word.

4.3 Using the 'Get it right' boxes

The *Macmillan English Dictionary* includes over 100 'Get it right' boxes at individual dictionary entries. These boxes deal with many different issues that cause difficulties for learners. About 40 of these boxes give advice about complementation.

The following 'Get it right' boxes deal with complementation problems:

Source of error
verb patterns: *accept, access, afford, agree, approve, arrive, ask, attend, avoid, contribute, discuss, enter, help, make, marry, mean, pay, prevent, provide, risk, say, spend, stop, suggest, tell, think*
noun patterns: *ability, decrease, desire, difference, difficulty, increase, influence, interest, knowledge, need, possibility, reason, right, risk, solution, tendency*
adjective patterns: *capable, dependent, guilty, harmful, independent, responsible, worth*

FIGURE 14.7 Advice on avoiding errors using the *Macmillan English Dictionary* (section about Complementation, *MED2*)

Topics and Related Ideas, (9) Listing Items, (10) Reformulation: Paraphrasing or clarifying, (11) Quoting and reporting and (12) Summarizing and Drawing conclusions. The Writing Sections were compiled on the basis of findings from a detailed analysis of 350 EAP markers (see Paquot 2007) in the *ICLE* corpus and in the native speaker corpus used in the project. As well as explaining the main strategies that writers can use to perform the 12 functions in formal writing, the sections also specifically address the wide range of problems advanced learners can be shown to experience when performing these functions. Special 'Be careful' notes, 'Get it right' boxes and Collocation boxes are included in the Writing Sections to help learners deal with problems of frequency, register, positioning, semantics and phraseology. For a detailed description of the Writing Sections and the method used to compile them, see Gilquin et al. (2007), Gilquin and Paquot (2008) and Granger and Paquot (2008).

14.3 Conclusion

The outcomes of the collaborative project reported on in this chapter are a good example of the invaluable contribution learner corpus-based research can make to MLDs. In this project trained lexicographers have worked in very close collaboration with learner corpus-based researchers (who are also experienced ELT teachers) to produce materials that can be seen to help boost the confidence of the MLD as a truly productive tool.

Notes

[1] 'Extracts, reproduced by kind permission of Macmillan Publishers Limited, taken from *Macmillan English Dictionary*, Second Edition, published 2007. © A&C Black Publishers Ltd 2007.'

References

De Cock, S. and Granger, S. (2004), 'Computer learner corpora and monolingual learners' dictionaries: The Perfect Match', in Teubert, W. and Mahlberg, M. (eds), *The Corpus Approach to Lexicography*. Special issue of *Lexicographica*, 20, 72–86.

De Cock, S., Gilquin, G., Granger, S., Lefer, M-A., Paquot, M. and Ricketts, S. (2007), 'Improve your writing skills', in Rundell, M. (editor in chief) Macmillan English

Dictionary for Advanced Learners (second edition). Oxford: Macmillan Education, IW1–IW50.

Gillard, P. (ed.) (2003), *Cambridge Advanced Learner's Dictionary*. Cambridge: Cambridge University Press.

Gilquin, G. and Paquot, M. (2008), 'Too chatty: Learner academic writing and register variation'. *English Text Construction*, 1, (1), 41–61.

Gilquin, G., Granger, S. and Paquot, M. (2007), 'Learner corpora: The missing link in EAP pedagogy', in Thompson, P. (ed.) *Corpus-based EAP Pedagogy* . Special issue of *Journal of English for Academic Purposes*, 6, (4), 319–335.

Granger, S. and Paquot, M.(2008), 'From dictionary to phrasebook?', in Bernal, E. and DeCesaris, J. (eds), *Proceedings of the XIII EURALEX International Congress*, Barcelona, Spain, 15–19 July 2008, 1345–1355.

Paquot, M. (2007), 'Towards a productively-oriented academic word list', in Walinski, J., Kredens, K. and Gozdz-Roszkowski, S. (eds), *Corpora and ICT in Language Studies. PALC 2005*. Lodz studies in LANGUAGE 13. Frankfurt am Main: Peter Lang, pp. 127–140.

Rundell, M. (1999), 'Dictionary use in production', *International Journal of Lexicography*, 12, (1), 35–53.

Rundell, M. (ed.) (2007), *Macmillan English Dictionary for Advanced Learners* (second edition). Oxford: Macmillan Education.

Summers, D. (ed.) (1995), *Longman Dictionary of Contemporary English*. Harlow: Longman.

—(ed.) (1993), *Longman Language Activator*. Harlow: Longman.

—(ed.) (1997), *Longman Essential Activator*. Harlow: Longman.

Chapter 15

Advanced Learner Corpus Data and Grammar Teaching: Adverb Placement

Tom Rankin
Institute for English Business Communication, Vienna University of Economics and Business Administration

15.1 Introduction

Learner corpus researchers often note the practical applications of their work for foreign language teaching (e.g. Granger 1998, Granger et al. 2002, O'Keeffe et al. 2007). Granger and Tribble (1998: 201) describe the advantage of using corpus data as follows:

> Until recently the selection of words, phrases and structures for form-focused instruction was largely based on teachers' intuitions. While this approach has its merits, it suffers from one major weakness: teachers' intuitions fail to provide a complete picture of learners' problems.

The answer to this weakness, they suggest, is the use of authentic learner data in the foreign language classroom. This idea has started to filter through to publishers and there are now a number of resources for language learners based on authentic data from learner corpora (e.g. *Macmillan English Dictionary for Advanced Learners*). Learner corpus data has been restricted mostly to use in the field of pedagogical lexicography. The learner dictionaries produced as a result tend to have a focus on phraseology, especially for EAP, highlighting appropriate collocations, and drawing the learner's attention to issues of over- and underuse. Dictionaries such as Macmillan and others provide invaluable tools and aids to improve advanced learners' writing in terms of complexity and fluency, but grammar can still pose a problem at more advanced levels, and ELT grammars based on learner corpus data are rare.

This chapter seeks to add to the field of learner corpus-based ELT by focusing solely on the possible application of learner corpus data to the teaching of grammar and syntax. The grammar point to be investigated is adverb placement. The aim is to show how the errors produced by advanced learners can provide a basis for more focused grammar exercises and teaching materials in a university English course.

15.2 Adverb Placement in Second Language Acquisition

Misplaced adverbs have been much studied in generative second language research as a diagnostic for verb raising (White 1990/91, Eubank 1993/94, Eubank et al. 1997, and others). The hypothesis, at least in the Government and Binding incarnation of generative theory, is that the verb raising parameter may be carried over from a learner's L1 into a non-raising target language such as English, producing errors where the verb is raised over the adverb, which then surfaces in an ungrammatical position between verb and object.

Some of the work in this tradition (White 1990/91, Haegeman 1992) has looked at possible pedagogical implications/applications of the research. However, in a study of how classroom input might reduce the tendency to produce this type of error, White concludes that resetting this parameter successfully is resilient to long-term correction by formal instruction.

Problems with this approach, especially in terms of possible applications to language teaching, are that it does not usually take authentic learner data into account, preferring elicited production data or grammaticality judgements. Secondly, it concentrates on one type of learner error, the VAO (verb adverb object) sequence. While this is obviously a problem for particular groups of learners of English (or indeed a developmental problem learners from various L1 backgrounds, see Osborne 2008), it does not take into account any more general problems with adverb placement which might be related and could therefore be dealt with as a whole in grammar instruction.

By using learner corpus data here, it will be shown that there are regularities in the occurrence of misplaced adverbs in advanced German speaking learners' writing. By comparing these to the materials used to teach adverb grammar in the university English course which the students followed, it is shown that the corpus data provides valuable information which could be exploited in the course's grammar teaching materials and exercises. The notion implicit in this, and in learner corpus-based teaching methods as a whole, is that studying authentic learner language and incorporating the findings into teaching will produce more effective teaching materials.

15.3 The Study

15.3.1 Corpus and method

As part of a pilot study for a doctoral dissertation, a corpus of student work was collected at the Vienna University of Economics and Business Administration (WU). The part of the corpus used in this study is made up of 37 seminar essays by 37 different students and comprises a total of 103,073 words. The students are all native speakers of (Austrian) German and they each completed a learner profile giving demographic information and information about length of time spent studying English and living in an English-speaking country. The average age of the students in the study is 24.2 years and they had an average of 13.6 years of formal English teaching.

The materials used to teach adverb grammar to the group of students in the study were also collated and studied. This should not, however, be construed as an attempt to illustrate any straightforward link between grammar teaching and eventual attainment in a second language. Rather, the idea is to show how it might be possible to use advanced students' writing to provide for the design of more subtle and appropriate grammar teaching materials.

Using the UAM Corpus Tool (O'Donnell 2006), the corpus was annotated for word order errors with adverbs. The error taxonomy was kept maximally simple and based on the fact that in English certain semantic types of adverbs prefer certain syntactic positions and that adverbs can modify different phrasal categories. The error categories are illustrated in Table 15.1 (1).

While the absolute number of errors is low at 86 (8.03 per 10,000 words), it should be borne in mind that the students are at a relatively advanced level. They wrote the essays in their own time with access to reference tools and the work was submitted for assessment. It can therefore be assumed

Table 15.1 Error taxonomy

Error Code	Description	Number of Occurrences
Post V	Adverb of wrong semantic category occurs after lexical verb (and its complements).	30
VAO	Adverb is placed between verb and its object.	
Pre V	Adverb of wrong semantic category precedes a lexical verb.	20
Pre Aux	Adverb is placed before modal or aspectual auxiliary.	10
Modification	Adverb modifies wrong phrasal category in clause.	26

that those errors which remain are true reflections of deficiencies in the students' knowledge rather than mistakes due to time pressure or online processing constraints.

In order to confirm that the errors coded do in fact deviate from native usage, certain aspects of adverb usage in the WU Corpus were compared to the Louvain Corpus of Native English Essays (LOCNESS). Both corpora were tagged using the CLAWS tagger in WMatrix (Rayson) to facilitate recovery of instances of adverbs.

15.3.2 English Course

A brief description of the English course the students followed is included to provide a point of comparison and to highlight where corpus data might suggest improvements.

The English course at WU is made up of three compulsory courses WIKO 1–3 (English business communication). The first of these courses deals explicitly with grammar points, and has a unit devoted to adjectives and adverbs. The second and third courses have a more ESP flavour and deal more specifically with business English terminology and the sort of language tasks necessary for business communication such as letter writing, report writing, etc.

There are standard course books produced in-house with teaching examples and exercises which the students complete. More detailed grammar explanations are left to individual teachers. Teachers can and do provide additional materials for their class but this obviously cannot be reconstructed here and so only the 'official' exercises are taken into account.

Altogether there were 14 exercises throughout the course which dealt with adverb grammar as illustrated in Table 15.2.

As is obvious, the focus of these exercises is the morphological distinction between adjectives and adverbs. It is understandable that there should be a

Table 15.2 Adverb Grammar Exercises in WU English course

Grammar Point	**Type of Exercise**	**Number of Exercises in Course**
Adjective/Adverb Distinction	Gapfill, Transformation, Editing	6
Comparative/Superlative Formation	Gapfill	2
Adverb Function	Gapfill, Transformation	2
Position	Placement in sentence	2
Other	Gapfill, Editing	3

concentration on this aspect of adverb grammar in a course designed mainly to cater for German native speakers as German does not regularly distinguish morphologically between adjectives and adverbs. Indeed, this distinction continues to cause some problems for advanced learners as in example (8) below. The analysis of the errors which follows includes suggested changes to grammar teaching materials.

15.4 Results

15.4.1 Post-verbal placement/VAO

Altogether, adverbs wrongly placed after the lexical verb were the most frequent type of error at 2.9 occurrences per 10,000 words. This also includes instances of VAO, which alone accounted for 2.2 occurrences per 10,000 words. These errors were relatively widely distributed in the corpus, occurring in 15 papers.

VAO is of course possible in native English in instances of heavy object shift. This is however rare in LOCNESS with a frequency of 0.06 per 10,000 words in the British component of the corpus. Given the lack of coverage of heavy NP shift in grammar teaching and the fact that the same effect does not occur in the learner population's L1, the use of this structure is simply taken here to be an error rather than an approximation of a native structure.

From a generative theoretical perspective, the occurrence of this sort of error is unsurprising in a learner population whose L1 allows verb raising. However, an examination of the instances of VAO reveals some striking similarities in the sort of errors made, which may be exploitable in grammar teaching.

Fifty-two per cent of VAO sequences occur with a verb in the infinitive as in (1).

(1) 'In this period it is crucial to provide rapidly information to the public.'

Why this particular structure may induce the misplacement of adverbs perhaps warrants study in its own right. Perhaps the students have picked up during their studies that the split infinitive is wrong and this is a method used to avoid splitting the infinitive with an adverbial. Whatever the reason for misplacement in this particular syntactic environment, it provides valuable information for the grammar teacher and shows how learner corpora can provide added value in materials design. All teachers do of course have an instinct for what sorts of errors their students make. It is, however, unlikely that this sort of specific insight about a problematic syntactic environment would occur to a teacher without having examined a corpus of authentic learner writing. As it seems that this structure causes problems for advanced learners, perhaps this sort of sentence should be given special attention in teaching examples or in gap fill and adverb placement exercises. This sort of corpus data would also be valuable as a source of editing or error correction exercises, providing as it does the sort of errors which the learners themselves might be prone to making.

The additive adverbs 'also' and 'as well' seem more prone to misplacement after a verb than adverbs of other semantic categories. Overall these two adverbs account for 37 per cent of these errors, and 30 per cent of VAO errors. In fact, these adverbs accounted for 21 per cent of total errors in the corpus and focusing adverbs for a further 23 per cent. Again then, corpus data provides a helpful way to approach grammar teaching and suggests perhaps that these types of adverbs should be given more attention.

On the basis of corpus evidence, it seems that there are some syntactic and semantic regularities in the production of VAO errors in this learner population. This could be taken into consideration in the design of teaching materials to target those areas which cause specific difficulties for learners, rather than considering this error as the manifestation of an abstract grammatical constraint which is not amenable to correction by teaching.

15.4.2 Modification

Errors connected to what phrasal categories an adverb may modify provide perhaps the most interesting category and could probably be dealt with most effectively in formal instruction. Modification errors occurred in 17 papers with a frequency of 2.4 per 10,000 words. In this category, focusing adverbs are particularly likely to be misplaced – 73 per cent of modification errors involve focusing adverbs, an example is provided in (2).

(2) 'On the one hand MNCs cannot only be blamed for these negative developments.'

From the context, the intended meaning is something like 'MNCs *alone* cannot be blamed . . .' However the focusing adverb has been placed in a position where it focuses the proposition ‘be blamed.’ This problem of focusing one element in a proposition recurs as shown below.

(3) ‘For a lot of people, it might be even difficult to describe their own culture.’

(4) ‘The United States have passed a number of different Acts and Laws to protect the national security . . . only to name a few.’

In (3) ‘even’ modifies ‘difficult’ too narrowly giving rise to an interpretation where one would expect ‘difficult’ to be in contrast with something in the previous discourse, whereas the intended and more natural usage would have been ‘. . . it might even be difficult.’ By contrast, in (4) ‘only’ modifies the whole proposition in the final infinitive clause rather than focusing ‘few’ as in ‘to name only a few.’

It seems then that the main difficulty here is with discourse and pragmatic influences on adverb placement. Indeed, this could just as easily apply to the cases mentioned above involving post-verbal placement of additive adverbs such as ‘also’. Example (5) is typical of the sort of error that occurs with ‘also’.

(5) ‘Therefore, separate computer installations caused also a lot of separate software solutions which became more and more difficult to maintain.’

This placement of ‘also’ is possible in native English but with a special discourse function. The interpretation of this example would be that a number of solutions caused by separate computer installations had already been evoked in the previous discourse and this sentence is then adding to that specific information, focusing this new special subset in some way. The more natural order here would have been to place ‘also’ before the verb as it relates better to the previous discourse where there had been a discussion of what else separate computer installations *did* but not what they had already *caused.*

It is possible that some errors in this category are due to overgeneralization of the adverb placement ‘rule’ in English, i.e. before the lexical verb. Some of these examples are at best borderline errors, in example (6) for instance we have a focusing adverb in this position modifying the verb, and indeed no doubt many native speakers would find this acceptable but it seems strange in the context of the student essay, where '. . . especially when . . .' would have been the more natural choice.

(6) 'Such failures especially occur when there is a lack of experience with the method.'

Analysing these examples rigorously would necessitate an extensive comparison with native data which is not possible here. In any case, it becomes clear when looking at this data that modification is a difficulty for learners and the learner data again provides pointers for materials design. The focusing adverbs which pose problems can be used to modify various phrasal categories and perform specific discourse and pragmatic functions such as adding and focusing information. It would seem then to be helpful to include some instruction which makes explicit the semantic and pragmatic differences between different adverbial positions for these types of adverbs as in (7).

(7) 'Even I like grammar classes.'
'I even like grammar classes.'
'I like grammar classes even.'

This could be especially relevant for advanced learners who are beginning to write longer pieces of work in English and so could be added to the familiar exercises and drills designed to practise using connectors and other discourse organizing phrases to create coherent discourse.

15.4.3 Pre-verbal placement

Misplacement of adverbs before a lexical verb does not seem to be a major problem for the students in this study. The occurrence at 1.94 per 10,000 words is slightly lower than the first two error categories but one quarter of the pre-verbal errors occurred in one paper. As mentioned briefly above, the tendency where the error did occur was to overgeneralize the pre-verbal position to those semantic categories of adverb which prefer post-verbal position. The following example is informative:

(8) 'The word metaphor is well known but often different seen by different people.'

Here 'different' (lacking adverb morphology) follows the pattern of the other adverbs in the sentence in pre-verbal position. This sort of authentic example, as with all the others taken from the learner corpus, could be useful in editing and error correction exercises.

15.4.4 Pre-auxiliary placement

Reference grammars note a difference for this adverbial position between British and American varieties of English, stating that it is a possible unmarked position in American English (Swan 2002). In order to provide a valid comparison to the learner corpus, instances of pre-auxiliary placement of adverbs were extracted from both the British and American components of LOCNESS. Perhaps surprisingly, this position occurred marginally more frequently in the British part of the corpus (1.09 per 10,000 words) than the American (0.83 per 10,000). These low frequencies would seem to suggest that this is a marked position for adverbs in both British and American English, at least in as far as the LOCNESS corpora can be taken to be representative of these varieties.

