Demystifying Corpus Linguistics for English Language Teaching

Edited by
Kieran Harrington · Patricia Ronan

palgrave
macmillan

Editors
Kieran Harrington and Patricia Ronan

Demystifying Corpus Linguistics for English Language Teaching

palgrave
macmillan

Editors
Kieran Harrington
Faculty of Cultural Studies, TU Dortmund University, Dortmund, Germany

Patricia Ronan
Faculty of Cultural Studies, TU Dortmund University, Dortmund, Germany

ISBN 978-3-031-11219-5 e-ISBN 978-3-031-11220-1
https://doi.org/10.1007/978-3-031-11220-1

© The Editor(s) (if applicable) and The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the

material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cover illustration: © Alex Linch / shutterstock.com

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This book is dedicated to the inspirational Günter Nold, Emeritus Professor of Applied Linguistics and ESL/EF Education at TU Dortmund University

Contents

1 Demystifying Corpus Linguistics for English Language Teaching
Kieran Harrington and Patricia Ronan

2 Learning to Teach English as a Foreign Language with Corpus Linguistic Approaches: A Survey of Teacher Training Students' Attitudes
Patricia Ronan

3 A Flexible Framework for Integrating Data-Driven Learning
Jane Templeton and Ivor Timmis

4 Speaking and Listening: Two Sides of the Same Coin
Michael McCarthy and Jeanne McCarten

5 Corpus Linguistics and Writing Instruction
Eric Friginal, Ashleigh Cox and Rachelle Udell

6 Corpus Affordances in Foreign Language Reading Comprehension
Alejandro Curado Fuentes

7 Corpus Linguistics and Grammar Teaching
Christian Jones

8 Corpus Linguistics and Vocabulary Teaching
Leo Selivan

9 Culture in English Language Teaching: Let the Language Do the Talking
Kieran Harrington

10 World Englishes and the Second Language Classroom: Why Introducing Varieties of English Is Important and How Corpora Can Help
Sarah Buschfeld and Emily Rose Weidle

11 Annotating VOICE for Pedagogic Purposes: The Case for a Mark-up Scheme of Pragmatic Functions in ELF Interactions
Stefanie Riegler

12 Detecting and Analysing Learner Difficulties Using a Learner Corpus Without Error Tagging
Gerold Schneider

13 The Potential Impact of EFL Textbook Language on Learner English: A Triangulated Corpus Study
Elen Le Foll

14 Conclusion
Patricia Ronan and Kieran Harrington

Index

List of Figures

Fig. 6.1 Number of works dealing with corpora and reading including and excluding DDL

Fig. 6.2 Concordance reading of “advertising” to discern it as either an adjective or a noun

Fig. 6.3 Drag-and-drop activity with collocations derived from the concordances

Fig. 8.1 Lextutor concordance screen

Fig. 8.2 Collocation forks

Fig. 9.1 Concordance screen (www.lextutor.ca, last accessed January 2022)

Fig. 9.2 Concordance lines for *sorry*

Fig. 9.3 Concordance lines of *sorry* sorted alphabetically one word to the left

Fig. 9.4 Expanded text for concordance line 84

Fig. 12.1 Hits for *make decision*

Fig. 12.2 Hits for *if woman*

Fig. 12.3 Hits of *just few* in the ICLE corpus

Fig. 12.4 Hits for *understand problems* in the ICLE corpus

Fig. 12.5 Hits of *take the advantage* in the ICLE corpus

Fig. 12.6 Hits of *been a widely* in the ICLE corpus

Fig. 12.7 Relative frequencies of Determiners by L1

Fig. 12.8 Frequency of determiners in spoken and written genres of the ICE corpora

Fig. 12.9 Hits for *discuss about* from ICLE

Fig. 13.1 Screenshot of the advanced concordance search function on sketchengine.eu

Fig. 13.2 Screenshot of a random sample of query results with three selected concordance lines

Fig. 13.3 Screenshot of the text type options when querying the OCLC

List of Tables

Table 2.1 What do respondents consider positive about working with corpora

Table 2.2 What do the respondents consider negative about working with corpora?

Table 3.1 Possible initial searches on the embedded COCA tool

Table 6.1 Scheduled activities and tasks in each course (* activities carried out in pairs or/and groups)

Table 6.2 Statistical significance in test comparisons

Table 7.1 Occurrences per million words of *What are you going to do?* in a TV corpus

Table 9.1 Discoursal functions of *sorry* after a pause and in two-word collocations

Table 9.2 Extraction of *sorry*

Table 12.1 Verb-Object-PP collocations, sorted by decreasing O/E

Table 12.2 Top 70 missing determiners in ICLE

Table 12.3 Extra determiners in ICLE, top 60 sorted by decreasing collocation score (O/E)

Table 12.4 Verb + Preposition overuse in ICLE, sorted by decreasing T ratio

Table 13.1 Intermediate-level causative constructions with MAKE

Table 13.2 Extract from *New Missions 2^e^*, p. 208 (emphasis original)

Table 13.3 Causative constructions in the TEC-T

Table 13.4 Most significant results from the collexeme analysis of [X MAKE Y V_{inf}] in the TEC-T (for full results see Appendix 3)

Table 13.5 Composition of the Textbook English Corpus (TEC)

Notes on Contributors

Sarah Buschfeld
is Full Professor of English Linguistics (Multilingualism) at TU Dortmund University (Germany), after previous appointments at the universities of Regensburg and Cologne. She has worked on postcolonial and non-postcolonial varieties of English (e.g., English in Cyprus, Greece, Namibia, Singapore, and St. Maarten) and in the field of language acquisition and multilingualism in and outside the classroom. She has written and edited several articles and books on these topics (including *English in Cyprus or Cyprus English: An Empirical Investigation of Variety Status*, 2013; *Children's English in Singapore: Acquisition, Properties, and Use*, 2020) and explores the boundaries between such disciplines and their concepts. Her research focus is empirical, with a strong corpus linguistic orientation.

Ashleigh Cox
is a graduate student in the Department of Applied Linguistics and ESL at Georgia State University in Atlanta, Georgia, USA. Her research interests include corpus linguistics, EAP, TESOL, sociolinguistics, and intercultural communication. Her research focuses on analyzing patterns in corpora of discipline-specific academic texts. She has also been involved in other projects on EAP writing, second language teaching, and English for aviation. In addition to research, she is also interested in language teaching. She has experience teaching English as a second language in programs with academic and non-academic foci, and she has taught both international students and adults living in her community. Because of her background in language pedagogy, she is interested in applied linguistics research that has practical implications for language learning and teaching.

Alejandro Curado Fuentes

is a senior lecturer at University of Extremadura, Spain, where he teaches English for specific purposes (ESP) and applied linguistics at graduate and post-graduate levels. He belongs to GexCALL (Extremadura's Research Group for Computer-Assisted Language Learning). He has participated as a main researcher and assistant in several national and international projects related to the use of information technology (IT) in English as a foreign language (EFL) teaching and learning. He has published extensively on corpus-based analyses and IT developments in ESP/EFL in different journals and edited volumes. He is the president of AELFE (European Association of Languages for Specific Purposes).