The WU corpus contains more instances of adverbs in this position but the frequency is still low at 1.06 per 10,000 words. It is perhaps possible that the slight overuse of this position is due to teaching effects if it is taught as a possible adverb position in American English. The learners do however seem to treat pre-auxiliary as a marked position and use it sparingly, if still quantitatively more compared to LOCNESS. Qualitatively there is a preference for modal and connective adverbs in pre-auxiliary position in the native corpora. Adverbs of these two types together account for 50 per cent of instances in LOCNESS but do not occur at all in this position in the non-native corpus.

This adverb position probably requires much more research to provide any concrete results. On the basis of the corpora used here, it seems that it might be wise not to make a simple distinction between British and American varieties but to provide students with information about what sort of adverbs are found in both varieties in this position. Here an explicit comparison between authentic native and non-native usage might be illuminating for students.

15.5 Conclusion

It has been shown that adverb placement continues to pose problems for the learners in this study and that adverb syntax is not dealt with to a significant extent in the grammar teaching materials used in the courses followed by these students. All the adverb grammar exercises do of course deal in some way with adverbial syntax as students have to choose which type of adverb to put in a particular position in all the gap-fill exercises for example. It is argued, however, that the course could be improved with some more attention paid to adverb placement. In particular, it seems that the residual problems with adverb placement are not due to any major deficiencies in basic grammar but rather to the fact that appropriate variation in adverb placement for specific discourse and pragmatic contexts has not been mastered, and indeed has not been given significant attention in teaching. This could and perhaps should be introduced in formal grammar/writing instruction at more advanced levels. Corpus data of the sort presented here could give pointers for the selection and sequencing of these sorts of topics at advanced levels of acquisition and provide practical help in choosing which type of semantic and syntactic features prove most problematic for the learners and should therefore be included in teaching examples and exercises. This would be particularly useful at more advanced levels when students have mastered basic concepts and require a more targeted and subtle approach to help eliminate persistent errors.

References

Eubank, L. (1993/1994), On the transfer of parametric values in L2 development. *Language Acquisition*, 3, 183–208.

Eubank, L., Bischof, J., Huffstutler, A., Leek, P. and West, C. (1997), '"Tom eats slowly cooked eggs": Thematic verb-raising in L2 knowledge', *Language Acquisition*, 6, (3), 171–99.

Granger, S. (ed) (1998), *Learner English on Computer*. London & New York: Addison Wesley Longman.

Granger, S. and Tribble, C. (1998), 'Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning,' in Granger S. (ed.), *Learner English on Computer*. London and New York: Addison Wesley Longman, pp. 199–209.

Granger S., Hung J., Petch-Tyson S. (eds), (2002), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam and Philadelphia: Benjamins.

Haegeman, L. (1992), 'Adverbial Positions and Second Language Acquisition', *Technical Reports in Formal and Computational Linguistics*, No. 3. University of Geneva, Geneva, Switzerland.

O'Donnell, M. (2006), 'The UAM Corpus Tool'. Paper presented at the 18th European Systemic Functional Workshop. University of Trieste, Gorizia, Italy. Retrieved 20 May 2007, from: http://www.wagsoft.com/CorpusTool/.

O'Keeffe, A., McCarthy, M., and Carter, R. (2007), *From Corpus to Classroom.* Cambridge: Cambridge UP.

Osborne, J. (2008), 'Adverb placement in post-intermediate learner English: A contrastive study of learner corpora', in Gilquin, G. Papp, S. and Diez-Bedmar, M. B. (eds), *Linking up Contrastive and Learner Corpus Research.* Amsterdam and New York: Rodopi, pp. 127–146.

Rayson, P. (2008), Wmatrix: A web-based corpus processing environment, Computing Department, Lancaster University. Retrieved 5 May 2007, from: http://ucrel.lancs.ac.uk/wmatrix/

Rundell, M. (editor-in-chief) (2007), *Macmillan English Dictionary for Advanced Learners* (second edition). Oxford: Macmillan Education.

Swan, M. (2002), *Practical English Usage* (second edition). Oxford: Oxford University Press.

White, L. (1990/91), 'The verb-movement parameter in second language acquisition', *Language Acquisition,* 1, (4), 337–60.

Chapter 16

FL Students' Input in Higher Education Courses: Corpus Methodology for Implementing Language Representativeness

Izaskun Elorza and Blanca García-Riaza
University of Salamanca

16.1 Introduction

Since Krashen put forward his Input Hypothesis in 1985, many language acquisition theories stress the importance of input in L2 acquisition, especially its critical role in our understanding of how learners create linguistic systems (Van Patten 2003: 25). The kind of input that teachers of higher education courses consider for learners of English Studies at tertiary level includes texts dealing with literary or cultural aspects of the L2 which also show great variation in terms of topic, style or length. One of the strategies to minimize this variability is to adjust the length of the texts which will be used in the higher education course according to the type of learning activity in which the text will be used, such as the practice of reading comprehension skills or even of text-copying by means of dictations.

Our main aim in this chapter is to analyse how corpus methodology can help us define the model of language we use in the classroom. The rationale for this is that the texts selected as class input may be considered in terms of their authenticity; as Johns points out, '[i]t is by now generally accepted by practitioners of EAP that texts used to teach reading should be tampered with as little as possible' (Johns 1994: 103). There is also the consideration of the kind of language those texts are meant to show to the learners. In this respect, the texts might be taken by the learners as a model of the language used for communication in the 'real world'.

However, one of the problems we face here is that the texts chosen for class input are often adjusted in order to facilitate learners' processing, or shortened to a more desirable length. When this happens, we may argue that the model of language presented to the learners may be different from language as it is used in real communication. In this chapter we suggest that corpus analysis software such as WordSmith Tools (Scott 2004) may help us identify when adaptations or modifications result in different language patterns from the linguistic representation expected and, therefore, it may help us implement the selection of texts for class input. Our underlying assumption is that better input availability may enhance students' self-control of their learning, e.g. the detection of new words to learn or the particular use of an expression in different contexts.

16.2 Written Input Availability: The Use of Pedagogic Corpora

All theories of language acquisition consider input a basic component of the acquisition process. The underlying assumption for this is that, as language is never produced in a vacuum, when acquiring a foreign language, learners need data which can be used as linguistic evidence in order to formulate hypotheses about the foreign language. In a broad sense, 'input' is the term given to the language which is available to the learner through any medium (Gass and Mackey 2006: 4), so '[c]onceptually, one can think of the input as that language (both spoken and written forms) to which the learner is exposed' (Gass and Selinker 2001: 260).

Nevertheless, the language available is considered to be input for acquisition only if it has some kind of communicative intent, i.e. if it consists of 'meaning-bearing utterances' which can be used by learners for the creation of an implicit linguistic system (Lee and VanPatten 2003: 26–27). The construction of such a system is the result of the learner's formulation of hypotheses about the foreign language, which depends on the learner's experience in the foreign language. This experience is unique to each learner and includes the internalization of patterns of use of the language. Michael Hoey's recent theory of lexical priming (Hoey 2005) is based on the assumption that exposure to words in use produces a cumulative effect in the internalization of their use in such a way that the particular use we make of a word is conditioned by our previous experience in our encounters with it in texts and communication. In this respect, the features of the input used in the classroom are extremely important.

In the instructional environment, learners are confronted with a linguistic input which is 'controlled and structured by the teacher [. . .] and by the materials used' (Gass 1990: 42); it is the teacher's task, then, to construct or compile the pedagogic corpus which will serve the course aims. The concept of corpus has been traditionally used in linguistics to refer to 'a collection of naturally occurring examples of language, consisting of anything from a few sentences to a set of written texts or tape recordings, which have been collected for linguistic study' (Hunston 2002: 2), but more recently corpora tend to be described as 'a collection of texts (or parts of text) that are stored and accessed electronically' (Ibid.). The emphasis on the electronic storage of texts is justified by the fact that, unlike other collections of texts such as libraries or electronic archives, a corpus is 'stored in such a way that it can be studied non-linearly, and both quantitatively and qualitatively' (Ibid.). The reason for this is that the purpose of the compilation is not just to be able to access the set of texts in order to read them but also to process the information contained in them with different aims, e.g. to find out how frequent an expression is in a particular genre.

A pedagogic corpus is defined by Hunston as 'a corpus consisting of all the language a learner has been exposed to' (Hunston 2002: 16). Input can be made available as a pedagogic corpus by collecting and compiling the transcription in an electronic format of 'all the course books, readers etc. a learner has used, plus any tapes etc they have heard' (Ibid.). Hunston also explains that this kind of corpus is useful for FL teaching and learning because it can be used 'to collect together for the learner all instances of a word or phrase they have come across in different contexts, for the purpose of raising awareness' (Ibid.).

Teachers often rely on two criteria for selecting texts for class input: brevity – it is easier to handle a shorter text than a longer one in class – and topic – according to its relevance to the course interests (Dubin et al. 1983: 138). But we also tend to base our decisions on other parameters, often of a qualitative nature, such as whether the text is a good example of a certain type or a good sample of real language; in short, whether it is representative enough to illustrate the point we want to teach. Here the representativeness of the text is estimated from its identification as belonging to a certain text type as well as from its particular textual features. Therefore, we interpret that the text is representative by estimating how typical it is as a sample of a type and also how successfully it fulfils the standards of textuality (De Beaugrande 1980; De Beaugrande and Dressler 1981), especially in relation to text-centred standards, i.e. cohesion and coherence. The question of the authenticity of texts is also related to this issue. The texts chosen for the written input of the course can also be categorized according to whether they are authentic or non-authentic (Harmer 1983: 146). Authentic texts are designed for native speakers, such as an editorial published in an English newspaper. Those texts are called natural texts by Dubin, Eskey and Grabe, who define them as 'examples of the target language as used by native speakers for authentic communicative goals' (Dubin et al. 1986: 137). Non-authentic texts, however, are written either 'to illustrate particular language points' or 'to appear authentic even though there has been some language control of the rough-tuning type' (Harmer 1983: 146).

One of the problems of non-authentic texts in relation to their representativeness is that, when modifying an authentic text or when writing an *ad hoc* one, the resulting text can present features which are artificial when compared to natural language, particularly those related to cohesion and coherence. The text titled 'The Next Morning' presented by Harmer (1983: 147) is an example of this. Even if coherence is achieved in the text, it presents features which cannot be associated with a good example of English language as it is used by native speakers in real communication, such as the sudden changes of topic, as the following excerpt illustrates: 'The two girls are drinking coffee. It is very good and very black. Kate is a bad cook but she can make good coffee' (Ibid.). In this sense, we cannot take 'The Next Morning' text as a good representation of English language; it is not a text which we would include in the pedagogic corpus. On the other hand, we should also emphasize that a pedagogic corpus for higher education students cannot be easily compiled in terms of textual representativeness related to pertaining to a text type. Students are often provided with a very wide range of written material which may include a great number of text types within a course, so it is more realistic to aim at compiling a pedagogic corpus based on a notion of representativeness related to textual quality or communicative success.

In order to analyse the features of texts as candidates for a pedagogic corpus by means of corpus linguistics methodology, we have searched in our own personal experience with higher education courses. One of the written production activities which is often practised (at least in Spain) is taking down messages from the dictation of short texts, which involves the reading of a text as spoken input and the students' re-production of the text as written output. Narrative texts are usually chosen for this kind of activity, dealing very often with descriptions of some cultural, literary or historical issue. Brevity is a crucial criterion here, as this activity is time-consuming and tiring, so the texts selected are frequently modified versions of authentic material, *ad hoc* texts for FL teaching, or also shortened versions of about 150–180 words. This modification usually involves selecting only a few paragraphs from a longer text, typically the opening ones. Sometimes the modification is based on the criterion of topic, by changing some words which are considered too difficult or irrelevant. The resulting texts are then quite homogeneous in length, although they vary in their degree of difficulty according to the familiarity of lexical words and their density in the text. Until recently, this selection was based solely on teachers' intuitions about what should be maintained and what should be taken out but the question was raised as to whether those texts really presented features of authenticity. In the following section we present some of our results regarding this matter.

16.3 What Corpus Methodology Can Do for Input Visibility, Accessibility and Language Representativeness

Corpus linguistics is a framework which provides more advantages for enhancing input visibility and accessibility than introspective approaches (introspection and elicitation) because corpus-based observations can be verified better and more accurately than judgements based on introspection (McEnery and Wilson 2001: 14) so, in this sense, this methodology is potentially more efficient when compiling a pedagogic corpus than mere intuition.

The use of vocabulary lists derived from word counts of corpora is not new in foreign language pedagogy (McEnery and Wilson 2001: 4) but our claim here is that the use of more sophisticated tools for text processing logically leads to more accuracy and visibility of the written input. For example, teachers may be interested in producing a list of the vocabulary present in the texts used as input in class in order to control the amount and type of vocabulary practised and learnt. It is much easier and less time-consuming to produce wordlists with the software developed for corpus processing than to try to keep control of the relevant vocabulary from the input without them. These advantages do not only benefit teachers but students as well. If the written input is made visible in different types and degrees of processing (as a database of texts, as a collection of wordlists or any other kind of product), access to the information may not only be easier but may also result in a deeper understanding of how language is organized because, as Stubbs points out, '[i]ntuitive judgements are particularly untrustworthy with respect to the frequency and distribution of different forms and meanings of words, and to the interaction of lexis, grammar and meaning' (Stubbs 1996: 31).

Nevertheless, if we envisage the compilation of the written input in terms of corpus construction, we must take into consideration which texts are to be included. As we have discussed in the previous section, this choice is usually based on their size, topic, authenticity and representativeness.

Paltridge (2006: 162) points out that the sampling and representativeness of the corpus must be considered in relation to the definition of the target population that the corpus represents. Regarding this aspect, Biber suggests that, while any selection of texts is a sample, the representativeness of the corpus further depends on 'the extent to which it includes the range of linguistic distribution in the population. That is, different linguistic features are differently distributed (within texts, across texts, across text types)' (Biber 1994: 378, as quoted by Paltridge 2006: 163). However, as we have already mentioned, when dealing with the corpus compilation of the written input, we cannot ignore the great variety of the texts used in higher education courses. The need for using texts from different types seems to impede the very possibility of compiling a representative corpus in terms of typological representativeness (representative of which type in particular?). Therefore, even if we are aware that this kind of representativeness is crucial for the value of the results we may obtain, we are not dealing with it here due to the practical difficulties already mentioned in relation to the compilation of the pedagogic corpus. Thus, this chapter only considers texts as samples of 'real language', rather than representations of a particular text type.

We know that the frequency of words in a corpus presents a structural pattern which can be analysed by means of a list of the word types in the corpus, which typically presents three types: (a) high-frequency words, which are prepositions, determiners, pronouns and conjunctions (*grammar* or *function words*), although some can play a lexical role as well, (b) medium-frequency words, most of which are lexical items and (c) words which appear only once in the corpus (*hapax legomena*) (Scott and Tribble 2006: 23).

In order to check whether the texts we use for written input in our English Language module are representative as samples of natural language, we analysed their features in terms of their word frequency. Underlying this, we established a hypothetical link between the frequency of words in a wordlist and the cohesive features of the text, especially in relation to the use, frequency and distribution of conjunctions and other grammar words typically used for signalling clause relations in text (Winter 1971, as quoted by Hoey 1983: 18). The rationale was that even on a small scale, the structural pattern of the wordlists of natural texts should be similar to that of a large corpus of natural texts except that the lists would probably present a higher percentage of *hapax legomena*. If this assumption was true, we expected to be able to identify some differences between the wordlists of natural texts and the wordlists of defective or non-authentic texts. For this purpose, we used the BNC as a reference, which presents a number of about 40 per cent of *hapax legomena*, according to Scott and Tribble (2006: 26).

We started by studying a wordlist produced from an excerpt of 181 running words which was the first part of a narrative text titled 'The Origins of Policing' included in the textbook 'Initiative' (wordlist TA1), and another wordlist from an excerpt of 187 running words which was the first part of the entry 'Edgar Allan Poe: Biography' in Wikipedia (wordlist TB1). As expected, both wordlists presented a very short part consisting of types with a frequency higher than one and a much longer part consisting of *hapax legomena*, as Table 16.1 below shows.

It was soon observed that the first part of the list containing the high-frequency types was much shorter (less than 30 per cent) than the equivalent part in the wordlist of a large corpus (about 60 per cent in the BNC World Edition) and that the grammar or function words in the list were scattered throughout the list and not just concentrated in the first part of the list. Similarly, there were lexical words with a much higher frequency than grammar words. It should be noted here that, in reference to the use of those texts as written input in dictation activities, we think it is important to keep a balance between known and new (lexical) words as well as between words which are more frequently used and words which are less frequent. The reason for this is that a very high percentage of *hapax legomena* involves a very high lexical variety and, therefore, the text is potentially more difficult to the learners than a text presenting a lower lexical variety.

In order to see the effects of text size on the structural patterns of the wordlists, we produced two wordlists from the whole texts from which the excerpts had been taken. In the case of 'The Origins of Policing', we found that the text had been taken from a British website titled 'The Story of Our Police', and we used this text as a reference (wordlist TA2), whereas in the case of 'Edgar Allan Poe: Biography' we took the whole entry in Wikipedia as a reference (wordlist TB2).

Our aim was also to determine whether this criterion of comparison could be useful for identifying defective texts in relation to their authenticity and for this we produced a wordlist of the text discussed by Harmer (1983: 147), 'The Next Morning' (wordlist NM) and compared its features to the other wordlists in order to identify differences. In Table 16.1, the results of this comparison are shown.

As it was not possible to identify any significant difference between the TA, the TB wordlists and the NM wordlist in the features analysed, we decided to analyse them again using this time a reference criterion. By making use of

Table 16.1 Wordlist statistics in relation to text length, completeness and representativeness

Wordlist TA1	Wordlist TA2	Wordlist TB1	Wordlist TB2	Wordlist NM	
181	6,493	187	8,065	166	TOTAL OF RUNNING WORDS
97	1,320	114	1,939	90	TOTAL OF WORD TYPES
69	676	90	1,173	62	NUMBER OF HAPAX LEGOMENA
71.14%	51.21%	78.95%	60.49%	68.90%	% HAPAX LEGOMENA IN TOTAL LIST
30% (43.3% HAPAX)	61% (ANALYSIS RESTRICTED TO THE FIRST 10% OF THE LIST)	24% (33.3% HAPAX)	57% (ANALYSIS RESTRICTED TO THE FIRST 10% OF THE LIST)	27% (48.15% HAPAX)	% WORDS PRESENT FROM THE MOST FREQUENT 100 WORDS (BNC)

a reference wordlist – the list of the most frequent 100 words in the BNC (World Edition) quoted by Scott and Tribble (2006: 24) in this case – this time we were able to identify a different distribution of word types which had been listed in the BNC as most frequent. In Tables 16.2 and 16.3, wordlists TA1 and TB1 show a sequential pattern of the most frequent items in the BNC wordlist which reproduces on a small scale the typical distribution of high-frequency items in a wordlist structure.

In the first ranks of the list (frequency higher than one), we can observe a high presence of items listed as most frequent in the BNC wordlist and a decrease along the ranks corresponding to *hapax legomena.*

As in Table 16.2, we can observe in the first ranks of wordlist TB1 a high presence of items listed as most frequent in the BNC wordlist and their practically total absence from rank 24 ('with') down. Even if wordlist TB1 presents the highest percentage of hapax legomena (78.95 per cent) and the lowest percentage of items coinciding with the BNC wordlist (24 per cent) (see Table 16.1), their distribution is consistently located within the high- and medium-frequency words in the list. In contrast, the pattern in wordlist NM shows a lower density of items in the first part of the list than in wordlists TA1 and TB1 and a tendency to scatter along the hapax legomena, as can be observed in Table 16.4. The density in the hapax legomena coinciding with the items in the BNC wordlist is the highest of the three wordlists, if we compare the percentage of hapax legomena in the list, which is the lowest (68.90 per cent), to the percentage of the hapax coinciding with the BNC wordlist, which is the highest (48.15 per cent).

16.4 Conclusions

It is our belief that we can enhance the accessibility to the written input by making the pedagogic corpus of the course available to the students, and that a better access may help students gain control over their learning because they may be able to cope more efficiently with reading comprehension and vocabulary learning even without the help of the teacher. But corpus methodology should be used prospectively rather than retrospectively, as is done in linguistic research, so that texts can be selected which meet the requirements of authenticity and representativeness more adequately.

In the experience described here, the difficulties for compiling a pedagogic corpus representative enough in terms of the text typologies included involved that many of the choices made in the selection of texts were still based on intuition. When better representations of text types are made available, we will be able to analyse wordlists of singles texts also in relation to typological representativeness.

Table 16.2 Wordlist TA1: Rank and frequency of word types and comparison to the 100 most frequent words in the BNC and rank

Rank	Word	Frequency	BNC Rank
1	THE	21	1
2	AND	6	3
3	IN	6	6
4	OF	6	2
5	TO	6	4
6	WAS	6	12
7	AGAINST	5	
8	ANGLO	4	
9	PEACE	4	
10	THEY	4	30
11	WERE	4	35
12	A	3	5
13	CRIME	3	
14	IT	3	10
15	SAXON	3	
16	THAT	3	8
17	THEIR	3	42
18	ALL	2	40
19	COMMUNITY	2	
20	CRIMES	2	
21	DUTY	2	
22	IF	2	43
23	KING'S	2	
24	LAW	2	
25	NOT	2	26
26	PEOPLE	2	89
27	SETTLED	2	
28	THIS	2	23
29	ACCORDING	1	
30	ACT	1	
31	AGES	1	
32	AMONG	1	
33	AN	1	33
34	BE	1	17
35	BETWEEN	1	
36	BRITAIN	1	
37	BROKE	1	
38	BROKEN	1	
39	BROUGHT	1	
40	BUT	1	25
41	BY	1	21
42	CALLED	1	
43	CATCH	1	
44	CITIZENS	1	
45	COMMUNITIES	1	
46	CUSTOM	1	
47	CUSTOMS	1	
48	DIFFERENT	1	
49	EARLY	1	
50	ENGLAND	1	
51	EXPECTED	1	
52	FOR	1	11
53	GOOD	1	
54	GRADUALLY	1	
55	GREW	1	
56	IDEA	1	
57	INVADERS	1	
58	IS	1	9
59	JUST	1	81
60	KEEP	1	
61	KEEPING	1	
62	KINGS	1	
63	LAST	1	
64	LAWS	1	
65	LIVED	1	
66	MALES	1	
67	MORE	1	52
68	OFFENDERS	1	
69	ORDER	1	
70	ORIGINS	1	
71	OWN	1	
72	POLICING	1	
73	PROPERTY	1	
74	PROTECT	1	
75	RATHER	1	
76	RESPONSIBLE	1	
77	RULE	1	
78	SAID	1	55
79	SAXONS	1	
80	SEE	1	
81	SERIOUS	1	
82	SIXTY	1	
83	SMALL	1	
84	SOME	1	59
85	SOMEONE	1	
86	THAN	1	72
87	THEMSELVES	1	
88	THESE	1	84
89	TIMES	1	
90	TOWNS	1	
91	TWELVE	1	
92	UNDER	1	
93	VICTIM	1	
94	VILLAGES	1	
95	WAVES	1	
96	WHEN	1	53
97	WHOLE	1	

Table 16.3 Wordlist TB1: Rank and frequency of word types and comparison to the 100 most frequent words in the BNC and rank

Rank	Word	Frequency	BNC Rank
1	POE	12	
2	THE	9	1
3	AND	8	3
4	OF	7	2
5	A	5	5
6	HIS	5	29
7	IN	5	6
8	WAS	5	12
9	VIRGINIA	4	
10	#	3	7
11	FOR	3	11
12	RICHMOND	3	
13	TO	3	4
14	AFTER	2	91
15	ALLAN	2	
16	AS	2	16
17	AT	2	20
18	BORN	2	
19	BUT	2	25
20	EDGAR	2	
21	HE	2	18
22	LEFT	2	
23	ONLY	2	
24	WITH	2	15
25	ACADEMY	1	
26	ACTOR	1	
27	ACTRESS	1	
28	AN	1	33
29	APPARENTLY	1	
30	APPOINTMENT	1	
31	ARMY	1	
32	ATTAINING	1	
33	ATTENDING	1	
34	AUNT	1	
35	BALTIMORE	1	
36	BAPTIZED	1	
37	BEFORE	1	
38	BEGAN	1	
39	BIOGRAPHY	1	
40	BOSTON	1	
41	BOTH	1	
42	CLEMM	1	
43	COMPEL	1	
44	DAUGHTER	1	
45	DAVID	1	
46	DECEMBER	1	
47	DELIBERATELY	1	
48	DIED	1	
49	DISCHARGED	1	
50	DISMISSAL	1	
51	DISOBEYED	1	
52	EDITING	1	
53	ELIZA	1	
54	ENGLAND	1	
55	ENLISTED	1	
56	FATHER	1	
57	FICTION	1	
58	HER	1	36
59	HIM	1	66
60	HIMSELF	1	
61	HOME	1	
62	INTO	1	64
63	ISSUE	1	
64	JOHN	1	
65	JR	1	
66	LITERARY	1	
67	MAJOR	1	
68	MARA	1	
69	MARYLAND	1	
70	MASSACHUSSETTS	1	
71	MAY	1	
72	MEANS	1	
73	MERCHANT	1	
74	MESSENGER	1	
75	MILITARY	1	
76	MOTHER	1	
77	MOVED	1	
78	NAME	1	
79	NEXT	1	
80	ON	1	13
81	ONE	1	37
82	OERS	1	
83	PARENTS	1	
84	PERRY	1	
85	POINT	1	
86	PRIVATE	1	
87	RANK	1	
88	RECEIVED	1	
89	REGISTERED	1	
90	S	1	
91	SCHOOLS	1	
92	SERGEANT	1	

(*Continued*)

Rank	Word	Frequency	BNC Rank
93	SERVING	1	
94	SON	1	
95	SOUTHERN	1	
96	STAYED	1	
97	SUCCESSFUL	1	
98	SUPPORTING	1	
99	TAKEN	1	
100	THOMAS	1	
101	THREE	1	
102	TWO	1	62
103	U	1	
104	UNIVERSITY	1	
105	US	1	
106	USED	1	
107	USING	1	
108	W	1	
109	WEST	1	
110	WHEN	1	53
111	WHITE	1	
112	WIDOWED	1	
113	YEAR	1	
114	YEARS	1	

Table 16.4 Wordlist NM: Rank and frequency of word types and comparison to the 100 most frequent words in the BNC and rank

Rank	Word	Frequency	BNC Rank
1	THE	17	1
2	IS	10	9
3	AND	8	3
4	TO	6	4
5	ARE	5	22
6	IT	4	10
7	KATE	4	
8	VERY	4	87
9	A	3	5
10	FLAT	3	
11	GOING	3	
12	HAVE	3	24
13	ON	3	13
14	ALL	2	40
15	BUT	2	25
16	CLEAN	2	
17	COFFEE	2	
18	GIRLS	2	
19	GLASSES	2	
20	GOOD	2	
21	GOT	2	
22	IN	2	6
23	LAST	2	
24	NIGHT	2	
25	PENNY	2	
26	ROOM	2	
27	SHE	2	
28	THERE	2	39
29	ASHTRAYS	1	
30	BACK	1	96
31	BAD	1	
32	BLACK	1	
33	CAN	1	51
34	CHAIRS	1	
35	CLOUDS	1	
36	COMING	1	
37	COOK	1	
38	CUPS	1	
39	DAY	1	
40	DIDN'T	1	
41	DIRTY	1	
42	DRINKING	1	
43	DULL	1	
44	EVERYTHING	1	
45	FLOOR	1	
46	FOR	1	11
47	FRANKLIN	1	
48	FRIEND	1	
49	FULL	1	
50	GUITAR	1	
51	HARD	1	
52	HE	1	18
53	HER	1	36
54	HERE	1	
55	INTO	1	64
56	ITS	1	63

(*Continued*)

Table 16.4 Continued

Rank	Word	Frequency	BNC Rank
57	JOHNSON	1	
58	JOYCE	1	
59	KITCHEN	1	
60	LEFT	1	
61	LOT	1	
62	MAKE	1	
63	MUCH	1	
64	NEW	1	82
65	OF	1	2
66	OUTSIDE	1	
67	OVER	1	76
68	PLACE	1	
69	PLATES	1	
70	PUT	1	
71	QUICKLY	1	
72	RAINING	1	
73	RODNEY	1	

Rank	Word	Frequency	BNC Rank
74	SAUCERS	1	
75	SHELVES	1	
76	SKY	1	
77	SLEEP	1	
78	SLOWLY	1	
79	TABLE	1	
80	TAKE	1	
81	THEIR	1	42
82	THEY	1	30
83	TIDY	1	
84	TIRED	1	
85	TWO	1	62
86	UNDER	1	
87	UNTIDY	1	
88	WALKING	1	
89	WASH	1	
90	WET	1	

Our contribution is tentative in its application of corpus methodology for the study of textual features of single texts. Even if the structure of wordlists is not reliable as an absolute criterion to accurately differentiate authentic texts from artificial, non-authentic or defective ones, the results obtained in our analysis seem to suggest that our procedure may be useful to identify if a text is closer or further from what we can consider representative of real language in terms of the frequency of words and the balance and distribution of the function and the lexical words in the wordlist. In spite of this, further research is needed in order to establish a more accurate relationship between a sample text and its corresponding text type/s, as well as of the kind of relationship that may (or may not) hold between the structure of a wordlist and the text cohesion.

References

De Beaugrande, R. A. (1980), *Text, Discourse and Process.* Norwood: Ablex.

De Beaugrande, R. A. and Dressler, W. U. (1981), *Introduction to Text Linguistics.* New York: Longman.

Dubin, F., Eskey, D. and Grabe, W. (1983), *Teaching Second Language Reading for Academic Purposes.* Reading, MA: Addison-Wesley.

Gass, S. M. (1990), 'Second and foreign language learning: Same, different or none of the above?', in Van Patten, B. and Lee, J. F. (eds), *Second Language Acquisition-Foreign Language Learning.* Clevendon: Multilingual Matters, pp. 34–44.

Gass, S. M. and Mackey, A. (2006), 'Input, interaction and output: an overview', in Bardovi-Harlig, K. and Dörnyei, Z. (eds), *Themes in SLA Research. AILA Review 19.* Amsterdam: John Benjamins, pp. 3–17.

Gass, S. M. and Selinker, L. (2001), *Second Language Acquisition: An Introductory Course.* Mahwah, NJ: Lawrence Erlbaum Associates.

Harmer, J. (1983), *The Practice of English Language Teaching.* London: Longman.

Hoey, M. (1983), *On the Surface of Discourse.* London: George Allen and Unwin.

—(2005), *Lexical Priming: A New Theory of Words and Language.* Abingdon, Oxon: Routledge.

Hunston, S. (2002), *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Johns, T. (1994), 'The text and its message', in Coulthard, M. (ed.), *Advances in Written Text Analysis.* London: Routledge, pp. 102–116.

Krashen, S. D. (1985), *The Input Hypothesis: Issues and Implications.* New York: Longman.

Lee, J. F. and Van Patten, B. (2003), *Making Communicative Language Teaching Happen.* New York: McGraw Hill.

McEnery, T. and Wilson, A. (2001), *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Paltridge, B. (2006), *Discourse Analysis.* London: Continuum.

Scott, M. (2004), *WordSmith Tools.* Oxford: Oxford University Press.

Scott, M. and Tribble, C. (2006), *Textual Patterns.* Amsterdam: John Benjamins.

Stubbs, M. (1996), *Text and Corpus Analysis.* Oxford: Blackwell.

Van Patten, B. (2003), *From Input to Output.* Boston: McGraw-Hill.

This page intentionally left blank

Part Four

Multimodality: Corpus Tools and Language Processing Technology

This page intentionally left blank

Chapter 17

A Generic Tool for Annotating Tei-Compliant Corpora: An ELT-Based Approach to Corpus Annotation[1]

José Maria Alcaraz Calero, Pascual Pérez-Paredes and
Encarnación Tornero Valero
University of Murcia, Spain

17.1 Introduction: Annotation and Language Pedagogy

Annotation plays a significant role in Data Driven Learning (DDL). If annotation is pedagogically oriented, this role may be even more relevant. In cognitive-mediated learning processes, such as foreign language learning, it is usual to find tools, materials and pedagogic media which guide and help the learner to understand the information she is trying to learn. These pedagogic tools may be of different nature. They could be elements such as a slide, a photograph or a figure; teaching media such as a board, a computer or an overhead projector, and learning tools such as computer programs, books or any other conventional source of knowledge. In this context, a computer could be considered as a tool, a medium as well as a piece of material.

Annotation tools make it possible that the user enrich texts or a linguistic corpus with additional information or meta-information which might be of interest to researchers or applied linguists. This possibility may be helpful in developing high-quality pedagogic materials ranging from plain texts to well-designed annotation of information. This was a very important concern in SACODEYL.

Leech (1993) states that generic annotation does not necessarily lead to high quality standards. It is necessary first to design and select the characteristics of the text we want to annotate depending on what the annotation is going to be used for. However, the challenge for a generic tool is tremendously ambitious as the possible applications that linguists could expect are galore. For example, Levy (1997, 49–50) has classified the general field of CALL in 24 sub-fields, such as language data processing, language teaching methodology, linguistics or second language acquisition among others. The scope of application of a given tool, such as a test, will necessarily determine the way and the qualities of the annotation process. In other words, the scope of the application of a text or a corpus precedes the design of the annotation approach. Thus, in the field of language teaching topic-driven approaches (Braun 2006) highlight the most relevant characteristics from a pedagogic perspective.

If we analyse some of the current annotation tools (Pérez-Paredes and Alcaraz 2007; Pérez-Paredes et al. 2009, forthcoming), we can observe that most of them carry out the annotation process from the perspective of the researcher in linguistics, who needs textual annotation on morpho-syntactic level. Hence, almost all the annotators are specialized in carrying out a morphological and syntactic text annotation in a manual, semi-automatic or automatic way with, inter alia, Xeros-EAGLES (Cutting and Pederson 1993), TreeTagger (Schmid 1995), CLAWS (Garside 1987) and FreeLing (Atserias et al. 2006).

Despite this majority of tools, there are now annotation solutions which allow the use of a more general and open process of annotation, not only a specific, closed set of labels or tags. This is the logical consequence of changing the representation system of the annotation process from the traditional way described above to a new approach for representing the information based on XML. Therefore, there is an increasing amount of new tools based on XML-aware annotation, among which we find Calisto (Bayer et al. 2006), LACITO (Jacobson 2006) and LT XML (Grover et al. 2006).

However, even though there are now more XML-based tools which allow an annotation process, these tools do not generally enable the user to define which characteristics to annotate. It would be then ideal that these XML tools were generic and extensible. We can find some proposals in this context, such as Dexter (Garretson 2006) or EXMERaLDA (Schmidt 2004). These tools are a step forward in getting annotation tools to be used in a generic way in different contexts of knowledge representation, e.g. foreign language teachers who want to prepare teaching materials. However, these tools are designed with a research aim in mind and it is not easy to use them pedagogically. Furthermore, these solutions code the annotation of the linguistic corpora using their own XML schema, not following any standardized[2] mechanism to represent the linguistic annotation of data and metadata.

In the language teaching context, we need thereby a tool which enables a pedagogic annotation to facilitate the development of high-quality pedagogic resources to be used by both learners and teachers in the context of DDL. These pedagogic resources may be used later for data-driven teaching or just for creating CALL exercises. What is more, it would be advisable to produce pedagogic resources which could be reused by other applications, which implies using a universally accepted academic standard to represent the information of the annotation (Ward 2002). This way, any application using standard-compliant XML annotated corpora could reuse these new pedagogic Cushion (2004).

The bottom line here is that if pedagogy can be annotated, language learning resources which make use of corpus-based materials are more likely to be implemented in the classroom.