Eric Friginal
is Professor of Applied Linguistics and Head of Department of English and Communication at The Hong Kong Polytechnic University. Before moving to Hong Kong, he was Professor and Director of International Programs at the Department of Applied Linguistics and ESL and the College of Arts and Sciences at Georgia State University, USA. He specializes in applied corpus linguistics, quantitative research, language policy and planning, technology and language teaching, sociolinguistics, cross-cultural communication, discipline-specific writing, and the analysis of spoken professional discourse in the workplace. His recent publications include *Corpus Linguistics for English Teachers: New Tools, Online Resources, and Classroom Activities* (2018) and *Advances in Corpus-Based Research on Academic Writing: Effects of Discipline, Register, and Writer Expertise*, co-edited with Ute Römer and Viviana Cortes (2020). He is the founding co-editor-in-chief of *Applied Corpus Linguistics* (ACORP) Journal (with Paul Thompson, University of Birmingham, UK).

Kieran Harrington
has been involved in English language teaching for over thirty-five years, as a teacher at all levels of education in Spain and in Ireland, and

as a trainer for EFL certification and lecturer in third level TESOL programs. He began his university teaching career at University College Cork, where he taught Spanish and Galician. He worked as senior lecturer and acting professor of Applied Linguistics and Language Teaching at TU Dortmund University before moving to a senior lecturer position in English Linguistics, also at TU Dortmund University in 2020. Since 2006 he has been a senior civil servant in the Galway and Roscommon Education and Training board (Ireland), leading projects in literacy and language teaching provision; he was also director of CELT training and coordinator of English language education for refugees and asylum seekers. Apart from corpus linguistics, his main research interests are linguistic ethnography, conversation analysis, and language disorders (dyslexia and aphasia). He is author of *Survival Communication: The Role of Corpus Linguistics in the Ethnography of a Closed Community* (2018).

Christian Jones
is a Reader in TESOL and Applied Linguistics in the Department of English at the University of Liverpool. His main research interests are connected to spoken language and he has published research related to spoken corpora, lexis, lexico-grammar, and instructed second language acquisition. He is the co-author (with Daniel Waller) of *Corpus Linguistics for Grammar*: *A Guide for Research* (2015), *Successful Spoken English: Findings from Learner Corpora* (with Shelley Byrne and Nicola Halenko) (2017), editor of *Literature, Spoken Language and Speaking Skills in Second Language Learning*, (2019), and author of *Conversation Strategies and Communicative Competence* (2021).

Elen Le Foll
is a research associate and English education lecturer at Osnabrück University (Germany). Having taught French and English in adult education for many years, she is now a passionate teacher trainer and has been teaching on the university's teacher training program since

2016. Her primary research interests include applications of corpus linguistics in language teaching and learning, materials development and evaluation, language learners' use of online resources, and the development of (student) teachers' critical digital competences. Her research project involves the corpus-based analysis of the language of English as a foreign language (EFL) textbooks used in European secondary schools. As an Open Education advocate, she is keen to explore alternative science communication channels and to involve student and expert language practitioners—for example, in the co-creation of Open Educational Resources—to bridge the gap between applied linguistics research and language teaching and learning.

Jeanne McCarten
taught English in Sweden, France, Malaysia, and the UK, before starting a publishing career with Cambridge University Press. As a publisher, she had many years' experience of commissioning and developing ELT materials and was involved in the development of the spoken English sections of the Cambridge International Corpus, including CANCODE. A freelance ELT materials writer, corpus researcher, and occasional teacher, her ongoing professional interests lie in successfully applying corpus insights to learning materials. She is co-author of the corpus-informed print, online and blended courses *Touchstone* and *Viewpoint*, and *Grammar for Business*, published by Cambridge University Press.

Michael McCarthy
is Emeritus Professor of Applied Linguistics, University of Nottingham, UK, and Teaching Officer at ADTIS, Cambridge University, UK. He is author/co-author/editor of 58 books, including *Touchstone, Viewpoint, the Cambridge Grammar of English, English Grammar: The Basics, From Corpus to Classroom, The Routledge Handbook of Corpus Linguistics, McCarthy's Field Guide to Grammar*, and titles in the *English Vocabulary in Use* series. He is author/co-author of 120 academic papers. He was co-founder of the CANCODE and CANBEC spoken English corpora

projects. His current research focuses on spoken grammar. He has taught in the UK, Europe, and Asia, has given talks and workshops in 46 countries, and has been involved in language teaching and applied linguistics for 56 years.

Stefanie Riegler
is a university assistant and PhD candidate in Applied Linguistics at the English Department of the University of Vienna. Her research interests include the pragmatics of English as a lingua franca (ELF) communication, (pragmatic) corpus annotation, the implications of ELF research for language pedagogy, and language (education) policy. In her dissertation, she works toward a methodology for annotating pragmatic functions in the *Vienna-Oxford International Corpus of English* (VOICE). She was part of the VOICE CLARIAH project team which developed and released VOICE 3.0 Online (https://voice3.acdh.oeaw.ac.at/). Her article 'Normativity in Language Teacher Learning: ELF and the European Portfolio for Student Teachers of Languages (EPOSTL)' has been published in the *Journal of English as a Lingua Franca* (https://doi.org/10.1515/jelf-2021-2048).

Patricia Ronan
holds a chair of English Linguistics at TU Dortmund University. She received her PhD at Maynooth University (Ireland) and she has held further positions in Spain, Sweden, Switzerland, and Germany. Her main research interests are in language variation and language contact. At the time of writing, she is co-authoring an Introduction to Multilingualism (with Sarah Buschfeld and Manuela Vida-Mannl), and is co-editing volumes on Language and Migration (with Evelyn Ziegler) and on Corpus Pragmatics (with John Kirk). She is further working on projects on media language, on variationist pragmatics, and on the linguistic inclusion of migrants.

Gerold Schneider
is a senior lecturer, researcher, and computing scientist at the Department of Computational Linguistics at the University of Zurich, Switzerland. His doctoral degree is on large-scale dependency parsing, his habilitation on using computational models for corpus linguistics. His research interests include corpus linguistics, statistical approaches, digital humanities, text mining, automated content analysis, and language modeling. He has published over 120 articles on these topics. He has published a book on statistics for linguists (https://dlf.uzh.ch/openbooks/statisticsforlinguists/). He leads the Text Crunching Center, a University platform which offers services in text processing and Digital Humanities. His Google scholar page is https://scholar.google.com/citations?user=l_8L7NYAAAAJ.

Leo Selivan
started his ELT career with the British Council, as a teacher, before moving into materials and course development, and teacher training. Today, he is a lecturer offering courses to both pre- and in-service teachers as well as language editors. His professional interests include second language acquisition, corpus linguistics, and lexico-grammatical interface—the topics he has written on not just for his own blog *Leoxicon*, but also several well-known teacher journals. Known on social media as 'Lexical Leo', he is passionate about the Lexical Approach, which was the subject of his first book, *Lexical Grammar* (2018). His second book, *Activities for Alternative Assessment*, came out last year. No less importantly, he is still a practicing EFL teacher with almost 20 years of experience.

Jane Templeton
is Lecturer in EAP at the University of Leeds. Her main research interests are in the practical application of corpus linguistics in the language classroom and the development of learner autonomy. She has presented on these topics at conferences and seminars. She is the co-

author, with Ivor Timmis, of 'DDL for English Language Teaching in Perspective' in the *Routledge Handbook of Corpora and Language Learning and Teaching*.

Ivor Timmis
is Emeritus Professor of English Language Teaching at Leeds Beckett University. His main pedagogic research interests are in corpus linguistics and materials development, and the relationship between the two fields. He has published extensively in these areas. More recently, he has become interested in historical spoken corpora. He is the co-author, with Freda Mishan, of *Materials Development for TESOL* and the author of *Corpus Linguistics for ELT*.