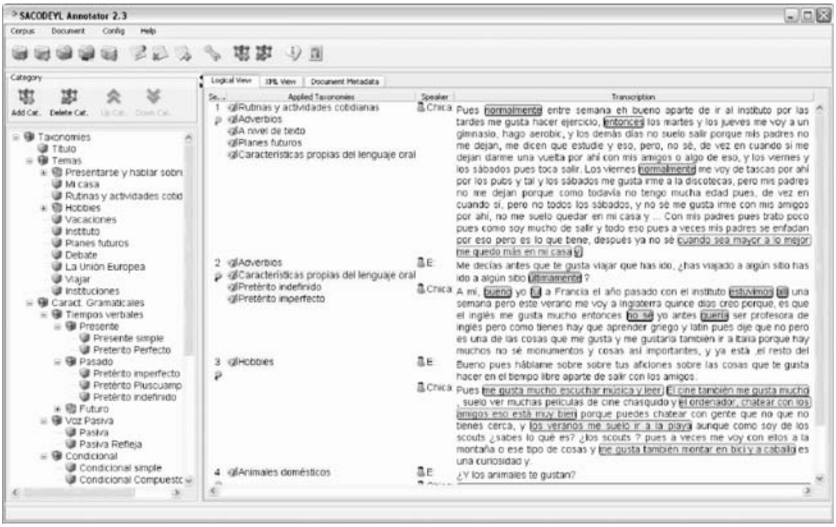
17.2 SACODEYL Annotator

17.2.1 Using SACODEYL Annotator

SACODEYL Annotator[3] has been developed as part of undergoing work on System Aided Compilation and Open Distribution of European Youth Language. The aim of the tool is to give the multilingual and multinational SACODEYL team the means to annotate seven different corpora. The main distinctive feature of the tool is that it has been developed originally to implement pedagogical annotation, which means that we have not adapted or converted other tools that may have been developed with other aims. Another design principle has been that of providing ease of use for the annotators as well as power and robustness in terms of the output data.

It is expected that different users will have an interest in the tool, from a computational linguist interested in annotating texts to a language learner that wishes to navigate the features annotated in a corpus, and thus become more acquainted with the sort of meta-information that has been included by the annotators. Certainly language teachers will show a natural inclination towards material selection and/or development. All of these users will find a very friendly interface that greatly facilitates both the annotation as well as the navigation process. Let us examine this interface in detail.

Figure 17.1 shows the distribution of the main window of the tool. On the left, we can find the annotation structure established by the annotator(s). On the right, we find the annotation performed on a text, in the example



Se...	Applied Taxonomies	Speaker	Transcription
1	Rutinas y actividades cotidianas Adverbios A nivel de texto Planes futuros Características propias del lenguaje oral	Chica	Pues normalmente entre semana eh bueno aparte de ir al instituto por las tardes me gusta hacer ejercicio, entonces los martes y los jueves me voy a un gimnasio, hago aerobic, y los demás días no suelo salir porque mis padres no me dejan, me dicen que estudie y eso, pero, no sé, de vez en cuando si me dejan darme una vuelta por ahí con mis amigos o algo de eso, y los viernes y los sábados pues toca salir. Los viernes normalmente me voy de tascas por ahí por los pubs y tal y los sábados me gusta irme a la discotecas, pero mis padres no me dejan porque como todavía no tengo mucha edad pues, de vez en cuando sí, pero no todos los sábados, y no sé me gusta irme con mis amigos por ahí, no me suelo quedar en mi casa y ... Con mis padres pues trato poco pues como soy mucho de salir y todo eso pues a veces mis padres se enfadan por eso pero es lo que tiene, después ya no sé cuando sea mayor a lo mejor me quedo más en mi casa y
2	Adverbios Características propias del lenguaje oral Pretérito indefinido Pretérito imperfecto	E	Me decías antes que te gusta viajar que has ido, ¿has viajado a algún sitio has ido a algún sitio últimamente ?
		Chica	A mí, bueno yo fui a Francia el año pasado con el instituto estuvimos allí una semana pero este verano me voy a Inglaterra quince días creo porque, es que el inglés me gusta mucho entonces no sé yo antes quería ser profesora de inglés pero como tienes hay que aprender griego y latín pues dije que no pero es una de las cosas que me gusta y me gustaría también ir a Italia porque hay muchos no sé monumentos y cosas así importantes, y ya está ,el resto del
3	Hobbies	E	Bueno pues háblame sobre sobre tus aficiones sobre las cosas que te gusta hacer en el tiempo libre aparte de salir con los amigos.
		Chica	Pues me gusta mucho escuchar música y leer. El cine también me gusta mucho , suelo ver muchas películas de cine chasquido y el ordenador, chatear con los amigos eso está muy bien porque puedes chatear con gente que no que no tienes cerca, y los veranos me suelo ir a la playa aunque como soy de los scouts ¿sabes lo qué es? ¿los scouts ? pues a veces me voy con ellos a la montaña o ese tipo de cosas y me gusta también montar en bici y a caballo es una curiosidad y.
4	Animales domésticos	E	¿Y los animales te gustan?

FIGURE 17.1 SACODEYL Annotator: a multi-purpose generic tool

above an oral text. Let us concentrate first on the left–hand side of the application.

This area of the tool clearly shows the potential of SACODEYL Annotator to become a truly generic tool for problem-oriented tagging (McEnery and Wilson 1996). This is possible as annotators are given the chance to decide on the tags they want to work with, and the tool takes care of the rest, i.e. the application performs management, extension, addition, modification and suppression functions on this set of tags. This is a key point in the development of a multi-purpose application that seeks to meet the needs of a wide range of language professionals.

On the right-hand side we can find the annotation of an oral text. The area is divided in four different columns. In the first column we find information as to the *section* of the text that is being annotated (section$_1$...... section$_n$). In SACODEYL this is a crucial issue, as the different texts of a corpus are segmented bearing a didactic exploitation in mind. For a further discussion on the idea of section see Braun (2005, 2007), Pérez-Paredes et al. (2009, forthcoming) and Pérez-Paredes and Alcaraz (2007). On the second column (under Applied Taxonomies) we find the annotation that has been assigned to a section, while on the third it is possible to identify the speaker or contributor. Finally, on the fourth column we can see the text proper. Here you can note some highlighting which matches the relationship established between the tags and the stretch of text that motivated the adscription of a particular sub-taxonomy or tag on a section. This is optional, i.e. a tag can be assigned and the annotator may decide not to establish a link between the language data and the annotation. In SACODEYL we call this highlighted stretch of text a *keyword*.

Figure 17.2 shows how a search tool may render the annotation performed on section 1 of the text previously displayed in Figure 17.1.

As seen above, the tool has been developed to meet the needs of a very wide range of users, and as a consequence no *a priori* knowledge of CL is needed in order to start annotation right away. The tool is very easy to use: tag assignment is performed through drag and drop and keyword assignment through select and click basic operations. To facilitate this process the application filters out the information shown on screen and so users can decide which highlighted keywords they want to see or hide. Secure deleting of the annotation is also provided. The tool is so intuitive that even learners with no CL background whatsoever might use it to navigate the annotation.

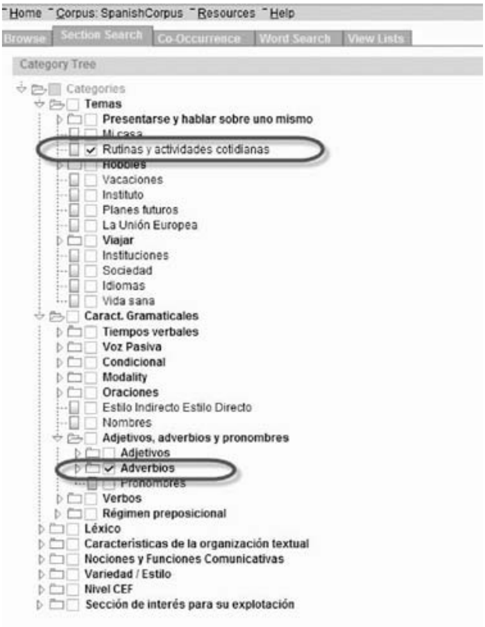
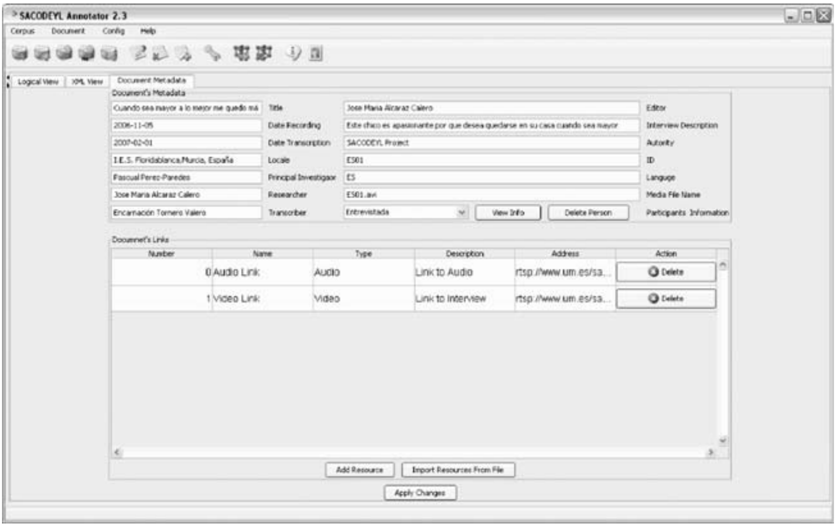


FIGURE 17.2 Taxonomy tree as rendered by SACODEYL Search Tool

But there is more to the tool. SACODEYL Annotator allows for the management of multiple corpus files, an underlying principle in SACODEYL. Users can thus create, import or select different corpora and perform the same or different annotation schemes on each of them. The user may work on texts of different nature, spoken or written, monologic or dialogic. The tool has used UNICODE standard (Needleman 2000) which gives it truly multilingual power. For SACODEYL this means that all seven language corpora (DE, EN, ES, FR, IT, LT, RO) can be annotated with the same tool, but for the generic potential discussed above it means that any corpus of any language could be annotated with it, from Chinese to Korean, just to cite two important non-Western languages. Also, it must be stressed that, apart form the language of the corpus, SACODEYL Annotator will read files encoded according to different standards: ANSI, ASCII, ISO, UFT-8 and Unicode.

A key issue in CL is how meta-data are handled. SACODEYL Annotator allows that the different XML entities in a corpus, i.e. texts, be assigned all kinds of meta-information such as title, author, editor, date, participants, description, language, etc. Figure 17.3 shows how this is done.

An interesting issue is the possibility for annotators to incorporate external resources or data to a particular section. In the framework of SACODEYL, it has been envisaged that this particular feature will be used



Document's Metadata

Cuando sea mayor a lo mejor me quedo má	Title	Jose Maria Alcaraz Calero	Editor
2006-11-05	Date Recording	Este chico es apasionante por que desea quedarse en su casa cuando sea mayor	Interview Description
2007-02-01	Date Transcription	SACODEYL Project	Autority
I.E.S. Floridablanca,Murcia, España	Locale	ES01	ID
Pascual Perez-Paredes	Principal Investigaor	ES	Languge
Jose Maria Alcaraz Calero	Researcher	ES01.avi	Media File Name
Encarnación Tornero Valero	Transcriber	Entrevistada	Participants Information

Documnet's Links

Number	Name	Type	Description	Address	Action
0	Audio Link	Audio	Link to Audio	rtsp://www.um.es/sa...	Delete
1	Video Link	Video	Link to Interview	rtsp://www.um.es/sa...	Delete

FIGURE 17.3 SACODEYL Annotator meta-data screen

to enrich the corpus pedagogically and feed the DDL web system with links to web services such as pages, multimedia, textual resources and FLT activities. In SACODEYL we have for the most part used this feature to feed our DDL web system, although learners or teachers may very well use it to enrich their language experiences in different ways.

It is worth mentioning that most annotation tools will not let the users modify or edit the linguistic data that is being annotated. In the framework of SACODEYL this power feature has performed a very important role in securing consistency and accuracy. These textual alterations can be easily done preserving the annotated tags, which no doubt facilitates the transcription-annotation-data delivery process. This is another feature that will be of interest to different professionals in a wide array of fields.

So far we have discussed the usefulness of SACODEYL Annotator in the annotation of corpus-based learning resources. However, the tool has been designed with a generic use in mind. SACODEYL Annotator is language input-independent, as different languages and text typologies can be annotated. A case in point is spoken language where different contributors can be represented by the tool interface, making the annotation and navigation process more intuitive. Also, the tool is discipline-independent due to the fact that annotators are given flexibility to establish the use that the corpus will be put to and, in accordance, the discipline where the annotated corpus is to be delivered. It is interesting to underscore the relevance that may have for non-XML-aware users, the fact that both annotation and 'taxonomy definition' can be performed with the same tool and on the same screen interface.

Within the field of language and linguistics, SACODEYL Annotator allows very refined uses and applications. Some of these include translations and interpretation studies, general and specific language learning purposes, computational studies, creation of folksonomies and the generation of ontologies. Last but not least, SACODEYL Annotator is multi-user oriented as it may cater for different and simultaneous needs, ranging from those of teachers, learners and materials developers.

Having discussed the generic potential of the tool, let us move to gloss over the technology that makes these generic uses possible in SACODEYL Annotator: the Text Encoding Initiative (Burnard 1995).

17.2.2 TEI as standardization method

One of the challenges to be met by system developers is that of standards and normalization. In our case the main issue was to decide on the way in which our linguistic data and our annotation should be stored. As discussed by Pérez-Paredes et al. (2009, forthcoming), our aim was to develop tools and products that could be reusable and, in this way, contribute significantly to the ever-growing movement of open-content. We were aware of the fact that existing *ad hoc solutions* could provide us with tools that could do the job, but we still felt that, given the nature of our initiative, we should strive for standardization as a goal.

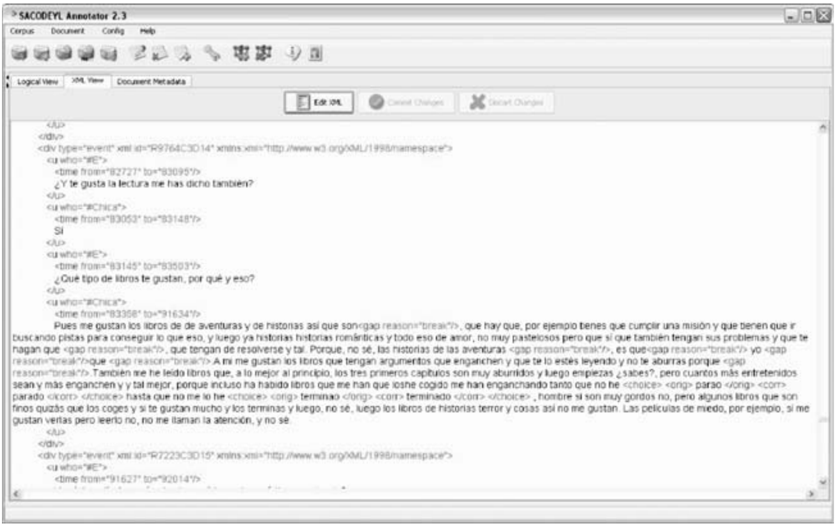
Having such a goal in mind, we decided to use the standard XML representation of the Text Encoding Initiative. TEI is a widely spread standard for text encoding that provides an XML schema for storing corpus and the metadata information associated to them. The main target of TEI is to offer a common framework for text encoding and cover all the different aspects and features that could be associated with any text or corpus. This way, spoken discourse features such as pauses or breaks, can be treated uniformly across different software applications. In the case of written texts, structural divisions of a text at different layers such as, documents, sections, paragraphs, sentences or words, bibliography description, tables of content, tagset description, and metadata can be conveniently stored in standard XML with a wide range of tools.

The number of XML tools that support TEI is increasing by days: oXygen by SyncRO Soft (2007), OpenOffice (Haugland and Jones 2002), TEI E-macs (Lease 2005), Anastasia Scholarly, by Digital Editions (2004), TEI Publisher (Lease 2005) and, inter alia, Xaira (Burnard 1995). SACODEYL annotator benefits from this standard coding at the same time that provides users with an extremely intuitive interface. Our SACODEYL XML files are corpus files that contain the language data, the language data structure information and the annotation proper (Pérez-Paredes et al. 2009, forthcoming).

It must be stressed that the tool easily adapts to the needs of advanced users or computational linguists who wish to work on the XML code itself. This feature allows advanced XML-aware users the possibility to perform changes on the very code. This can be better appreciated in Figure 17.4.

17.3 Annotation in the Foreign Language Classroom

Adapting texts and corpora to the needs of the language classroom is an area where SACODEYL Annotator may be instrumental. It is well-known that the text encoding initiative allows the subdivision of a text into meaningful fragments for analytic purposes, a feature which has been



```
    </u>
  </div>
  <div type="event" xml:id="R9764C3D14" xmlns:xml="http://www.w3.org/XML/1998/namespace">
    <u who="#E">
      <time from="82727" to="83095"/>
      ¿Y te gusta la lectura me has dicho también?
    </u>
    <u who="#Chica">
      <time from="83053" to="83148"/>
      Sí
    </u>
    <u who="#E">
      <time from="83145" to="83503"/>
      ¿Qué tipo de libros te gustan, por qué y eso?
    </u>
    <u who="#Chica">
      <time from="83358" to="91634"/>
      Pues me gustan los libros de de aventuras y de historias así que son<gap reason="break"/>, que hay que, por ejemplo tienes que cumplir una misión y que tienen que ir
buscando pistas para conseguir lo que eso, y luego ya historias historias románticas y todo eso de amor, no muy pastelosos pero que sí que también tengan sus problemas y que te
hagan que <gap reason="break"/>, que tengan de resolverse y tal. Porque, no sé, las historias de las aventuras <gap reason="break"/>, es que<gap reason="break"/> yo <gap
reason="break"/>que <gap reason="break"/> A mi me gustan los libros que tengan argumentos que enganchen y que te lo estés leyendo y no te aburras porque <gap
reason="break"/>.También me he leído libros que, a lo mejor al principio, los tres primeros capítulos son muy aburridos y luego empiezas ¿sabes?, pero cuantos más entretenidos
sean y más enganchen y y tal mejor, porque incluso ha habido libros que me han que loshe cogido me han enganchando tanto que no he <choice> <orig> parao </orig> <corr>
parado </corr> </choice> hasta que no me lo he <choice> <orig> terminao </orig> <corr> terminado </corr> </choice> , hombre si son muy gordos no, pero algunos libros que son
finos quizás que los coges y si te gustan mucho y los terminas y luego, no sé, luego los libros de historias terror y cosas así no me gustan. Las películas de miedo, por ejemplo, si me
gustan verlas pero leerlo no, no me llaman la atención, y no sé.
    </u>
  </div>
  <div type="event" xml:id="R7223C3D15" xmlns:xml="http://www.w3.org/XML/1998/namespace">
    <u who="#E">
      <time from="91627" to="92014"/>
```

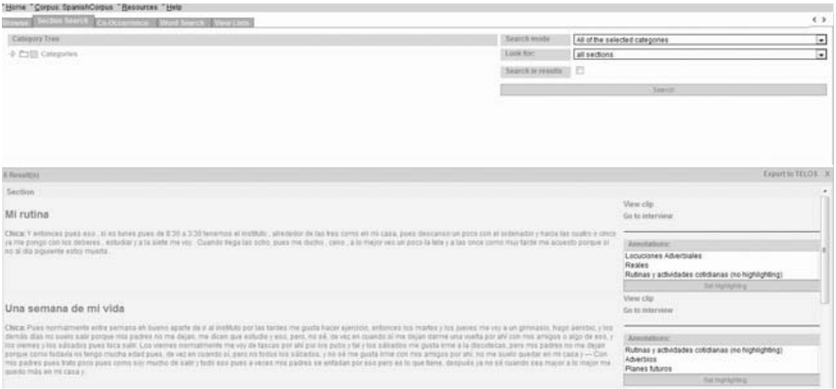
FIGURE 17.4 SACODEYL Annotator XML edition and exploration screen

conveniently adapted into SACODEYL Annotator for the representation of our own section element. The annotation categories are declared in a *<classDecl>* element which allows for creating extensible subcategories as deemed by the annotators. The section element has been integrated into the *<div>* tags. An example of the categories annotated on a section of the Spanish corpus follows:

> *<div decls="#routinesTopic #Adverbios #TextOrganizationFeatures #futurePlan-Topic #Tipical OffSpokenLang" type="event" xml:id="R2738C0D1">*
> *<head>Una semana de mi vida</head>*

An important feature of the SACODEYL system is that every corpus can be looked upon and searched dynamically in the sense that each corpus informs our search tool about the different annotated categories that have been applied to the corresponding sections. Figure 17.5 exemplifies this point.