Rachelle Udell
is a doctoral student of applied linguistics at the Department of Applied Linguistics and ESL at Georgia State University, Atlanta, GA, USA. Her research specializations include English for specific purposes (ESP), ESP-Aviation English, L2 literacy, corpus-based discourse analysis, and intercultural communication. She also focuses on various language policy and planning implications of Aviation English, particularly on written documents such as technical manuals and texts from other data sources such as Controller-Pilot Data Link Communications (CPDLC).

Emily Rose Weidle
is a student of Applied Linguistics at TU Dortmund University (Germany). She works for the Chair of English Linguistics (Multilingualism) and has shown an interest in empirical applied linguistics right from the beginning of her studies. She has supported project work on 'English in multilingual St. Maarten' and 'English for touristic purposes: An investigation in Cuba and Croatia' by helping in the data collection and preparation processes and has thus gained

valuable insights into empirical, corpus linguistic practices and methods at an early stage of her career. Her areas of interest include multilingualism, English language teaching in Germany, World Englishes, and sociolinguistics.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_1

1. Demystifying Corpus Linguistics for English Language Teaching

Kieran Harrington[1] ✉ and Patricia Ronan[1] ✉
(1) Faculty of Cultural Studies, TU Dortmund University, Dortmund, Germany

✉ **Kieran Harrington (Corresponding author)**
Email: kieran.harrington@tu-dortmund.de

✉ **Patricia Ronan**
Email: patricia.ronan@tu-dortmund.de

Keywords Corpus linguistics – Concordancing – Keywords – POS tagging – Data-driven learning – Corpus Query Language (CQL)

1.1 Introduction

Are you interested in language teaching? Do you feel that you would like to kindle the students' or pupils' interest in 'real' English and work with authentic language data? Then this book is for you. Especially if you have been thinking about using corpora for this, but also if you have not yet considered this possibility, we would like to use the opportunity to describe how corpora can be employed in various aspects of English language teaching.

With this volume we aim to demystify corpus linguistics for use in English language teaching (ELT). We advocate the inclusion of corpus linguistics in the classroom as part of an eclectic approach to language teaching, the main benefit of which is the engagement of students with naturally occurring language.

Before we present an outline of the chapters, we first provide a brief introduction to corpus linguistics, directed principally at those readers who may not be familiar with the field or may look at the field with certain apprehension. This will be followed by a review of the use of corpus linguistics in ELT.[1]

1.2 Corpus Linguistics

Modern corpus linguistics refers to the collection of spoken and written texts and their analyses by software programmes. These texts may come from recorded casual conversation or formal interactions such as meetings, and from written material such as media discourse and fiction writing. These collections are known as *corpora*, or *corpus* in the singular. They are a principled collection of texts in the sense that they are representative of a particular language, discourse variety or genre and the design matrix captures all the necessary variables, such as age and gender.

The use of electronic corpora in linguistics started in the 1960, partly as a reaction to the surge in theoretical linguistic approaches, such as that of Noam Chomsky's (1957) work on syntax. In contrast to many theoretical linguists, corpus linguists believe that the nature of language is best studied and described on the basis of real, naturally occurring language, and that how speakers actually perform, including their errors, offers the best perspective of not only their language performance, but also their competence. Corpus linguistic methodologies are based on verifiable and repeatable approaches and different researchers can run the same or comparable experiments—or indeed can perform the same experiments on different data sets und use the results to describe differences. We can use corpus data to provide us with insights into how speakers of any variety of a language, be they native speaker or learner varieties, use a feature or an item, with which frequency the items are used, and what is specific to any particular varieties of a language in comparison to other varieties. All these arguments make corpora a prime resource for language teaching: showing teachers and their students alike whether a structure is found in a target language, in which contexts it is found, how frequent it is, and how it might differ in different varieties of the target language,

makes corpus-based approaches prime resources for language teaching. Being able to access these sources of knowledge with quick and efficient computerized approaches makes this even better.

A first landmark in the development of computer-based corpora was the compilation of the Brown Corpus by Kučera and Francis at Brown University in the United States and its release in 1961. It contains one million words of text samples of American English from a broad variety of genres of written American English. It comprises 500 texts, each of them about 2000 words long, from different genres of press writing, religious texts, skills and hobbies, popular lore, *belles lettres*, academic texts, different genres of fiction writing, romance and humour. The genre balance that is found in this corpus—and other corpora—allows researchers to obtain a good overview of how genres differ from each other in terms of grammar, lexis or style.

In the 1970s, Geoffrey Leech and his team at Lancaster University joined forces with teams from Oslo and Bergen universities and created a one million word corpus of British English, the Lancaster Oslo Bergen (LOB) Corpus, designed to be comparable with the Brown Corpus in order to facilitate comparative research between the two varieties of English. The texts making up this corpus also stem from 1961 and follow the same genre distribution as the Brown corpus. Since then, additional corpora have been added to these for comparison purposes: the Freiburg Brown Corpus (Frown) and the Freiburg LOB Corpus (FLOB), based on American and British English texts from 1992 respectively, as well as the B-LOB and B-Brown Corpora, based on materials from 1931, produced at Lancaster and Zurich Universities respectively, which give researchers the possibility to compare these varieties also across time.

The Brown and LOB family corpora are joined by another family of 1 million word corpora, the International Corpus of English (ICE) family, conceived in the late 1980s by Sidney Greenbaum. The idea behind the creation of the ICE corpora was to make varieties of English from across the globe comparable by creating similarly sampled corpora from different English speaking countries world-wide (Nelson et al. 2002). The ICE corpora contain 1 million words each, 600,000 words of transcribed spoken English from different contexts, 400,000 from different written genres. At the time of writing in 2022, ICE corpora are

available from Canada, East Africa, Great Britain, Hong Kong, India, Ireland, Jamaica, New Zealand, Sri Lanka, Nigeria, the Philippines, Singapore and for the USA, the written corpus component.

These smaller corpora have been joined by a number of large scale corpus projects. In 1980 John Sinclair initiated a project later known as COBUILD (Collins Birmingham University International Language Database) which originally contained 7.3 million words of balanced written and transcribed spoken data from both British and American English. In the course of time more materials were added to it to reflect the continued development of the English language. The Bank of English Corpus now contains 650 million words; these form the basis of a series of dictionaries and grammars published by Collins.

In the early 1990s, teams from Oxford University, Lancaster University and the British Library, together with the publishers Longman, Chambers and Oxford University Press, started to produce the British National Corpus, released in 1994. Its 100 million words also comprise ten million words deriving from spoken interaction. The written texts stem from different sources of fiction and non-fiction materials. In 2018, the 11.5 million-word spoken component of the ongoing BNC2014 project was released. This project is designed with the intention of facilitating comparison with the earlier BNC data in order to capture changes that may have occurred since then. The BNC can be accessed through various online corpus interfaces free of charge.[2]

The most used corpus of American English is the Corpus of Contemporary American English (COCA), created by Mark Davies at Brigham Young University in the United States.[3] It contains over one billion words from between 1990 and 2019. It can be accessed freely or by subscriptions through its web-interface. It, too, contains a broad range of genres from different sources: transcriptions of spoken data, largely from media sources, fiction, popular magazines, newspapers and academic writing. In addition to COCA, the Brigham Young corpus suite also contains the Corpus of Historical American English (COHA), consisting of more than 475 million words of English from between the 1820s and the 2010s. We further find various other corpora, predominantly of media and online language, which make interesting and attractive sources of data[4]: the GloWbE Corpus of Global Web-

based English, the News on the Web (NOW) corpus, the corpus of Soap Operas (SOAP), the MOVIE corpus, based on international movies, to name but a few.