This is a major breakthrough in the customization of corpus-based language learning and teaching. To date, language professionals have been prompted to make use of materials whose primary orientation was linguistic research. In this sense, annotation can give language learning and teaching stakeholders the chance to adapt corpus methods and



Mi rutina

Chica: Y entonces pues eso , si es lunes pues de 8:30 a 3:30 tenemos el instituto , alrededor de las tres como en mi casa, pues descanso un poco con el ordenador y hacia las cuatro o cinco ya me pongo con los deberes , estudiar y a la siete me voy . Cuando llega las ocho, pues me ducho , ceno , a lo mejor veo un poco la tele y a las once como muy tarde me acuesto porque si no al día siguiente estoy muerta.

Annotations:
Locuciones Adverbiales
Reales
Rutinas y actividades cotidianas (no highlighting)

Una semana de mi vida

Chica: Pues normalmente entre semana eh bueno aparte de ir al instituto por las tardes me gusta hacer ejercicio, entonces los martes y los jueves me voy a un gimnasio, hago aerobic, y los demás días no suelo salir porque mis padres no me dejan, me dicen que estudie y eso, pero, no sé, de vez en cuando sí me dejan darme una vuelta por ahí con mis amigos o algo de eso, y los viernes y los sábados pues toca salir. Los viernes normalmente me voy de tascas por ahí por los pubs y tal y los sábados me gusta irme a la discotecas, pero mis padres no me dejan porque como todavía no tengo mucha edad pues, de vez en cuando sí, pero no todos los sábados, y no sé me gusta irme con mis amigos por ahí, no me suelo quedar en mi casa y — Con mis padres pues trato poco pues como soy mucho de salir y todo eso pues a veces mis padres se enfadan por eso pero es lo que tiene, después ya no sé cuando sea mayor a lo mejor me quedo más en mi casa y.

Annotations:
Rutinas y actividades cotidianas (no highlighting)
Adverbios
Planes futuros

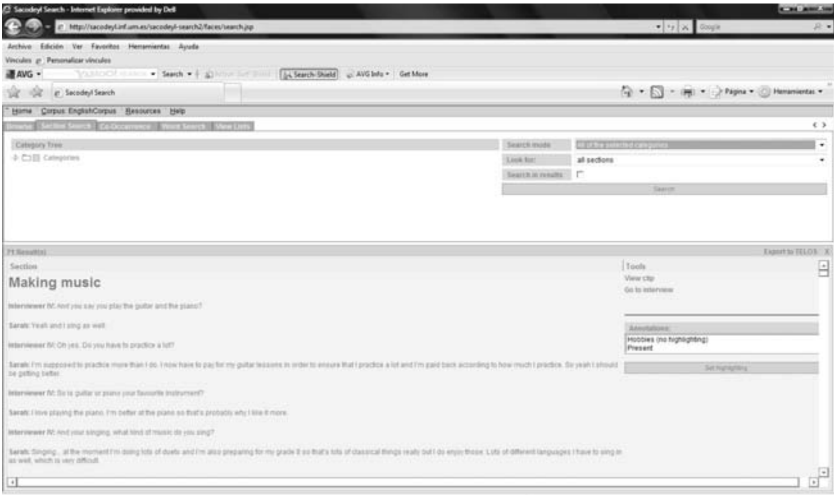
Figure 17.5 Annotation in action as displayed by SACODEYL Search Tool

resources to the type of authenticity that is sought after in the language classroom. The search interface shown on Figure 17.5 can be reached from the SACODEYL website http://www.um.es/sacodeyl or from the SACODEYL dedicated server http://www.purl.org/sacodeyl/search. This search interface dynamically reads the annotation and renders a query tree based on the information which has been provided by the annotators. This is a ready-to-use example of how pedagogic annotation can be used in varied language learning contexts. In the case of SACODEYL the aim was to develop annotation which could serve as a pedagogic mediator in the process of foreign language learning of young Europeans. Although the possibilities are unlimited, the annotation categories which were used by the seven language teams in SACODEYL focused very significantly on topics, CEFRL levels and the features of spoken language.

Learners and teachers interested in evidence-driven language learning can use the power of annotation to query multimodal corpora. Say, a group of learners is interested in learning more about the hobbies topic area.

The Search interface (Figure 17.6) displays 71 results for this corpus, which is probably way too many. Learners may want now to refine their search and establish technology as a subset within the results (Figure 17.7).

Now the learners get seven sections where Hobbies > Using technologies are used. In a way, we have applied CL methods to the notion of topic and pedagogic section, which we expect to be of usefulness in most FLT contexts. Learners have now sections which deal with a very restricted thematic area and which can be further searched. Figure 17.8 shows how one of these sections has been annotated as displaying modality, while retaining the thematic feature.


Sacodeyl Search - Internet Explorer provided by Dell
http://sacodeyl.inf.um.es/sacodeyl-search2/faces/search.jsp
Archivo Edición Ver Favoritos Herramientas Ayuda
Vínculos Personalizar vínculos
AVG Search Search-Shield AVG Info Get More
Sacodeyl Search
Página Herramientas
Home Corpus: EnglishCorpus Resources Help
Browse Section Search Co-Occurrence Word Search View Lists
Category Tree
Categories
Search
71 Result(s)
Export to TELOS
Section
Tools
View clip
Go to interview
Annotations:
Set highlighting

Search mode: All of the selected categories
Look for: all sections
Search in results ☐

Hobbies (no highlighting)
Present

Making music

Interviewer IV: And you say you play the guitar and the piano?

Sarah: Yeah and I sing as well.

Interviewer IV: Oh yes. Do you have to practice a lot?

Sarah: I'm supposed to practice more than I do. I now have to pay for my guitar lessons in order to ensure that I practice a lot and I'm paid back according to how much I practice. So yeah I should be getting better.

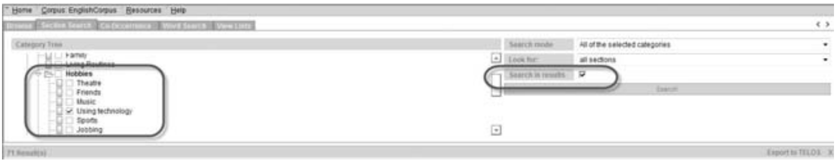
Interviewer IV: So is guitar or piano your favourite instrument?

Sarah: I love playing the piano. I'm better at the piano so that's probably why I like it more.

Interviewer IV: And your singing, what kind of music do you sing?

Sarah: Singing... at the moment I'm doing lots of duets and I'm also preparing for my grade 8 so that's lots of classical things really but I do enjoy those. Lots of different languages I have to sing in as well, which is very difficult.

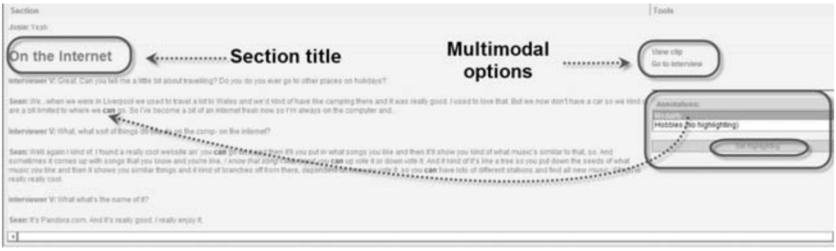
Figure 17.6 Searching for sections where 'Hobby' has been annotated



☑ Using technology

Search mode: All of the selected categories
Look for: all sections
Search in results ☑

Figure 17.7 Refining a search



Josie: Yeah

Interviewer V: Great. Can you tell me a little bit about travelling? Do you do you ever go to other places on holidays?

Sean: We... when we were in Liverpool we used to travel a lot to Wales and we'd kind of have like camping there and it was really good. I used to love that. But we now don't have a car so we kind of are a bit limited to where we **can** go. So I've become a bit of an internet freak now so I'm always on the computer and...

Interviewer V: What, what sort of things do you do on the comp- on the internet?

Sean: Well again I kind of, I found a really cool website an' you **can** go on it and then it'll you put in what songs you like and then it'll show you kind of what music's similar to that, so. And sometimes it comes up with songs that you know and you're like, I know that song! And then you **can** up vote it or down vote it. And it kind of it's like a tree so you put down the seeds of what music you like and then it shows you similar things and it kind of branches off from there, dependent on how you vote it, so you **can** have lots of different stations and find all new music. Which is really really cool.

Interviewer V: What what's the name of it?

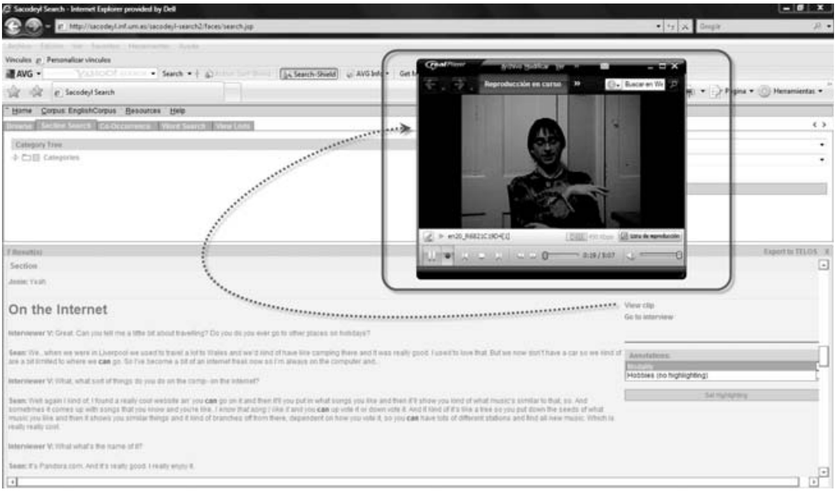
Sean: It's Pandora.com. And it's really good. I really enjoy it.

Figure 17.8 A section in the SACODEYL Search Tool

This section called 'On the Internet', can be viewed in isolation or in the context of the whole interview/text and, interestingly, in red, displays a feature which the annotation team has found of interest from a pedagogic perspective: modality. Now the search has been expanded into Hobbies > Using technologies > Modality by way of the suggested features added by the annotators. If the learner is interested in the section, she can watch it, as shown in Figure 17.9.

Of course, word search is central to the application. Figure 17.10 shows 'Facebook' search in SACODEYL.

Using this search, learners could build up a sense of the contexts where one could expect to find Facebook in discourse, i.e. being a member of Facebook, find Facebook really good, Facebook is for slightly older people, go on Facebook, etc., which while not being representative of English


Sacodeyl Search - Internet Explorer provided by Dell
http://sacodeyl.inf.um.es/sacodeyl-search2/faces/search.jsp
Vínculos Personalizar vínculos
AVG Search Search-Shield AVG Info
Sacodeyl Search
Home Corpus: EnglishCorpus Resources Help
Browse Section Search Co-Occurrence Word Search View Lists
Category Tree
Categories
RealPlayer
Reproducción en curso
Buscar en Web
en20_R6821C19D-[1]
0:19 / 5:07
Lista de reproducción
Export to TELOS
View clip
Go to interview
Annotations
Hobbies (no highlighting)
Set highlighting
Página Herramientas

7 Result(s)

Section

Jessie: Yeah

On the Internet

Interviewer V: Great. Can you tell me a little bit about travelling? Do you do you ever go to other places on holidays?

Sean: We... when we were in Liverpool we used to travel a lot to Wales and we'd kind of have like camping there and it was really good. I used to love that. But we now don't have a car so we kind of are a bit limited to where we **can** go. So I've become a bit of an internet freak now so I'm always on the computer and...

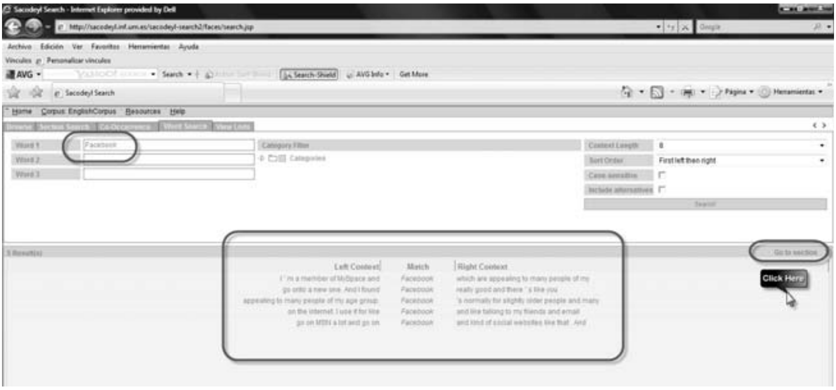
Interviewer V: What, what sort of things do you do on the comp- on the internet?

Sean: Well again I kind of, I found a really cool website an' you **can** go on it and then it'll you put in what songs you like and then it'll show you kind of what music's similar to that, so. And sometimes it comes up with songs that you know and you're like.. I know that song I like it and you **can** up vote it or down vote it. And it kind of it's like a tree so you put down the seeds of what music you like and then it shows you similar things and it kind of branches off from there, dependent on how you vote it, so you **can** have lots of different stations and find all new music. Which is really really cool.

Interviewer V: What what's the name of it?

Sean: It's Pandora.com. And it's really good. I really enjoy it.

FIGURE 17.9 A multimodal section


Sacodeyl Search - Internet Explorer provided by Dell
http://sacodeyl.inf.um.es/sacodeyl-search2/faces/search.jsp
Archivo Edición Ver Favoritos Herramientas Ayuda
Vínculos Personalizar vínculos
AVG Search Search-Shield AVG Info Get More
Sacodeyl Search
Página Herramientas
Home Corpus: EnglishCorpus Resources Help
Browse Section Search Co-Occurrence Word Search View Lists
Category Filter
Categories
Search
Go to section
Click Here

Word 1: Facebook
Word 2:
Word 3:

Context Length: 8
Sort Order: First left then right
Case sensitive: ☐
Include alternatives: ☐

5 Result(s)

Left Context	Match	Right Context
I ' m a member of MySpace and	Facebook	which are appealing to many people of my
go onto a new one. And I found	Facebook	really good and there ' s like you
appealing to many people of my age group.	Facebook	's normally for slightly older people and many
on the internet. I use it for like	Facebook	and like talking to my friends and email
go on MSN a lot and go on	Facebook	and kind of social websites like that . And

FIGURE 17.10 Word search

discourse as BNC, still can compensate for the important weaknesses of representative corpora in pedagogic contexts.

17.4 Conclusions and Future Work

SACODEYL Annotator has already enabled the SACODEYL team to accomplish DDL-oriented pedagogical annotation (Tornero et al. 2007). However, it is our intention to refine and improve the tool to make it as generic and flexible as possible.

The tool may contribute to building knowledge in many disciplines and provide textual resources with different kinds of annotated enrichment. In special, the tool could be helpful in CALL-related fields providing high-quality pedagogical materials stored in a standard format. Furthermore, these materials could be also reused by a wide amount of tools that support TEI.

SACODEYL is then the first major effort where pedagogic DDL has been implemented. By using TEI standardization, we hope to make this effort even more meaningful to the FLT and linguistic community. This environment can be viewed as a language learning platform which integrates multimodal search facilities, including section search and browse plus the more traditional concordance lines.

So far we have implemented P5 version of the TEI guidelines, which were released on 1 November 2007. Future work on SACODEYL Annotator is focused on the dissemination of the tool in connection with the Text Encoding Initiative tools and utilities such as XAIRA (Burnard 2004), *TAPoR*[4], *PhiloLogic*[5] *or Wordhoard*[6].

A wiki[7] has been established to attract the interest of fellow researchers in pedagogical annotation, and we expect to continue to develop SACODEYL Annotator into a more powerful device and system independent tool to store and process texts and corpora that can be used in the language classroom.

Notes

[1] System Aided Compilation and Open Distribution of European Youth Language research funded by the European Commission under the Socrates-Minerva initiative (225836-CP-1-2005-1-ES-MINERVA).

[2] The importance of standards in computer science lies beyond the scope of this article. Suffice it to say that if standard XML is used more and more users and applications will reuse the annotated resources.

[3] SACODEYL Site, http://www.um.es/sacodeyl, URL last accessed 15.07.2009)

[4] http://portal.tapor.ca/portal/portal, URL last accessed 15.07.2009)

[5] http://www.lib.uchicago.edu/efts/ARTFL/philologic/, URL last accessed 15.07.2009)

[6] http://wordhoard.northwestern.edu/userman/index.html, URL last accessed 15.07.2009)

[7] http://www.tei-c.org/wiki/index.php/Sacodeyl_Annotator, URL last accessed 15.07.2009)

References

Anastasia Scholarly Digital Editions. (2004), 'Anastasia: Analytical System Tools and SGML/XML Integration Applications'. Available through http://anastasia.sourceforge.net/whatis.html

Atserias, J., Casas, E., Comelles, M., González, L., Padró and Padró, M. (2006), 'FreeLing 1.3: Syntactic and semantic services in an open-source NLP library', in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06).* Genoa, Italy.