In addition to this selection of general corpora of different varieties of English and special themed corpora, we also find corpora that have been compiled with a particular view to being used in language teaching. Here we can mention the International Corpus of Learner English (ICLE, Granger et al. 2020), which has been compiled by the Université catholique de Louvain and released initially in 2002. The current version contains 5.5 million words and comprises subcorpora of essay writing by upper and intermediate learners of English by learners with different first languages, for example Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish and Swedish. Though the different corpus components are rather small (roughly 200,000 words each), this corpus is very useful for comparing typical features of the language of learners of a specific first language. It thus offers the possibility of identifying specific language acquisition features and allows for the production of targeted teaching materials.

Other landmark developments in the compilation of corpora include the Cambridge and Nottingham Corpus of Discourse English (CANCODE), a corpus of 5 million words of spoken British and Irish English which has been used extensively in ELT studies, and the Vienna International Corpus of English (VOICE), compiled by the Department of English at the University of Vienna, which was the first corpus exclusively comprising English as a lingua franca (ELF). A further good source for the study of English is the Cambridge English Corpus (CEC), formerly the Cambridge International Corpus (CIC), which contains many billions of words. It comprises texts from a wide range of genres of spoken and written language (natural-occurring talk, interviews, meetings, magazines, newspapers, academic articles, student essays, etc.). This corpus further includes the 40 million-word Cambridge Learner Corpus, compiled from the written exam responses of English language learners.

Much smaller corpora can also be compiled by researchers, teachers and students alike, who are interested in specific uses of language in particular communities or contexts. As McCarthy and Carter (2001) say,

‘size isn’t everything’, pointing out that small but carefully constructed corpora can produce fruitful research-specific results and answers to research questions. The size of the corpus depends as Timmis (2015: 2) comments, on *fitness for purpose*. Usually such corpora contain fewer than 1 million words. Examples of individual researchers compiling such corpora include Harrington (2018), who compiled a corpus of 98,000 words of the discourse of asylum seekers confined in an institution, who used a very reduced form of English to communicate, and Buschfeld (2020), who compiled a corpus of 48,360 words of children’s discourse to inquire into the acquisition and characteristics of Singapore English as a first language.

1.3 Corpus Tools and Software

When using corpora, we can do so in two basically different ways. First, we can approach corpora with a fixed hypothesis in mind, such as we might expect that in connection with the temporal adverbial *since*, we would typically find verb forms in the present or past perfect, and in the context of the temporal adverbial *yesterday*, we would typically find past tense verb forms, and we would then use corpora to verify—or not—this hypothesis. This approach is typically referred to as a corpus-based approach (Tognini-Bonelli 2001). Alternatively, we can approach a corpus without any fixed expectations in mind and be informed by the data that we find. Such an approach is known as a corpus-driven approach. Either of these may be used in classroom settings, of course. Corpus based approaches are likely to be used if we want to find out what is ‘right or wrong’ in a given context. Corpus-driven approaches would be employed if corpus users want to obtain an overview of possible structures that can be found in a given context.

Whichever approach we adopt, when using corpus linguistic methodologies, these are now largely associated with computer and software programs that can perform a multitude of tasks on what are now known as machine-readable corpora. These programs can facilitate simple word frequency searching, concordancing, key word analysis at its most basic, but also sophisticated statistical research and querying.

There are many software programs available both for purchase and download, such as Wordsmith Tools (Scott 2020), or as freeware which can be used online through a browser. Wordsmith Tools was one of the earliest software packages and was developed by Mike Scott at the University of Liverpool as an elaboration of MicroConcord, originally co-developed by Mike Scott and Tim Johns and published by Oxford University Press in 1993. A widely used freely accessible concordancing tool is AntConc (Anthony 2022), developed by Laurence Anthony at Waseda University, Japan.

The larger corpora, such as COCA and the BNC have their own corpus software on their platforms and facilitate both simple concordancing and querying. As such they seem more suitable for work in English language teaching as they allow the user to investigate large corpora, both written and spoken of many different genres. Smaller platforms, such as Compleat Lextutor (Cobb 2021), offer the user the possibility of searching reduced versions of large corpora (such as COCA and the BNC), allowing for the display of concordance lines, consultation of the original text, frequency calculation and even the possibility of creating language exercises such as cloze tests. The corpus tool that is mainly used in English language teaching is the concordancing tool.

1.4 Concordancing

Concordancing tools find every occurrence of a word or string of words —referred to as the 'node' in corpus linguistics terminology—in a corpus. Once the node is input into the search box of the tool, the concordancing programme will present quantitative information (how many times the node occurs in the corpus) and display the concordance lines with the node in the centre—with an average of eight key words in (the) context (KWIC) to either side. These raw concordance lines will usually enlighten the user with regard to the function or meaning of the node-word in the particular case; however, for problematic cases, the larger source text can be uncovered. Concordancing programmes also facilitate alphabetical sorting of the lines (to the left and right of the node), which makes the discovery of patterns, an essential feature of corpus linguistics, much easier. Concordancing is also the main tool

used in what is called data-driven learning (DDL)—which at its most basic refers to learners engaging with large amounts of authentic language with the purpose of discovering patterns.

1.5 Word Frequency Listing

The use of word frequency listing is not a new phenomenon and was used well before the advent of super-fast micro-computers. Previously, the corpora were manually compiled and the most frequent words in the corpora manually counted and classified. A good example in the context of education is what is known as the *Dolch List*, first compiled in the 1930s by Edward Dolch and still used today in literacy studies and literacy grading. The latter compiled a corpus of children's literature of the era and then established the 220 most frequent 'service words' as they were called, which were (and still are) to be learned by children. It was estimated that this list covered 80% of the words found in typical children's books of the era and 50% of the words found in adult writing. The idea was that if the child knew such a high percentage of words then reading would be much easier. 63 years after Dolch's mention of his project in a journal article in 1936, a seminal article appeared by Michael McCarthy (1999), in which he posited, based on a computational analysis of the CANCODE, that 2000 (the most frequently used) words did most of the work (80%) in conversation. The implications of this and similar work done (for example Nation 2006 and Nation and Waring 1997) for language teaching has been immense.

1.6 Words and Key Word Analysis

Before continuing, it is appropriate to define what we mean by *words* in the context of corpus linguistics. Three terms that the reader will encounter when using corpus linguistics to investigate words are *lemma*, *token* and *type*. A lemma is the basic form of a word that appears in a dictionary. For example, the lemma *like* (or LIKE), appears as the main entry, but not the inflectional forms (*likes, liking, liked*). Another distinction that is important for corpus linguistics is that between *tokens* and *types.* Each occurrence of a particular word or a

particular lemma in a corpus is called a token. This is also what software programs, such as Wordsmith Tools or AntConc, report as word count. *Types* are the unique word forms in a text. The type/token ratio is determined by dividing the number of types by the number of tokens, which can inform the user of the range of vocabulary in a text and as such may be useful for teachers in determining the level of vocabulary difficulty in a reading exercise.

Key words are those which have *unusually* high or low frequency. The article *the*, for example, is always high (as in the most frequent word) in word lists compiled from large English corpora, but that is not surprising. This follows for all function words. The objective is doing keyword analysis then is to discover what words in a corpus might be particularly but unusually frequent in comparison to a larger reference corpus. Such *keyness* is tested by means of a statistical association metric (such as the Loglikelihood Test, the t-test and the Fisher Test) which is built into the software program. The important thing for English language teaching is that key word analysis can be used by material developers and by teachers themselves to create more focussed frequency lists for English for Specific Purposes (ESP)—for fields such as business English, aeronautical English, etc.

Another concept that is basic to corpus linguistics in its treatment of words is *collocation,* which as one can deduce from its form, has something to do with words co-locating or being located close to one another. In the literature on corpus linguistics and in corpus-informed academic literature the concept of collocation has varying definitions, but for the purposes of introductory work on corpus linguistics in English Language Teaching it can be taken fundamentally as the phenomenon of certain words frequently appearing together or close together. The verb *perform,* for example, collocates with *operation* but not with *discussion.* The verb *do* collocates with *damage, duty* and *wrong,* but not with *trouble, noise* or *excuse.* Concordancing software can search through corpora rapidly and find such collocations, which in turn provides useful information for teaching.

Further, the concept of *chunk* needs to be considered. Timmis (2015: 27), for example, defines a chunk as "a frequent meaningful sequence of words which may include both lexical and grammatical words", which he distinguishes from a *lexical bundle*, "a sequence of

words found together without a clear semantic or pragmatic meaning". An example of a chunk would be *at the end of the day*, and an example of a lexical bundle would be *it was a.* All corpus software provide a tool for finding, listing and ranking chunks according to frequency. However, the software does not distinguish them. It may find *that sort of thing* as a frequent four-word chunk, which is of course meaningful and recognisable. It is a string of words that people tag on to the end of an utterance. But it may also find *you know what I* as one of the most frequent strings of words. It is not a meaningful sequence of words and as such may not be considered useful for vocabulary instruction; however, that is not to say that it is not a string of words that we can throw on the rubbish pile, because it obviously is frequent in spoken language and will combine with words such as *heard, saw,* and *think.*

1.7 Corpus Annotation and Query Language

In order to be fully exploitable, corpora should be annotated with sufficient information to allow the corpus user to interpret the data. Thus corpora typically provide meta annotation concerning where the data come from, and often even demographic information on each data source. For example a learner corpus may provide the first language of a speaker, their age, their gender, their language level, etc. Further, both corpora built for specific research purposes and larger corpora (such as the BNC) can be used in a raw state—that is, the user can simply search for words or groups of words and check their frequency or observe their behaviour in concordance lines, this is normal procedure in what we call data-driven learning (see Sect. 1.8 below) in the classroom. However, if there is a need to check on the presence, frequency or behaviour of specific parts of speech, then the user needs to use a linguistically annotated corpus. In a linguistically annotated corpus, all the words have a tag or label attached indicating what part of speech they belong to.

If information on specific parts of speech is needed from a corpus, then the researcher will carry out more sophisticated grammatical querying, which is a process of the researcher telling the computer what to find; but the computer can only find this information if the corpus has been tagged. This is a task that is carried out in the

background during the compilation of the corpus and is facilitated by software tagging tools such as CLAWS (the Constituent Likelihood Automatic Word-tagging System), which was developed in the 1980s at Lancaster University by the University Centre for Computer Corpus Research on Language (UCREL). A list of labels, such as AT (article), JJ (general adjective), NN (singular common noun), BEZ (code for the word *is*) used to indicate the part of speech (POS) is known as a *tagset.* The labels are added to a word—or coded—by means of an underscore symbol. The first CLAWS software produced had 132 basic wordtags—but these are consistently updated. The following is an example of such POS tagging:

hospitality_NN is_BEZ an_AT excellent_JJ virtue_NN.

More sophisticated tagging includes grammatical parsing, semantic tagging, prosodic annotation, and pragmatic tagging. Grammatical parsing, which facilitates more sophisticated investigation of syntax, assigns a phrase-marking label to each sentence and distinguishes, for example, between singular noun phrases and plural noun phrases and between adjective phrases and adverb phrases. Parsing refers to the process of syntactic analysis and a *parser* is a software tool used for this purpose. The principal goal of semantic tagging is to distinguish the particular sense of a particular word in discourse. Furthermore, pragmatic annotation identifies the pragmatic function of language in use dependent on context variables such as situation and time and identity. Finally, prosodic annotation provides information on intonation patterns, stress and pause in spoken corpora. A well-known example of a prosodically annotated corpus is the London-Lund Corpus of Spoken English, based on two projects, one started by Randolph Quirk in London in 1959 and the other by Jan Svartvik in Lund in 1975 (Svartvik 1990) . Other broader types of tagging include the annotation of corpora of different Englishes used around the world, and of English as a lingua franca, that is English used as a means of communication in cases where speakers do not share the same language, and learner English.

It should be noted that some linguists prefer to engage with corpora in their raw state, that is without annotation, because they consider

annotated corpora too interpretive, reflecting the errors and even the biases of the annotator. However, many linguists nowadays consider annotation an enrichment. It is vital for more sophisticated and complex corpus querying through the use of corpus query language (CQL), which is a powerful function that facilitates the searching of corpora for complex lexico-grammatical patterns.

1.8 Corpus Linguistics and English Language Teaching

Corpus linguistics can impact language teaching in two principal ways. Firstly, the analysis of naturally occurring language can provide a multitude of insights for materials development, such as for learner dictionaries, wordlists and course books. Secondly, corpus linguistics can be implemented pedagogically in the classroom. This approach is known as 'data-driven learning' (DDL). Students observe and examine naturally occurring language (the data) and reach conclusions with regard to usage. In a sense, the students become researchers themselves, but the pedagogical essence is that the learning is inductive and autonomous, and in line, for example, with the European Framework for Key Competences (2019), the students apply higher order thinking skills and create their own knowledge which promotes the development of basic skills for lifelong learning.

Since the first use of the term by Johns (1991), DDL has been firmly associated with a lexicogrammatical approach to language (see Boulton and Cobb 2017), where vocabulary and grammatical structures are seen as interdependent, and as a "new form of grammatical consciousness raising" (Hadley 2002: 8). However, data-driven learning also has huge potential in other areas where it has not been much utilized up to now, such as reading processes (see Curado Fuentes this volume), culture and language (see Harrington this volume) and the incorporation of World Englishes in the ELT classroom (see Buschfeld and Weidle this volume).

As far as actual classroom implementation is concerned, the teacher can print out concordance lines for the students to explore the discourse (language above and beyond the sentence), under his or her

guidance. This is referred to as the 'soft version' of DDL by Gabrielatos (2005). Otherwise, the students can use classroom computers or personal devices, such as laptops and smartphones, to do corpus linguistics themselves by using software tools to engage with the corpora. This is referred to as the 'hard version' of DDL by Gabrielatos (ibid.). The rationale of DDL, apart from being driven by the important consideration of inductive learning, is based on the potential that the observation and examination of authentic language has for the process of language learning and teaching. It could be argued that the concordance lines or stretch of expanded text lose their authenticity once they have been removed from their original context. However, such criticism might be misplaced when the objective in practice is simply to discover attested frequencies and patterns in naturally occurring language. It is true that these authentic texts may prove taxing for weaker students especially. The challenge is increased by the fact that the concordances are multimodally presented in the sense that we are not just reading from left to right anymore, but also from top to bottom. Yet, here is an opportunity for students to see the reality of language: in a soft version of DDL, the students can discover for themselves how structures can be used in different contexts. In a hard version of DDL, they can learn that language is not something that follows prescriptive rules, but is ruled by language users' performance as well as their creativity.