Bayer, S., Doran, C., Condon, S. and Gertner, A. (2006), 'Dialogue annotation as a correction task', in *9th International Conference on Intelligent User Interfaces,* 2006.

Braun, S. (2005), 'From pedagogically relevant corpora to authentic language learning contents', *ReCALL,* 17, (1), 47–64.

—(2006), 'ELISA – a pedagogically enriched corpus for language learning purposes', in Braun, S., Kohn, K. and Mukherjee, J. (eds), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods.* Frankfurt, a. M, Peter Lang, pp. 25–47.

—(2007), 'Integrating corpus work into secondary education: from data-driven learning to needs-driven corpora'. *ReCALL,* 19, (3), 307–328.

Burnard, L. (1995), 'The Text Encoding Initiative: An overview', in Leech, G., Myers, G. and Thomas, J. (eds), *Spoken English on Computer: Transcription, Markup and Applications.* Harlow: Longman.

Burnard, L. (2004), 'BNC-Baby and Xaira', in *Proceedings of the Sixth Teaching and Langauge Corpora conference,* Granada, p. 84.

Burnard, L. and Berglund, Y. (2007), 'Exploring BNC XML Edition with Xaira'. *28th Annual Conference of the International Computer Archive for Modern and Mediaeval English (ICAME).*

Cushion, S. (2004), 'Increasing accessibility by pooling digital resources'. *ReCALL,* 16, (1), 41–50.

Cutting D. and Pedersen, J. (1993), *The Xerox Part-of-Speech Tagger Version.* Via citeseer.ist.psu .edu/cutting93xerox.html

Garretson, G. (2006), 'Dexter: Free tools for analyzing texts'. *V International Congress of AELFE. Academic and Professional Communication in the 21st Century: Genres and Rhetoric in the Construction of Disciplinary Knowledge.*

Garside, R. (1987), 'The CLAWS word-tagging system', in R. Garside, F. Leech and G. Sampson (eds), *The Computational Analysis of English.* London: Longman.

Haugland, S. and Jones, F. (2002), *OpenOffice.Org 1.0 Resource Kit.* Prentice Hall PTR.

Jacobson, M. (2006), 'Le projet "Archivage" du LACITO'. *Langues et cité.* 6, 11

Lease, E. (2005), 'Creating and managing XML with open source software'. *Library Hi Tech Journal,* 23, (4), 526–540.

Leech, G. (1993), 'Corpus Annotation Schemes'. *Literary and Linguistic Computing,* 8, (4), 275–281.

Levy, M. (1997), *Computer-Assisted Language Learning: Context and Conceptualization.* Oxford: Oxford University Press.

McEnery, T. and Wilson, A. (1996), *Corpus linguistics.* Edinburgh: University of Edinburgh Press.

Mercader, A., Pérez-Paredes, P., Alcaraz, J. M. and Tornero, E. (forthcoming), 'The role of pedagogic annotation in DDL', in Bellés, B., Campoy-Cubillo M. C. and Gea-Valor, Ll. (eds) (forthcoming), *Exploring Corpus-Based Research in English Language Teaching.* Col·lecció Estudis Filologics.Publicacions de la Universitat Jaume I. Castelló.

Needleman, M. (2000), 'The Unicode Standard'. *Serial Review Journal,* 26, (2), 51–54.

Pérez-Paredes, P. and Alcaraz, J. M. (2007), 'Developing annotation solutions for online data-driven learning'. *ReCALL,* 21, (1), 55–75.

Pérez-Paredes, P., Alcaraz, J. M., Mercader, A. and Tornero, E. (2009, forthcoming), 'Extracting data from xml annotated corpora: Not so mysterious ways into data driven learning (DDL)', in Bellés, B., Campoy-Cubillo M. C. and Ll. Gea-Valor (eds) (forthcoming), *Exploring Corpus-Based Research in English Language Teaching.* Col·lecció Estudis Filologics.Publicacions de la Universitat Jaume I. Castelló.

Schmid, H. (1995), *TreeTagger – a Language Independent Part-of-Speech Tagger.* Institut fur Maschinelle Sprachverarbeitung, 1995.

SyncRO Soft Ltd. (2007), *oXygen XML Editor.* http://www.oxygenxml.com

Schmidt, T. (2004), Transcribing and annotating spoken language with EXMARaLDA, in *Proceedings of the LREC-Workshop on XML Based Richly Annotated Corpora,* Lisbon, 2004.

Tornero, E., Pérez-Paredes, P., Mercader, A. and Alcaraz, J. M. (2007), 'Annotating Spanish youngsters spoken language for DDL applications'. Paper presented at the *EUROCALL Conference.* University of Ulster at Coleraine, September 2007.

Ward, M. (2002), 'Reusable XML technologies and the development of language learning materials'. *ReCALL,* 14, (2), 285–294.

Chapter 18

Translation and Language Learning: AlfraCOVALT as a Tool for Raising Learners' Pragmatic Awareness of the Speech Act of Requesting[1]

Josep Roderic Guzmán Pitarch and Eva Alcón Soler
Universitat Jaume I, Castelló

18.1 Introduction

Applied linguistics has always been influenced by linguistic theories. In this sense, the shift from a concern with formal aspects of language (structural and generative linguistics) towards the study of language as communication created the conditions for adopting a pragmatic approach in linguistics and a large body of research on factors influencing learners' development of communicative competence. In the field of linguistics, Hymes (1972), Levinson (1983) and Leech (1983) encouraged a new emphasis away from Chomsky's notion of competence towards meaning in use, and different definitions of the term pragmatics were provided. Among them, Crystal's (1997) definition, and Leech's (1983) and Thomas's (1983) distinction between pragmalinguistic and sociopragmatic components within pragmatics may be relevant when we deal with pragmatics in language learning. On the one hand, Crystal (1997: 301) defines pragmatics as:

> The study of language from the point of view of users, especially of the choices they make, the constraints they encounter in using language in social interaction and the effects their use of language has on other participants in the act of communication.

Following Crystal (1997), pragmatics pays attention to meaningful interaction among users of language in particular sociocultural contexts. This perspective, also shared by LoCastro (2003), considers both speakers and hearers, or writers and readers, as users of language in context, whose actions may be directed towards conveying and interpreting communicative and interpersonal meanings. Their behaviour, in addition, is motivated by certain assumptions of universality concerning matters of linguistic politeness. Thus, research has shown that speech act categories and their realization strategies, such as indirectness, minimization and maximization of the pragmatic force, are found across languages. However, they do not apply in the same way to all languages. Leech (1983) and Thomas (1983) account for this fact by dividing pragmatics into two components: pragmalinguistics and sociopragmatics. The former refers to the resources for conveying communicative acts and interpersonal meanings, whereas the latter refers to the social perceptions underlying participants' interpretation and performance of communicative acts. Hence, when dealing with pragmatics we should consider knowledge of the means to weaken or strengthen the force of an utterance, that is pragmalinguistic knowledge, and knowledge of the particular means that are likely to be most successful for a given situation, i.e. sociopragmatic knowledge. Similarly, Goodwin and Duranti (1992) suggest that to understand meaning in interaction it is necessary to look beyond the event itself and consider situational factors because language in use does not only reflect context, but it shows how interlocutors negotiate which aspects of context are relevant for specific situations.

From this perspective, translation can not be understood as a linguistic procedure, but as an act of communicating across cultures. According to House (2008), translating always involves both languages and cultures because they are inextricably intertwined. Thus, translation could be defined as communication across cultures, which in turn involve using linguistic resources for conveying communicative acts and interpersonal meanings, while paying attention to the social perceptions underlying participants' interpretation and performance of communicative acts. The question is whether and how this definition of translation can be applied to language teaching. To answer this question in this chapter we will first provide a historical outline of how translation has been used in foreign language teaching. Secondly, we will raise the need to focus on pragmatics and review research dealing with learners' pragmatic awareness. Thirdly, we will illustrate how AlfraCOVALT is operated and make a number of suggestions as to how AlfraCOVALT might be used to improve learners' pragmatic awareness of the speech act of requesting.

18.2 Translation in Foreign Language Teaching

Translation has a long tradition in foreign language contexts. The basis of the grammar translation method consisted of translation from the foreign language and learning grammar rules and vocabulary through the translation of disconnected sentences. However, the direct method movement rejected the use of translation as a teaching technique and emphasized the importance of the spoken mode in foreign language teaching. Although criticism of the use of translation in foreign language teaching and learning continued at the beginning of the twentieth century, such criticism was emphasized with the advent of Audiolingual methodology, which was based on the assumption that oral communication is the main objective of language learning. The opposition of translation as a teaching technique was based on the belief that the mother tongue would avoid the learning of the target language. Finally, within the communicative approach the controversy about using translation in the language classroom is not settled. As far as the principles of communicative language teaching (CLT) are concerned, there seems to be a consensus on focusing on learners' development of communicative competence, as well as on the principle that communication is both an end and a means towards language learning. Concerning the former principle, speech act theory motivated the CLT content by designing functional-notional syllabi, which in turn influenced Hymes's (1972) notion of communicative competence. Hymes's original definition of communicative competence, which has been taken into account in several pedagogically communicative competence models (Canale and Swain 1980; Canale 1983; Bachman 1990; Celce-Murcia et al. 1995; Alcón 2000), have influenced the selection of the content of CLT, pragmatics being a key component. However, although pro-translation voices suggest using translation in CLT as a technique to raise awareness of contrasts between native and foreign language pragmatic competence, translation is often not related to the desired principles of CLT. In our opinion, the problem seems to be that in evaluating translation as a technique to increase learners' pragmatic competence, only pragmalinguistic is considered while sociopragmatic issues are neglected. The emphasis on pragmalinguistic issues results, as reported by House (2008), in failure to exploit the pedagogic usefulness of translation as a complex cross-linguistic activity. However, in line with House, we suggest that the strong pragmatic component in translation makes it potentially useful in raising learners' pragmatic awareness, an issue that has motivated current research in the field of interlanguage pragmatics.

18.3 Pragmatic Awareness and Language Learning

Analysing language use in context has provided language teachers and learners with a research-based understanding of the language forms and functions that are appropriate to the many contexts in which a language may be used. From this perspective, research in cross-cultural pragmatics has provided information on the interactive norms in different languages and cultures. Cross-cultural studies with a focus on speakers' pragmatic performance aim to determine whether the same speech act can be found in different cultures, and if so, to what extent it is performed. Likewise, explanations that account for those differences are provided. Among them, pragmatic transfer at the level of formal, semantic and speakers' perception of contextual factors seem to explain some of the differences between L1 and L2 speakers' use of the language. In addition, research from an interlanguage perspective takes into account acquisitional rather than contrastive issues, but in line with cross-cultural studies, it has focused on routines and pragmalinguistic realizations of different speech acts. A wide amount of studies now exist with a focus on request realizations (Hassall 1997; Li 2000; Rose 2000, among many others). Other speech acts that have received some attention on the part of scholars may be refusals (Félix-Brasdefer 2004), compliments (Rose and Ng 2001), and apologies (Trosborg 1995). We may also find exceptional studies in which sociopragmatic factors have been dealt with, but they usually refer to descriptions of situations presented to learners so that they acknowledge the most appropriate routine (Lorenzo-Dus 2001).

Although the sociopragmatic component has received less attention in interlanguage pragmatics, there is no doubt that sociopragmatics is relevant in L2 pragmatic development. On that account, Brown and Levinson's (1987) politeness variables – namely those of power, distance and ranking of imposition – and Scollon and Scollon's (1995) suggested politeness frameworks on the basis of face relationships have been used as a point of departure when dealing with pragmatics in foreign language learning and teaching. For instance, Scollon and Scollon's framework is considered in Safont's (2005) study devoted to examining the extent to which explicit instruction on learners' use of request formulae throughout one semester affected their use of peripheral modification devices. The training sessions consisted of description, explanation, discussion and practice on requests in context, and data were collected by means of a pre-test and post-test distributed before and after the instructional period. Results showed a positive effect of explicit instruction, since the use of the awareness-raising and production tasks employed in the study favoured learners' appropriate use of request peripheral modification devices after the treatment although, as claimed by the author, these elements had not been taught explicitly. Another example can be found in Martínez-Flor (2008). The author examined the effectiveness of an inductive-deductive teaching approach on learners' appropriate use of request modifiers in different situations that varied according to the three sociopragmatic factors described in Brown and Levinson's (1987) politeness theory, namely those of social distance, power and degree of imposition. Results from this study indicated that, after being engaged in the instructional period, learners (i) used a greater number of request modifiers; (ii) made use of a higher number of both internal and external modifiers and (iii) employed all different sub-types of internal and external modifiers, thus, including a wider variety of mitigating devices in learners' requestive behaviour.

In addition, pragmatic awareness seems to be particularly relevant in foreign language learning. Research on ILP has demonstrated that, in contrast to native speakers, who may not need to recognize speech act type consciously, foreign language learners' attention to pragmatic issues seems to be important due to the input difficulties found in foreign language contexts for pragmatic learning. Alcón and Safont (2008) illustrate how several investigations draw on Schmidt's (1993, 2001) noticing hypothesis to address awareness-raising as an approach to the teaching of pragmatics. These authors also point out that the studies conducted by Rose (2000) Grant and Starks (2001), Washburn (2001) and Alcón (2005) were motivated by the assumption that audiovisual input provides ample opportunities to address all aspects of language use in a variety of contexts. In addition, audiovisual input is reported to be useful to expose learners to the pragmatic aspects of the target language. Finally, the authors suggest that pragmatic judgement tasks based on audiovisual discourse analysis are useful to prepare learners for communication in new cultural settings.

From this perspective, corpora created and built with translations from audiovisual texts can be used to increase learners' pragmatic awareness. As stated by various scholars, learners' pragmatic awareness manifested in their ability to recognize and identify speech act types is limited. For instance, Kasper's (1984) investigation of the pragmatic comprehension of German-speaking English learners, suggested that failure to comprehend the illocutionary force of speech acts could be explained by learners' inability to produce those illocutionary devices in non-conventional indirect speech acts. In addition, the effect of language proficiency on learners' pragmatic awareness has been examined by Koike (1996), Cook and

Liddicoat (2002) and García (2004) pointing out learners' proficiency-related differences in the identification of speech acts. In our opinion, contextual knowledge and linguistic ability should be viewed as complementing variables that interact with each other in the comprehension of L2 culture. From this point of view, using translation in foreign language classrooms could be used as a first step to raise learners' sociopragmatic and pragmalinguistic awareness. As we will illustrate next, teaching particular pragmatic features, such as requests, can be achieved by means presenting learners with contextualized examples of requests in translation and using AlfraCOVALT together with and an inductive-deductive teaching approach.

18.4 Using AlfraCOVALT to Increase Learners' Pragmatic Awareness of the Speech Act of Requesting

The pragmatic feature selected to illustrate how to use AlfraCOVALT to increase learners' pragmatic awareness is the speech act of request. Trosborg (1995), Sifianou (1999), Márquez Reiter (2000) and Safont (2005) among others, have claimed that requests consist of two parts, (i) the core request or head act and (ii) the peripheral elements (see Safont, 2008, for a detail explanation of the speech act of request). On the one hand, the head act is the main utterance which has the function of requesting and can stand by itself. On the other hand, the peripheral elements are additional items which may follow and/or precede the request head act. They do not change the propositional content of the request head act but rather serve to either mitigate or aggravate its force. Since request modifiers accompany the request head act with the purpose of varying politeness levels and decreasing threatening conditions, they have notable importance when dealing with learning how to request. For the present study we followed Trosborg's (1995) typology of request realization strategies (Table 18.1) and the typology of peripheral request modification devices suggested by Alcón et al. (2005) and described in Table 18.2. An adaptation of Sifianou's taxonomy (1999) and the analysis of Spanish EFL learners' oral production data of request modification devices (Martínez-Flor and Usó, 2006) was taken into account in the design of the taxonomy provided in Table 18.2. Moreover, Brown and Levinson's (1987) sociopragmatic factors, summarized in Table 18.3, were also taken into account.