Although in direct data driven learning the students themselves engage in corpus linguistics, this does not mean that the teacher cannot scaffold the engagement, and it goes without saying, of course, that the corpora and the discourse should be carefully selected for both direct and indirect engagement.

1.9 Chapter Outline

In Chap. 2, Patricia Ronan investigates the attitudes of German and Swiss teacher training students towards the use of corpus linguistics in English language teaching. She finds that a considerable number of respondents point to the complexity of corpus use in classroom and to the extensive amount of time needed to prepare for its implementation. She calls for a greater use of indirect corpus methods and for more focussed training in corpus methods for trainee teachers. This sets the tone for what is to follow. In Chap. 3, Jane Templeton and Ivor Timmis address the reservations which have been expressed about the viability of data-driven learning in the classroom and outline a flexible framework for its integration in everyday classroom practice, a key feature of which is what they call 'systematic opportunism'. This consists of teachers introducing learners to relevant DDL techniques and resources as and when the opportunity or need arises in the classroom, the longer-term aim being the eventual autonomous use of DDL affordances. In Chap. 4, Michael McCarthy and Jeanne McCarten adduce corpus evidence to support the view that speaking and listening should not be taught as separate skills in English language teaching but be integrated as part of a 'fifth skill', *interaction*, into the syllabus. In Chap. 5, Eric Friginal, Ashleigh Cox and Rachelle Udell explore the ways in which corpora can be used to inform the teaching of writing in a second language, with a focus firstly on curriculum design and materials development, and then on ways of teaching learners to use corpora themselves in order to discover features of writing in the target language. In Chap. 6, Alejandro Curado Fuentes covers a topic that few corpus studies have addressed specifically: the potential usefulness of data-driven learning for supporting reading comprehension in English language teaching. He uses a case study of two courses in which learners availed of both web-based corpus tools and non-corpus digital resources to show how reading performance can be improved by this combination and its adaptation to the learning situation. In Chap. 7, Christian Jones focuses on the use of corpus linguistics in grammar teaching and shows how corpora can facilitate teachers' understanding of frequency in different genres and how common aspects of grammar are used in terms of form and meaning(s). He demonstrates how

grammar can be better understood by exploring language in context and how such analysis can inform materials design, methodology and common classroom practices such as teacher explanation. In Chap. 8, Leo Selivan, considering the everyday practical needs of language teachers, demonstrates how corpus tools can be employed in the classroom to enhance vocabulary teaching, firstly with regard to single words, their frequency and use, and secondly in relation to multi-word units or chunks. The suggestions presented in the chapter can be used by both teachers for lesson preparation and students themselves as part of classroom activities.

In the next three chapters we move away from the traditional and core strands of English language teaching to focus on how corpus linguistics can support pluricentric approaches. In Chap. 9, Kieran Harrington illustrates how data-driven learning can be used in English language teaching to raise awareness of the embeddedness of culture in language. He demonstrates how simple concordancing can illustrate how linguistic items function in the immediacy of everyday interaction, the kernel of the interface of culture and language, in the negotiation of meaning and the quest for intersubjectivity. In a step by step approach that can be imitated by both teachers and students, he describes the process using the platform *Compleat Lextutor*. In Chap. 10, Sarah Buschfeld and Emily Weidle explain why introducing different varieties of English and rethinking our old native speaker stereotypes is of crucial importance for successful and up-to-date ELT. They illustrate how corpora can be utilized to convey linguistic variation and to introduce students to different varieties of English and the connections between language and culture. In Chap. 11, Stefanie Siegler, on the subject of English as a Lingua Franca (ELF) and how ELF users exploit language resources strategically and pragmatically, discusses the provision of an annotation system for ELF corpora which would annotate tag communicative processes and pragmatic functions in ELF interactions in such a way as to highlight their pedagogic significance and make them accessible for use in language classrooms. In the final two chapters the focus is on the work linguists do in the background and how they provide insights not just for a greater understanding of language in general, but insights that can facilitate language teaching. In Chap. 12, Gerold Schneider reports on the detection of typical learner

errors and areas of difficulty in learner corpora, insights of which can be then used by materials developers and indeed by teachers to create targeted teaching material for students. In Chap. 13, Elen le Foll investigates how causative constructions are represented in secondary school EFL textbooks and how textbook representations may influence the output of EFL learners. To do this she uses a corpus of the language used in nine series of EFL textbooks used in France, Spain and Germany. Her chapter shows how corpus linguistics can be used in the classroom to bridge the gap between the textbook representation of language and authentic language use—an aim that is central in the use of corpus linguistics in the classroom.

References

Anthony, Laurence. 2022. *AntConc* (Version 4.0.5) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software.

Boulton, Alex, and Tom Cobb. 2017. Corpus Use in Language Learning: A Meta-analysis. *Language Learning* 67 (2): 348–393. https://doi.org/10.1111/lang.12224. [Crossref]

Buschfeld, Sarah. 2020. *Children's English in Singapore: Acquisition, Properties and Use*. New York: Routledge.

Chomsky, Noam. 1957. *Syntactic Structures*. The Hague: Mouton & Co. [Crossref]

Cobb, Tom. 2021. *Compleat Lextutor Tutor.* Lextutor 2021. Accessed 1 March 2022. https://www.lextutor.ca/.

Dolch, Edward W.A. 1936. Basic Sight Vocabulary. *The Elementary School Journal* 36 (6): 456–460. [Crossref]

European Commission, Directorate-General for Education, Youth, Sport and Culture. 2019. *Key Competences for Lifelong Learning*. Publications Office 2. https://data.europa.eu/doi/10.2766/291008.

Gabrielatos, Costas. 2005. Corpora and Language Teaching: Just a Fling or Wedding Bells? *The Electronic Journal for Teaching English as a Second Language* 8 (4): 1–37.

Granger, Sylviane, Maité Dupont, Fanny Meunier, Hubert Naets, and Magali Paquot. 2020. *The International Corpus of Learner English. Version 3*. Louvainla-Neuve: Presses universitaires de Louvain. https://dial.uclouvain.be/pr/boreal/object/boreal:229877.

Hadley, Gregory. 2002. An introduction to data-driven learning. *RELC Journal* 33 (2): 99–124

Harrington, Kieran. 2018. *Survival Communication: The Role of Corpus Linguistics in an Ethnography of a Close Community*. New York: Routledge.

Johns. Tim. 1991. Should You Be Persuaded—Two Samples of Data-Driven Learning Materials. *English Language Research Journal* 4: 1–16.

McCarthy, Michael. 1999. What Constitutes a Basic Vocabulary for Spoken Interaction? *Studies in English Language and Literature* 1: 233–240.

McCarthy, Michael, and Ronald Carter. 2001. Size Isn't Everything. Spoken English, Corpus and the Classroom. *TESOL Quarterly* 35 (2): 337–340.
[Crossref]

Nation, Paul. 2006. How Large a Vocabulary Is Needed for Reading and Listening? *Canadian Modern Language Review* 63: 59–82.
[Crossref]

Nation, Paul, and Robert Waring. 1997. Vocabulary Size, Text Coverage and Word Lists. In *Vocabulary: Description, Acquisition and Pedagogy*, ed. Norbert Schmitt and Michael McCarthy. Cambridge: Cambridge University Press.