In addition, learners were trained to use AlfraCOVALT in their language classroom. AlfraCOVALT is a query programme that processes two corpora: the Auvi corpus and the COVALT corpus. Auvi is a corpus *ad hoc* created

Table 18.1 Trosborg's typology (1995)

REQUEST REALISATION STRATEGIES		
Indirect	Hints: *Statement*	*I have to be at the airport in half an hour*
Conventionally Indirect (hearer-based)	Ability: *Could you…?/ Can you…?*	*Can you lend me your car?*
	Willingness: *Would you…?*	*Would you lend me your car?*
	Permission: *May I…?*	*May I borrow your car?*
	Suggestory formulae: *How about…?*	*How about lending me your car?*
Conventionally Indirect (speaker-based)	Wishes: *I would like…*	*I would like to borrow your car*
	Desires/ needs: *I want/ need you to…*	*I want you to lend me your car*
	Obligation: *You must…/ You have to…*	*You must lend me your car*
Direct	Performatives: *I ask you to…*	*I ask you to lend me your car*
	Imperatives	*Lend me your car*
	Elliptical phase	*Your car*

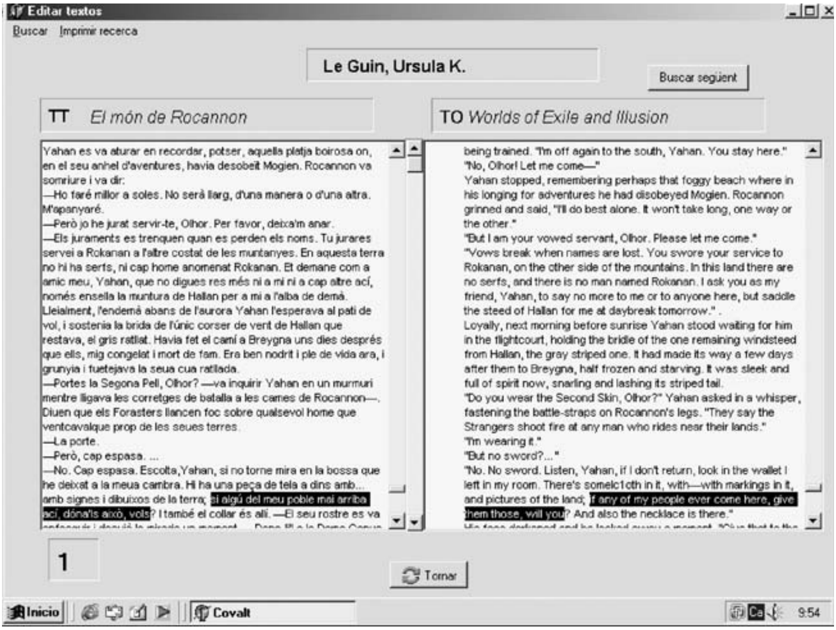
Table 18.2 Typology of peripheral modification devices in requests (Alcón et al. 2005)

TYPE	SUB-TYPE		EXAMPLE
Internal Modification	Openers		*Do you think* you could open the window? *Would you mind* opening the window?
	Softeners	Understatement	Could you open the window *for a moment*?
		Downtoner	Could you *possibly* open the window?
		Hedge	Could you *kind of* open the window?
	Intensifiers		You *really* must open the window *I'm sure* you wouldn't mind opening the window
	Fillers	Hesitators	I *er, erm, er* *I wonder* if you could open the window
		Cajolers	*You know, you see, I mean*
		Appealers	*OK?, Right?, yeah*
		Attention-getters	*Excuse me …; Hello …; Look …; Tom, …; Mr. Edwards …; father …*
External Modification	Preparators		*May I ask you a favour?* *Could you open the window*?
	Grounders		*It seems it is quite hot here.* Could you open the window?
	Disarmers		*I hate bothering you but* could you open the window?
	Expanders		Would you mind opening the window?… *Once again, could you open the window?*
	Promise of reward		Could you open the window? *If you open it, I promise to bring you to the cinema.*
	Please		Would you mind opening the window, **please?**

Table 18.3 Based on Brown and Levinson (1987)

FACTORS	POLITENESS EFFECT		
Social distance	Social distance increases	→	Politeness increases
Power	Power increases	→	Politeness increases
Imposition	Imposition is great	→	Politeness increases

and built with TV series like *Stargate SG-1* (with 94 episodes, of one hour), *The Berenstain Bears* (a cartoon serie of 40 episodes of half an hour each), and the movie *Two can play that game* (2001), directed by Mark Brown. The duration of all these programmes together is more than 116 hours. English is the language of the original texts, which are translated into Catalan for the Valencian TV. On the other hand the corpus COVALT (Guzman and Serrano, 2006) is built with full texts of narrative works and their translations into Catalan and Spanish published by Valencian press between 1990 and 2000 (204 works). As far as the interface is concerned, the AlfraCOVALT program has embedded a sentence alignment algorithm. It is a query programme that searches concordances between parallel texts using lexical information and certain heuristics. These searches can be carried out in the original text or in the translated text. Basically the programme works with internal and external sources. The external sources are three *Access* databases, each database with two fields, the first one with Catalan lemmatized words and the second one with the equivalence lemmatized word in English, French or German. All of them indexed and with duplicates. The German database has 54.581 entries, the English one 45.399 and the French one 73.474. The internal sources are *Paradox* databases with the texts' splitting sentences. The texts are split with a multilingual sentence boundary disambiguation algorithm using regular expressions and saved as registers in the *Paradox* databases. After typing the search string the alignment algorithm looks for the sentences where that string is embedded. The selected sentence (S1) will be tokenized, and each token lemmatized. A SQL query searches the Access Database (depending on the languages involved, the English, German or French Database) for the lemmas' translation, and then into the translated (or original) text looking for the sentences with the same words, of course with the morphological changes needed. So, in order to reduce time and economize resources, this search is done in a small window of the target text. This process is based on the idea that there is a relationship between the text length in terms of characters and the position of the searched string in both texts, original and translated. If the number of found words in a sentence of the target text (S2) is



Le Guin, Ursula K.

TT *El món de Rocannon*

Yahan es va aturar en recordar, potser, aquella platja boirosa on, en el seu anhel d'aventures, havia desobeït Mogien. Rocannon va somriure i va dir:
—Ho faré millor a soles. No serà llarg, d'una manera o d'una altra. M'apanyaré.
—Però jo he jurat servir-te, Olhor. Per favor, deixa'm anar.
—Els juraments es trenquen quan es perden els noms. Tu jurares servei a Rokanan a l'altre costat de les muntanyes. En aquesta terra no hi ha serts, ni cap home anomenat Rokanan. Et demane com a amic meu, Yahan, que no digues res més ni a mi ni a cap altre ací, només ensella la muntura de Hallan per a mi a l'alba de demà.
Lleialment, l'endemà abans de l'aurora Yahan l'esperava al pati de vol, i sostenia la brida de l'únic corser de vent de Hallan que restava, el gris ratllat. Havia fet el camí a Breygna uns dies després que ells, mig congelat i mort de fam. Era ben nodrit i ple de vida ara, i grunyia i fuetejava la seua cua ratllada.
—Portes la Segona Pell, Olhor? —va inquirir Yahan en un murmuri mentre lligava les corretges de batalla a les cames de Rocannon—. Diuen que els Forasters llancen foc sobre qualsevol home que ventcavalque prop de les seues terres.
—La porte.
—Però, cap espasa. ...
—No. Cap espasa. Escolta,Yahan, si no torne mira en la bossa que he deixat a la meua cambra. Hi ha una peça de tela a dins amb... amb signes i dibuixos de la terra; si algú del meu poble mai arriba ací, dóna'ls això, vols? I també el collar és allí. —El seu rostre es va

TO *Worlds of Exile and Illusion*

being trained. "I'm off again to the south, Yahan. You stay here."
"No, Olhor! Let me come—"
Yahan stopped, remembering perhaps that foggy beach where in his longing for adventures he had disobeyed Mogien. Rocannon grinned and said, "I'll do best alone. It won't take long, one way or the other."
"But I am your vowed servant, Olhor. Please let me come."
"Vows break when names are lost. You swore your service to Rokanan, on the other side of the mountains. In this land there are no serfs, and there is no man named Rokanan. I ask you as my friend, Yahan, to say no more to me or to anyone here, but saddle the steed of Hallan for me at daybreak tomorrow." .
Loyally, next morning before sunrise Yahan stood waiting for him in the flightcourt, holding the bridle of the one remaining windsteed from Hallan, the gray striped one. It had made its way a few days after them to Breygna, half frozen and starving. It was sleek and full of spirit now, snarling and lashing its striped tail.
"Do you wear the Second Skin, Olhor?" Yahan asked in a whisper, fastening the battle-straps on Rocannon's legs. "They say the Strangers shoot fire at any man who rides near their lands."
"I'm wearing it."
"But no sword?..."
"No. No sword. Listen, Yahan, if I don't return, look in the wallet I left in my room. There's somelc1cth in it, with—with markings in it, and pictures of the land; if any of my people ever come here, give them those, will you? And also the necklace is there."

FIGURE 18.1 Concordances example

bigger than five and its percentage is 20 per cent bigger than the words in S2, then we can say that S1 and S2 are equivalent. If any of these conditions are negative, then the query continues with the sentences before and after the analysed sentence. At the end, the programme returns the concordances between both texts (OT and TT) as shown in Figure 18.1.

The following procedure was used with the aim of raising learners' awareness of requests:

1 Searching for requests in the original version
2 A comparison of the original version with the one provided by means of AlfraCovalt by focusing on the following pragmalinguistic question: How many forms of requests modifiers did you find in the original version? Are they translated literally? If not, write down the equivalent.

> '**Would you mind** telling me, Agatha, what it was that you dreamed about me?' (A. C. Doyle, *The Parasite*)
>
> '[. . .] **et faria res** contar-me què has somiat de mi?'
>
> 'BROTHER (delicately) **Would you mind** if I took that book? I left it here by mistake.' (*The Berenstain Bears* 'Think of those in need' EPISODE 29A)
>
> GERMÀ (DE) **Li importa que** m'emporte este llibre? Me l'he deixat ací per error.

3 A comparison of the OT and TT to check if there is any difference in the quantity and type of request modifiers. If so, why?

(a) 'My dear sir,' said Mr. Otis, 'I really must insist on your oiling those chains, [. . . .]' (O. Wilde, *Lord Arthur Savile's Crime*)

Estimat senyor meu- digué el senyor Otis-, francament he d'insistir que greixe les cadenes

(b) SQUIRE (partly OS) 'I'm going to need your skills. You see, I need a gift for my wife's birthday. I'd like you to build her a very special chair.' (*The Berenstain Bears,* 'The hiccup cure' EPISODE 29B)

MONOCLE Necessitaré de les teues habilitats. Mira, voldria un regal per a l'aniversari de la meua dona. (ON/OFF) M'agradaria que li construïres una cadira molt especial.

(c) QUILTER #1 (to Brother) Thank you for the lemonade, Dear. (then to Sister) You know, my eyes aren't what they used to be. Do you think you could thread my needle for me? (*The Berenstain Bears,* 'Trouble with money' EPISODE 6A)

TEIXIDORA Gràcies per la llimonada, bonico./Filla, amb l'edat he perdut molta vista.(OFF) Creus que podries enfilar-me l'agulla?

(d) O'NEILL: 'Oh, stop it, will you?' (STARGATE SG-1 'Abyss' EPISODE #P653)

O'NEILL: Ai, deixa-ho ja, per favor.

(e) BROTHER Huh. You'll never let me forget that, will you? (*The Berenstain Bears* 'The talent show' EPISODE 9A)

GERMA: (G) No se t'oblidarà mai, veritat?

(f) 'Shut the door so that it don't fly open, will you? I can't stand a door banging. They've put a lot of rubbishy locks into . . .' (J. Conrad, *Typhoon*)

Tanque la porta de manera que no s'óbriga, vol?

(g) 'Just hand over that sapphire cross of yours, will you?' (G. K. Chesterton, *The Secret Garden*)

Done'm ara mateix la seua creu de safirs, entesos?

4 Teachers' explanation of the typologies in Table 18.1 and 18.2 with presentation of request head acts and the internal and external modification devices accompanying them in OT and TT are provided.

5 Teachers' explanation on the effect of sociopragmatic factors on politeness is provided.

6 An analysis of requests in OT and the TT to examine whether the linguistic realizations of the speech act of requesting is influenced

by sociopragmatic factors such as degree of familiarity, interlocutors' power or size of the request.

18.5 Conclusions

The above-mentioned procedure enables teachers to guide learners' attention to different linguistic formulae requests that are given and received in different languages, and how different realization strategies are used, taking into account social factors such as interlocutors' power, familiarity or status. These observation tasks based on translations may help students make connections between linguistic forms, pragmatic functions, their occurrence in different social contexts and their cultural meanings. In other words, students are guided to notice the information they need in order to develop their pragmatic competence in L2. Thus, we can claim that translation offers foreign language learners the opportunity to reflect on different pragmatic options in a communicative event. In addition, by encouraging students to explore and reflect on their experiences, observations and interpretations of translations as communication across cultures we might gain a better understanding of the meaning in the original text.

Notes

[1] Research for this article was funded by a grant from the Ministerio de Educación y Ciència and FEDER for the project HUM2006-11524, and HUM2004-04435/FILO, and Bancaixa P1•1B2006-13.

References

Alcón, E. (2000), 'Desarrollo de la competencia discursiva oral en el aula de lenguas extranjeras: perspectivas metodológicas y de investigación', in Muñoz, C. (ed.), *Segundas lenguas: Adquisición en el aula.* Barcelona: Ariel, pp. 259–272.

—(2005), 'Does instruction work for learning pragmatics in the EFL context?'. *System,* 3, 417–435.

Alcón, E. and Safont, M. P. (2008), 'Pragmatic awareness in second language acquisition', in Cenoz, J. and Hornberger, N. (eds), *Encyclopedia of Language and Education. V 6. Knowledge about Language* (second edition). New York: Springer, pp. 193–204.

Bachman, L. F. (1990), *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Brown, P. and Levinson, S. C. (1987), *Politeness. Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Canale, M. (1983), 'From communicative competence to communicative language pedagogy', in Richards, J. C. and Schmidt, R. W. (eds), *Language and Communication*. London: Longman, pp. 2–27.

Canale, M. and Swain, M. (1980), 'Theoretical bases of communicative approaches to second language teaching and testing'. *Applied Linguistics*, 1, 1–47.

Celce-Murcia, M., Dörnyei, Z. and Thurrell, S. (1995), 'Communicative competence: A pedagogically motivated model with content specifications'. *Issues in Applied Linguistics*, 6, 5–35.

Cook, M. and Liddicoat, A. J. (2002), 'The development of comprehension in interlanguage pragmatics: The case of request strategies in English'. *Australian Review of Applied Linguistics*, 25, 19–39.

Crystal, D. (1997) (ed.), *The Cambridge Encyclopaedia of Language* (second edition). New York: Cambridge University Press.

Félix-Brasdefer, J. J. (2004), 'Interlanguage refusals: Linguistic politeness and length of residence in the target community'. *Language Learning*, 54, (4), 587–653.

García, P. (2004), 'Pragmatic comprehension of high and low level language learners'. *Teaching English as a Second or Foreign Language*, 8, 1–10.

Goodwin, C. and Duranti, A. (1992), 'Rethinking context: An introduction', in Duranti, A. and Goodwin, C. (eds), *Rethinking Context. Language as an Interactive Phenomenon*. Cambridge: Cambridge University Press, pp. 1–42.

Grant, L. and Starks, D. (2001), 'Screening appropriate teaching materials. Closings from textbooks and television soap operas'. *International Review of Applied Linguistics*, 39, 39–50.

Guzman Pitarch, J. R. and Serrano, A. (2006), 'Alineamiento de frases y traducción: AlfraCOVALT y el procesamiento de corpus'. *Sendebar*, 17, 169–186.

Hassall, T. J. (1997), 'Requests by Australian learners of Indonesian'. Unpublished doctoral dissertation. Canberra: Australian National University

House, J. (2008), 'Using translation and interpreting to increase L2 pragmatic competence', in Alcón, E. and Martínez-Flor, A. (eds), *Investigating Pragmatics in Foreign Language Learning, Teaching and Testing*. Clevedon: Multilingual Matters, pp. 135–152.

Hymes, D. (1972), 'On communicative competence', in Pride, J. and Holmes, J. (eds), *Sociolinguistics: Selected Readings*. Harmondsworth: Penguin, pp. 269–293.

Kasper, G. (1984), 'Pragmatic comprehension in learner-native speaker discourse'. *Language Learning*, 34, 1–20.

Koike, D. A. (1996), 'Transfer of pragmatic competence and suggestions in Spanish foreign language learning', in Gass, S. M. and Neu, J. (eds), *Speech Acts Across Cultures: Challenges to Communication in a Second Language*. New York: Mouton de Gruyter, pp. 257–281.

Leech, G. (1983), *The Principles of Pragmatics*. London: Longman

Levinson, S. (1983), *Pragmatics*. Cambridge: Cambridge University Press.

Li, D. (2000), 'The pragmatics of making requests in the L2 workplace: A case study of language socialization'. *Canadian Modern Language Review*, 57, (1), 58–87.

LoCastro, V. (2003), *An Introduction to Pragmatics: Social Action for Language Teachers.* Michigan: Michigan Press.

Lorenzo-Dus, N. (2001), 'Compliment responses among British and Spanish university students: A contrastive study'. *Journal of Pragmatics*, 33, 107–127.

Martínez-Flor, A. (2008), 'The effect of an inductive-deductive teaching approach to develop learners' use of request modifiers in the EFL classroom', in Alcón, E. (ed.), *Learning How to Request in an Instructional Language Learning Context.* Bern: Peter Lang, pp. 191–225.

Martínez-Flor, A. and Usó-Juan, E. (2006), '"Learners" use of request modifiers across two University ESP disciplines'. *Ibérica*, 12, 23–41.

Rose, K. R. (2000), 'An exploratory cross-sectional study of interlanguage pragmatic development'. *Studies in Second Language Acquisition*, 22, 27–67.

Rose, K. R. and Ng, C. K. (2001), 'Inductive and deductive teaching of compliments and compliment responses', in K. R. Rose and G. Kasper (eds), *Pragmatics in Language Teaching.* Cambridge: Cambridge University Press, pp. 145–170.

Safont, M. P. (2005), *Third Language Learners. Pragmatic Production and Awareness.* Clevedon: Multilingual Matters.

—(2008), 'The speech act of requesting', in Alcón, E. (ed.), *Learning How to Request in an Instructional Language Learning Context.* Bern: Peter Lang, pp. 41–64.

Schmidt, R. (1993), 'Consciousness, learning and Interlanguage pragmatics', in Kasper, G. and Blum-Kulka, S. (eds), *Interlanguage Pragmatics.* New York: Oxford University Press, pp. 21–42.

—(2001), 'Attention', in Robinson, P. (ed.), *Cognition and Second Language Instruction.* New York: Cambridge University Press, pp. 3–33.

Scollon, R. and Scollon, S. W. (1995), *Intercultural Communication. A Discourse Approach.* Oxford: Blackwell.

Sifianou, M. (1999), *Politeness Phenomena in England and Greece. A Cross-Cultural Perspective.* Oxford: Oxford University Press.

Thomas, J. (1983), 'Cross-cultural pragmatic failure'. *Applied Linguistics*, 4, 91–112.

Trosborg, A. (1995), *Interlanguage Pragmatics: Requests, Complaints, and Apologies.* Berlin and New York: Mouton de Gruyter.

Washburn, G. N. (2001), 'Using situation comedies for pragmatic language teaching and learning'. *TESOL Journal*, 10, (4), 21–26.

Chapter 19

The Videocorpus as a Multimodal Tool for Teaching

Inmaculada Fortanet-Gómez and Mercedes Querol-Julián[1]
Universitat Jaume I, Castelló

19.1 Introduction

As Römer has extensively commented in Chapter 2 of this volume, the use of corpora for research started a long time ago mainly in order to make dictionaries. However, it was with the creation and dissemination of the computers that the large corpora began to be compiled and research using these corpora started. Some years later more specific corpora were needed which compiled the discourse of certain events or contexts to collect examples of spoken genres at American and British universities. In order to assist in the analysis of corpora, software programmes were developed to search concordances and collocations, and to tag the language according to their syntactic, semantic or morphological features.

Corpus linguistics can be used for syllabus design, materials development and classroom activities (Krieger 2003), as prove the experiences that have been reported in several publications. However, from our point of view, when trying to teach spoken discourse using a corpus-based learning, transcripts do not provide the real situation when and where language is used. There is a lack of general context and background, we do not know who says what, how, when and where. Language is accompanied by prosodic features such as intonation, accent or stress, among others; and kinesics such as gestures and body language, which are not present and cannot be discerned from a transcript, no matter how many labels are added to the text.