Nelson, Gerald, Sean Wallis, and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: Benjamins.
[Crossref]

Scott, Mike. 2020. *WordSmith Tools Version 8*. Stroud: Lexical Analysis Software.

Svartvik, Jan. 1990. The London Corpus of Spoken English: Description and Research*. Lund Studies in English*, 82, Lund: Lund University Press.

Timmis, Ivor. 2015. *Corpus Linguistics for ELT. Research and Practice*. London and New York: Routledge.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: Benjamins.
[Crossref]

Footnotes

1 We would like to thank Carolin Loock, TU Dortmund, for her valuable support and her help in formatting the articles of the volume.

2 BNC 1994: http://www.natcorp.ox.ac.uk/, https://www.english-corpora.org/coca/, BNC 2014: http://corpora.lancs.ac.uk/bnc2014/.

3 https://www.english-corpora.org/coca/, accessed 1 April 2022.

4 https://www.english-corpora.org, accessed 1 April 2022.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_2

2. Learning to Teach English as a Foreign Language with Corpus Linguistic Approaches: A Survey of Teacher Training Students' Attitudes

Patricia Ronan[1] ✉
(1) Faculty of Cultural Studies, TU Dortmund University, Dortmund, Germany

✉ **Patricia Ronan**
Email: patricia.ronan@tu-dortmund.de

Keywords Corpus linguistics – Corpora in the classroom – Indirect corpus use – Direct corpus use – TEFL – Second Language Acquisition – Language attitudes

2.1 Introduction

Corpus linguistics methodologies are, at least in theory, well suited for language teaching as they allow students and pupils to determine correct and incorrect language structures themselves by searching for attestations in corpora (O'Keeffe et al. 2007). That learner language could indeed profit from comparison with native speaker data can be shown, and has been shown, by comparisons of a broad range particularly of lexical, morphological, morpho-syntactic and syntactic items (Braun 2007; McEnery and Xiao 2010; Reppen 2010). However, in spite of the obvious benefits of corpus linguistics for classroom practice, its use also presents diverse challenges for the prospective

users of the methodology in the classroom setting. Here particularly language skills of the pupils, accessibility of materials, difficulty of query language and delimiting source materials can be mentioned (Bennett 2010). Further, the teachers who consider using the resources in their classrooms may not feel sufficiently confident themselves about using corpus linguistic methodologies to teach their students, and research has shown that large percentages of in-service teachers are not familiar with corpus linguistic methodologies (e.g. Mukherjee 2004; Callies 2019).

The current study addresses benefits and challenges of the use of corpus linguistic methodologies from the point of view of teacher training students, who learn about the methodology as students in the university setting themselves and then need to feel motivated and confident enough to pass on their knowledge in a different role, as teachers, in their own classrooms after their graduation. The research aim in this paper thus is to find out which advantages and disadvantages teacher training students see in the use of corpora in the classroom. This issue is investigated by means of an online questionnaire which was distributed amongst students who participated in corpus linguistics classes at two different universities in Germany and Switzerland. On the basis of the 65 responses to the questionnaire, the study shows where the prospective teachers experience difficulties and which benefits they see in the use of this methodology.

After this short introduction to the study, previous core research on using corpora in (English) language classrooms is briefly outlined. After this, the methodology employed in the current study is introduced. Following this, the perceived advantages and disadvantages of corpus work mentioned by the respondents to the survey are outlined. In the next step, the respondents' assessment of these benefits and challenges are presented and then differences between these assessments are thematised. The data is then discussed before conclusions are drawn.

2.2 Using Corpus Linguistics in the Language Classroom

Ample research exists which points out the benefit of using corpus linguistic approaches in the classroom. Thus O'Keeffe et al. (2007), Reppen (2010) and Bennett (2010), amongst others, make strong cases for the usefulness of corpora in language teaching and offer practical, and, in the case of Bennett (2010), even hands-on advice on how to carry out such work. Benefits that are offered by corpus linguistics in the language classroom are that, at least in theory, corpus work is well suited for language teaching and three main uses of corpora have been identified by Leech (1997). First, this is so-called indirect use of corpora, in which a teacher takes a mediating role between the corpus and the pupil. It includes e.g. publishing corpus-based references, developing corpus-based materials, or the use of corpora in language testing. Second, direct use can be made of corpora and the corpora themselves can be employed in the language teaching classroom. Third, in further teaching related developments, specialized corpora may be used for language instruction.

Corpus work has further been shown not only to be beneficial in foreign language learning, but also in the mastery of first languages. In a case study with 25 English as a Foreign Language students in Germany, Braun (2007) found that corpus work also showed beneficial effects on students' first language in secondary school classrooms. She shows that those students who used corpora outperformed the students in the control group, but that to carry out corpus-based work, various skills-related, methodological and logistical problems need to be overcome.

One approach of how to successfully implement corpus work with students is outlined by Bennett (2010: 18–21). She suggests the following structured approach to teachers planning to employ corpora in their classes.

1. Ask a research question.

2. Determine the register on which your students are focused.

3. Select an appropriate corpus for the intended register or compile authentic texts from that register.

4. Use a concordancing program for quantitative analysis of the data.

5. Engage in qualitative analysis.

6. Create exercises for the students.

7. Engage students in a whole-language activity.

As not all learners will have the same level of ability, Bennett (2010: 20) suggests that the exercises should be adapted to learners with weak language levels by asking simple research questions, with teachers finding their own subset of concordance lines, by adapting these for the students' levels, by giving few concordance lines to students and by encouraging group work amongst the students. In a recent study, Callies (2019) argues for the inclusion of corpus linguistics classes and modules in the teacher training curricula. He advocates introducing teacher training students to corpus use by means of hands-on-activities that illustrate researchers', teachers' and learners' perspectives. He stresses the particular value of using both learner corpora and native-speaker corpora in order to increase the teachers' capacity to identify and correct errors.

On the basis of such, and other, well-structured advice, as well as the benefits of the use of corpus linguistics in the classroom that have been outlined in the theoretical work mentioned above and in Chap. 1 (this volume), we might assume that the advantages of employing corpora in language teaching are beyond any doubt and the use of this methodology in classrooms should be uncontested. However, when surveying (student) teachers' experiences with and attitudes to using corpora in and for teaching, a different picture emerges.

Early work on teachers' attitudes on using corpora in teaching was conducted by Mukherjee (2004), who ran workshops on corpus linguistics with in-service English teachers in Germany. Mukherjee found that at the time, the large majority (79.4%) of teachers had been unaware of corpus linguistic approaches before the workshops, but 95% of the teachers found them useful for teaching purposes after attending the workshops. However, only a minority of the participating teachers (12%) were in favour of providing direct corpus access to

their pupils while the majority preferred for the teachers to use the corpus-based methods themselves. Based on these findings, Mukherjee recommends providing more extensive corpus linguistics training to teachers to improve the teachers' competence—as well as their confidence—in corpus linguistic approaches to enable them to use corpus linguistic approaches more extensively.

However, the use of corpus linguistic approaches is still not very widespread even today and in a bid to update Mukherjee's results, Callies (2019) finds that still only 34.6% of 26 in-service teachers participating in his study had heard of corpus linguistics, and also those who were aware of corpora did not necessarily use them inside or outside the classroom. Where corpora are used, the teachers predominantly use them to verify acceptability of language structures. Yet even though around 75% of the respondents partly or largely agree that corpus materials could be beneficial for their pupils and a comparable numbers of respondents would let students work with self-compiled or teacher compiled corpus resources, only 31% largely or partly agree with the statement that they use corpora to design teaching materials (Callies 2019: 251–252).