On the other hand, in order to provide this additional information, video recordings that try to exemplify situations accompany some textbooks of English. By experience, we know that these are not real situations, since they are performed by actors trying to produce a perfect discourse where characteristics of speech such as blending, hesitations, false starts or overlapping never occur.

The proposal we present in this chapter is the creation of a video corpus including recordings of real situations, in which the English language is used. The example we provide has been recorded and edited for its application to a course of teacher training for lecturing in English at Universitat Jaume I. Lectures in English were recorded for this purpose and excerpts were tagged to exemplify the different functions of these classroom events. This is only an example of the type of tagging or classification of speech events that can be done in video corpora, many other possibilities can be explored which can assist in the teaching of pragmatics, grammar, vocabulary, etc.

19.2 The Use of Corpora

The analysis of large amounts of texts started a long time ago with the aim of finding words and meanings for the creation of glossaries and dictionaries. However, it has been in recent times, since the late 1980s, but mostly during the 1990s, when most computerized large corpora have been compiled. According to Stubbs (1996: 231):

> Within a very short period of time, linguists have acquired new techniques of observation. The situation is similar to the period immediately following the invention of the microscope and the telescope, which suddenly allowed scientists to observe things that had never been seen before.

At the end of the 1990s, it was still difficult for a large amount of researchers to have access to these corpora, but in the latest years the popularization of the internet, in addition to the dissemination of software, has allowed the use of corpora, not only for research all over the world, but also for teaching.

Corpora can be used for syllabus design, materials development and classroom activities (Krieger 2003). One of the most well-known products of corpus linguistics is Biber et al.'s *Longman Grammar of Spoken and Written English* (1999) and the subsequent development of the student's book and workbook. The novelty about this grammar is that it is based on the analysis of a corpus of over 40 million words, including conversation, fiction, news and academic prose. For the first time, grammar rules are established from samples of authentic language rather than by the intuition of a native speaker linguist.

Although the beginning of corpus linguistics was marked by the exclusive analysis of written texts, eventually spoken discourse also drew the attention of researchers. Transcriptions of speech formed corpora such as COLT

(Bergen Corpus of London Teenage Language 2000), Corpus of Professional American English (Barlow 2000), MICASE (Michigan Corpus of Academic Spoken English, Simpson et al. 2002) or BASE (British Academic Spoken English 2007). The procedure for the compilation of these corpora consists in audio or video recording native or non-native speakers when speaking in certain situations. As a second step, these recordings are transcribed, including a certain annotation of circumstances that may affect the interpretation of the transcription, such as pauses, overlappings, noises or sounds, etc. It is these transcriptions that researchers and teachers have used as corpora for their research or tuition. However, transcriptions cannot show the reality of spoken discourse since all multimodal elements are ignored, reducing so the richness of the spoken discourse to the restrictive meaning of the written mode. We believe that multimodal characteristics of spoken discourse should not be obviated, and a complete analysis could not be considered as finished without attending to non-verbal elements of communication. There has been a tradition in semiotics and sociolinguistics in the study of NVC (Non-verbal Communication); however, only a few studies can be found in recent research to analyze these elements from the point of view of linguistics and together with the language (Räisänen and Fortanet 2006, Crawford-Camiciottoli 2007). Poyatos (2004) describes human communication as consisting of three basic types of elements: verbal, kinesic (body movement, gestures, face expression) and paralinguistic (accent, intonation, pauses, stress, rhythm). Apart from these, spoken discourse is produced in a certain context, and a visual image of that context can provide also additional information about the space, the time, the function or role of the speaker and hearer, or the relation between the participants in the speech event. Furthermore, speakers commonly use supportive materials in the form of computer presentations (e.g. ppt slides), white or blackboard, handouts, etc. All this information is not present when analysing transcriptions, no matter how detailed it may be. The importance of showing video recordings in the language classroom has been observed by several researchers from the early 1990s:

> Non-verbal behaviours are among the hardest to make learners aware of, yet we know their significance for communication, especially cross-culturally. (Candlin 1990: vii)

For this reason video recordings were created to accompany many English Language Teaching manuals, especially in the field of business (Comfort and Utley 1995; Jones and Richard 1996). Stempleski and Tomalin (1990: 3) encourage teachers to use video in the language classroom, since it improves students' motivation, makes students more ready to communicate in the target language, shows the non-verbal aspects that accompany language and gives opportunities to observe differences in cultural behaviour.

Even though we acknowledge the importance of video recordings, most of the materials that have been traditionally used in language courses have been recorded on purpose with this aim using previously written scripts and actors who say their words learnt by heart in the clearest possible way to help students' understanding. However, most language learners feel frustrated the first time they try to hear an authentic conversation, with background noises, interrupted sentences, ellipsis, strong accents, unusual body language, etc. A way to prevent this frustration is providing students with recordings of authentic language, not in transcripts but all of it. However, something that was very convenient with the artificial recordings accompanying teaching materials was that they matched what was being learnt at the moment. How could video recordings be prepared to obtain this advantage? During recent years written corpora as well as some transcripts of spoken discourse have been tagged, i.e. using electronic means a series of labels are added to the text to identify parts of the text, linguistic elements, speakers and their characteristics, etc. (Campoy 2002: 123).

Although multimodal resources and multimodal systems of annotation are developing in different fields, mainly in psychology or cross-linguistics studies of multimodal communication; they have seldom been used from a linguistic perspective as language resources per se. Apart from Baldry and Taylor (2004), and Martin et al. (2002) we have not found any references to the recording and tagging of lectures for being used as pedagogical material. Baldry and Taylor (2004) created an online multimodal concordancer the Multimodal Corpus Authoring System (MCA) for rich film analysis, subtitling and dubbing. Designed originally as a support for translators' research to examine and compare multiple contexts and texts, it is also used as a resource for distance language learning. Martin et al. (2002) worked on the annotation of a corpus of video-taped lectures and student working sessions to improve the existing online tutorial with multimodal and adaptive hypermedia features. However, these studies focus exclusively on the use of the annotated material in distance language learning. This chapter proposes the use of annotated real lecture recordings in face-to-face English for Academic Purposes (EAP) classes.

The group of research GRAPE (Group of Research on Academic and Professional English), at Universitat Jaume I, is currently working in a project to compile and design a multimodal corpus of academic events in

English that take place at the university. The corpus does not only include video recordings, but also files used as computer presentations displayed during the sessions, handouts and other materials related to the event, as well as full transcriptions of the recordings. The Multimodal Academic and Spoken Language Corpus (MASC) has a dual-purpose, to do research and to design materials to teach English for Academic Purposes (EAP). The aim of this chapter is to illustrate how the video recording of a lecture given in an e-business postgraduate course at Universitat Jaume I can be edited and tagged with standard editing software to be used in the EAP classes.

19.3 Method

The method we followed to create the teaching material is divided into two stages: recording, and edition.

19.3.1 Recording

The sessions were recorded using a Mini-DV Digital Video camera with an external wireless unidirectional microphone. We used 90 minutes LP tapes which allowed full recordings of the lectures without unnecessary cuts. It is important to record some seconds before the speaker starts, and to do the same after she/he finishes the session. This simple measure guarantees complete recordings. In addition, other aspects should be considered before video-taping, in order to prevent problems which may affect the quality of the film, especially those related to environmental factors, and speakers performance.

Regarding the environmental factors, the room size, and distribution of tables, blackboard, aisles, and other elements in the room should be observed before setting up the camera. We should cause little trouble to the speaker, so that she/he does not feel threatened by the camera, keeps an eye on it the whole time, or behaves unnaturally; the smaller the room the more difficult it will be to create a comfortable environment keeping the speaker in focus. On the other hand, the camera should neither prevent the audience from seeing the speaker, nor call their attention and distract them from the lecture. Light conditions are also important. Excessively dark rooms or light contrasts when using projectors on a screen may cause poor recordings. Furthermore, depending on the light conditions, writing or drawing on white or blackboards may not be visible. Finally, in some rooms which are designed for holding conferences or meetings, audio points are available. In these cases it is recommended to connect the camera to these points to avoid background noise. In the other situations, we can plug an external microphone into the camera.

As for the speakers' performance, they may be sitting during the whole event, stand up, or stand up and move (along the front row, or up and down the room). The three possibilities should be considered when setting up the camera to be able to have the speaker in focus all the time.

We should check in advance all these aspects to avoid poor recordings with the speaker out of focus; non-visible slides, whiteboard or blackboard; or audio problems.

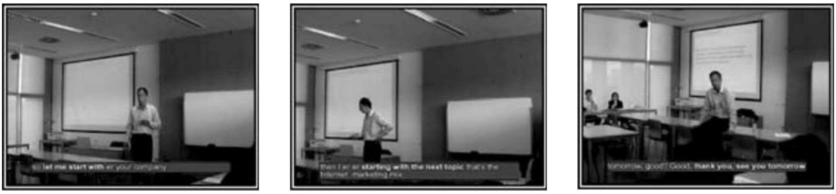
19.3.2 Edition

The edition stage is divided into three parts: tagging, subtitling, and creation of an interface. We use the video editing software Avid Liquid 7.0 to create and edit the DVDs converting the recordings into .avi files. One of the advantages of this software is it works with multiple tracks of audio, video or the combination of both. This feature opens a range of possibilities to edit the film for research and pedagogical purposes. In this chapter we make a proposal for a pedagogical use of the recording of a business lecture in EAP classes.

The first step in the edition is tagging. Tagging has been widely used for transcripts and written corpora, but not for video recordings up to now. There are many aspects that can be considered in an annotated corpus. We adopted Fortanet-Gómez et al.'s (2008) functional approach to tag the lecture. This work describes eight functions fulfilled by the teachers in university lectures: to start the lecture, to define concepts, to introduce a classification, to give examples, to explain a process, to set up objectives, to compare and contrast and to end the lecture. We selected one excerpt from the clip (90-minute lecture recording) to exemplify each function. Then the excerpts were spliced in a new clip of about 5 minutes. Each excerpt was tagged with the name of the function it represents.

The second stage is the subtitling. We did it with a tool available in the program to create titles. We only included in the subtitle the fragments that contained the most characteristic expressions and discourse markers used in each function. For the selection of these expressions and discourse markers we used as a source Bellés-Fortuño (2007) and Fortanet et al. (2008) (Figure 19.1).

Finally, we created an interface to access the multimodal clip in a practical way to be used in class. Thus, we designed a DVD menu with eight


so let me start with er your company
then I er er starting with the next topic that's the Internet marketing mix
tomorrow good? Good, thank you, see you tomorrow

To start a lecture To set up objectives To end the lecture

FIGURE 19.1 Video shots of tagged functions (pictures taken from MASC)

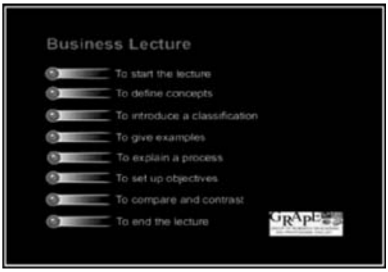


FIGURE 19.2 Video shot of the DVD menu

entries, one for each function. When we click on an entry the example of the function that we want to show is screened (see Figure 19.2).

The clip is exported as a .vob file type to avoid incompatibility problems and to keep the quality of the original film. The files have an extension of 186 MB, which allows making a copy in a DVD, but also in a portable hard disk.

19.4 Pedagogical Application

In recent years there has been a growing demand in Spanish universities for training courses in advanced English for faculty. The internationalization of higher education encourages and even obliges university faculty members to teach in other languages, mainly in English. At Universitat Jaume I these EAP courses are taught by lecturers from the Department of English Studies. One of the main difficulties of these courses is the lack of specific materials. Students in these courses are rather demanding and require authentic materials where they can observe the behaviour and language of other native, or non-native lecturers with a high proficiency in the English language, and in a similar situation to that they are going to find when teaching in English.

Some of the sessions that created more interest among the students were those related to the language used for the functions carried out in a lecture, such as starting the lecture, defining concepts or introducing examples. A deep analysis of the material we had recorded, as well as some research carried out by members of the research group (Bellés-Fortuño 2007, Querol-Julián 2007) provided us with the key language. However, there was a need for contextualized examples and video excerpts had to be carefully searched.

Searching for the most appropriate examples proved to be an exhausting task that would have to be repeated for every course, unless the video recordings were tagged and a search tool could be applied. The first stage, the tagging of the videocorpus has already been started and will continue in the next months with the creation of new software to search the recorded and tagged materials, completing therefore the second stage of the experience.

19.5 Conclusion

This chapter tries to prove how simple it can be to use a standard editing software to create teaching material for the class. In this way EAP students are provided with full examples of natural language (in opposition to the artificial scripts performed by actors in traditional materials) used in some of the functions accomplished by the teacher in lectures. With these examples students do not only listen to the most frequent expressions and discourse markers employed by an authentic teacher; but also watch how he speaks, behaves and moves (prosodic features and kinesics); how he uses hesitations, false starts, pauses, ellipsis; and how he interacts with the classroom elements.

However, we are currently working on a more complex task, to design a multimodal concordancer. The MASC (Multimodal Academic and Spoken Language Corpus) is presently constituted by three elements: (a) video recordings of English academic events from different disciplines (not only lectures, but also guest lectures, paper presentations, plenary speakers presentations, seminars, dissertation defences and students presentations), (b) full transcriptions of the events (some annotations have been already added such as identification of the speakers, pauses, overlaps, laughter, contextual events, reading passages, uncertain or unintelligible speech) and (c) supportive materials used by the speaker (e.g. slides, computer presentations or handouts). Nevertheless, the first step in the concordancer design is to properly annotate the corpus according to our needs.

Two big groups of annotations are to be made. The first group will include general features: (a) type of event (lecture, paper presentation, seminar, etc), (b) academic discipline (Business, Biology, Chemistry, etc) and (c) speaker profile (sex, genre, status, etc). The second group will cover discursive features: (d) speaker performance (sitting, standing up, standing up and moving), (e) linguistic functions (those used in this chapter and others more commonly employed in other types of events), (f) prosodic features (intonation, accent or stress), (g) kinesics (hand gestures, gaze, posture, facial expressions) and (h) use of supportive materials.

The multimodal concordancer will design queries to go into these two groups of annotations. Hence, the result of the query will be a multimodal outcome: audio, video, graphics, visuals and written text. Our intention is that when we search for instance how *to start a lecture* the concordancer looks for this function in all the lectures in the corpus and retrieves that particular excerpt from all of them with the four elements that will constitute the corpus: video recording, transcription, annotations and supportive materials.

Furthermore, these tagged corpora can be useful for the teaching of the English language from the point of view of pragmatics and intercultural communication, since language is usually accompanied by multimodal elements which often provide the most important clues for communication.

Notes

[1] The contribution of this author was supported by Universitat Jaume I under Grant Number PREDOC/2005/23.

References

Baldry, A. and Taylor, Ch. (2004), 'Multimodal concordancing and subtitles with MCA', in Partington, A., Morley, J. and Haarman, L. (eds), *Corpora and Discourse.* Bern: Peter Lang, pp. 57–70.

Barlow, M. (2000), *Corpus of Spoken Professional American English* (CD-Rom). Houston, TX: Athelstan.

BASE. *British Academic Spoken English (BASE) Corpus.* (2007), The corpus was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Retrieved 20 April 2007, from: http://www.warwick.ac.uk/go/base.

Bellés-Fortuño, B. (2007), *Discourse Markers within the University Lecture Genre: A Contrastive Study between Spanish and North-American lectures.* Ph.D. thesis. Castelló: Pubicacions de la Universitat Jaume I.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999), *Longman Grammar of Spoken and Written English.* London: Longman.

Campoy, M. C. (2002), 'Spoken Corpora and their Pedagogical Applications', in P. Safont and M. C. Campoy (eds), *Oral Skills: Resources and Proposals for the Classroom.* Castelló: Publicacions de la Universitat Jaume I, pp. 117–134.

Candlin, Ch. N. (1990), General Editor's Preface to S. Stempleski and B. Tomalin, *Video in Action. Recipes for Using Video in Language Teaching.* Hemel Hempstead: Prentice Hall, pp. vii–viii.

COLT. *The Bergen Corpus of London Teenage Language.* University of Bergen, Department of English (2000), Retrieved 20 April 2007, from: http://www.hd.uib.no/colt/

Comfort, J. and Utley, D. (1995), *Oxford Business English Skills Series.* Oxford: Oxford University Press.

Crawford-Camiciottoli, B. (2007), *The Language of Business Studies Lectures.* Amsterdam: John Benjamins.

Fortanet Gómez, I. (coord.), Bellés Fortuño, B., Giménez Moreno, R., Palmer Silveira, J. C., Ruiz Garrido, M. (2008), *Hablar inglés en la universidad: docencia e investigación.* Oviedo: Septem Ediciones.

Jones, L. and Richard, A. (1996), *New International Business English. Communication Skills in English for Business Purposes.* Cambridge: Cambridge University Press.

Krieger, D. (2003), 'Corpus Linguistics: What It is and How It Can Be Applied to Teaching'. *The Internet TESL Journal,* 9, (3). Retrieved 20 April 2007, from: http://iteslj.org/

Martin, J. Cl., Réty, J. H. and Bensimon, N. (2002), 'Multimodal and adaptive pedagogical resources'. *3rd International Conference on Language Resources and Evaluation* (LREC 2002), Las Palmas, Canary Islands, Spain, 29–31 May 2002. Retrieved 20 April 2007, from: http://www.lrec-conf.org/lrec2002/index.htlm

Poyatos, F. (2004), *Nonverbal Communication across Disciplines. Volume I: Culture, Sensory Interaction, Speech, Conversation.* Amsterdam: John Benjamins.

Querol-Julián, M. (2007), 'Narratives in English academic lectures'. Unpublished Master thesis. Castelló: Universitat Jaume I.

Räisänen, Ch. and Fortanet, I. (2006), 'Do genres have body language? Nonverbal communication in conference paper presentations'. Paper presented at the *Conference Homage to John Swales.* Ann Arbor, MI, June 2006.

Simpson, R. C., Briggs, S. L., Ovens, J. and Swales, J. M. (2002), *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan. Retrieved 20 April 2007, from: http://www.lsa.umich.edu/eli/micase/micase.htm

Stempleski, S. and Tomalin, B. (1990), *Video in Action. Recipes for Using Video in Language Teaching.* Hemel Hempstead: Prentice Hall.

Stubbs, M. (1996), *Text and Corpus Linguistics: Computer-Assisted Studies of Language and Culture.* Cambridge, MA: Blackwell Publishers.

Index

NOTE: Page references in *italics* denote figures and tables.