Where teacher training students are trained in corpus linguistics, studies describing such programs generally report positive results (e.g. Farr 2008; Leńko-Szymańska 2014; Zareva 2017). Thus Leńko-Szymańska (2014) describes that her students at the University of Warsaw appreciated the corpus-based tools and their benefits, but that more extensive time resources than are provided by a single university course are needed to provide the students with enough familiarity with the corpora to become comfortable enough to employ the corpora in classroom teaching. She identifies teachers' lack of knowledge of relevant corpus tools as a major obstacle and calls for more time resources to be allocated to teach how to use corpora in language teaching.

In a recent study, experiences with and attitudes to corpus use of Norwegian in-service teacher are recorded by Kavanagh (2021), who investigates both the familiarity with corpus linguistic approaches, which is comparatively high in his data, and experiences with them. Of his 193 informants, 15% had never heard of corpus linguistics, 39% had little or no idea about it, 28% claimed to be fairly familiar, and 18%

claimed they had already done some work. In in-depth interviews with three teachers who use corpus linguistics in their classes, Kavanagh charts these teachers' usage of the corpora. Concerning attitudes to corpus use, three main areas are reported as problematic (Kavanagh 2021: 16–18). These are issues pertaining to school levels, usability of the corpora and lack of teacher need. The informants pointed out that at a time when pupils would be best equipped to use corpora and their interfaces at higher grade levels, the curricula focus less on language structure for which corpus work would be appropriate than on communication about culture, society and news. The second issue is user-friendliness: corpus interfaces were described as complicated, unattractive and old-fashioned. Finally, his respondents did not feel a need to use corpora for their own information anymore. To address the issues that in-service teachers have with using corpora, Kavanagh (2021: 20) argues that it is not only corpus literacy that is a problem, but a lack of familiarity with corpus tools that have already been reported as working well in classrooms.

The current study is based on data from two universities, TU Dortmund University in Germany and University of Zurich in Switzerland. In spite of the fact that the students had been taught by instructors who are themselves passionate about the use of corpus linguistic approaches in the classroom, the participating students voiced concerns and reservations which make it clear that the use of corpus linguistics in the classroom is not as straightforward as we might have wished ourselves.

2.3 Data and Method

In order to survey experiences with and attitudes to the use of corpus linguistics in the classroom, an online questionnaire was compiled in the Qualtrics Survey programme and distributed amongst students who had participated in linguistics classes with either a focus on, or elements of, corpus linguistic approaches. Participants had attended one of the following four courses: a Corpus Linguistics in the Classroom course at TU Dortmund University, a Research Methods Course at TU Dortmund University, and two corpus-based courses at the University of Zurich. While not all students pursued a teacher training degree at

TU Dortmund University, an average of three quarters of the student population are enrolled in a teacher training programme. At the University of Zurich, as in other Swiss universities and similarly to the situation in countries like Ireland or the United Kingdom, a teacher training programme can only be entered after having successfully completed a degree programme at a university, and the largest percentage of university students of English will embark on a teaching career after their degrees. Many of the teacher training students in both countries are also already gaining teaching experiences as interns or as substitute teachers.

Data collection for the current study took place from February to June 2019. A total of 65 respondents took the survey, and 36 of those completed the survey fully while 29 completed the survey only partially to varying degrees. In the majority of classes, the questionnaire was distributed towards the end of one of the classroom sessions and the students were asked to complete the survey. In one class, the link was provided and it was left to the interested participants to complete the survey in their own time. This approach carries with it the danger that where the students were asked to complete the survey in class, they might have abandoned the survey prematurely either at the end of the class or due to lack of interest. By contrast, those who decided to complete the survey in their own time are likely to have been the more dedicated students, and this is likely to have had an impact on the answers. However, these disadvantages were consciously accepted as a spread across different classes and universities seemed desirable to neutralize effects which instructor-specific teaching approaches might otherwise have had on the results of the study.

The questions in the survey determined whether the students had taken a full corpus linguistics course or whether they had taken a linguistics course with a corpus linguistics component, and how likely they were to use corpora in either teaching or research themselves. Following this, the participants were asked open questions about which advantages and disadvantages they saw in corpus work in general, what they saw as advantages or disadvantages of corpus use for teachers, pupils and the lessons. Finally, the participants were asked to indicate on Likert scales what they considered advantageous or

disadvantageous about corpus use. The results of this survey are presented below.

2.4 Results

In general terms, the respondents to the study have relatively positive attitudes to the use of corpus work. When asked whether they are likely to use corpora for their own teaching or research, of the 75% of the respondents who chose to reply to this question (n = 49), 44% replied they might use it, 24% replied they were likely to use it and 32% thought they were likely to use corpora or had already done so.

The perceived advantages of corpus work in the classroom were manifold, and reiterate many of the points made in favour of corpus work by previous research on the use of corpora in the language classroom (e.g. Reppen 2010; O'Keeffe et al. 2007; Bennett 2010). In general terms, work with corpora is evaluated as interesting, and providing access to vast amounts of authentic data at one's fingertips. However, in line with various other studies (e.g. Leńko-Szymańska 2014; Kavanagh 2021), some respondents report reservations about the complexity of corpus use and corpus searches.

2.4.1 What Are the Advantages of Using Corpora in the Classroom?

The respondents to the survey were asked what they perceived as advantages of corpus use in the classroom for the teachers, for the students, and for the lessons. The questions were asked as open questions, no prompts were given to the respondents.

The perceived advantages of corpus work for the teachers were that by using corpus-based methodologies, non-traditional, student-centred teaching approaches could be introduced. Some respondents were of the opinion that corpus work was also easy to prepare, the materials were already available and they were authentic. Further, the use of media could be introduced into the teaching, and not only could vocabulary lists be readily compiled, and typical mistakes be found, but also answers to difficult questions could be obtained. Additionally, it was considered beneficial that corpus data could also enable pupils to work empirically.

Corresponding advantages were also assumed for the pupils who participated in classrooms where corpus linguistics was used. Corpus-based instruction was thought to be interesting for the pupils because they would encounter varied, real-life authentic materials, which would enable the students to improve their vocabulary and their syntax. Various respondents mentioned that the pupils would further encounter important chunks, phrases and collocations and learn important skills for their later professional life. One respondent sums up various points made by other respondents when stating that

1. Students will learn to use corpus software, use experimental approaches to learning, get to know frequent collocations and idioms through corpus use.

In terms of perceived advantages of using corpus linguistics in the classroom, the most frequently mentioned item (n = 5) was that it allowed for different and diversified teaching methods with topic-centred approaches, and that all the pupils would be drawn into the classroom work. It was further mentioned that checking vocabulary, phrases and collocations was done much more quickly with the help of corpora than when checking dictionaries.

2.4.2 What Are the Disadvantages of Using Corpora in the Classroom?

In addition to the perceived advantages, various perceived disadvantages were also mentioned. While some respondents had considered corpus-materials easy to prepare, a larger number (n = 7) stated that work with corpora would be work-intensive for the teachers. They argued that it would need a lot of additional explaining (n = 3), it would be difficult to predict how the pupils would manage, it would require in-depth skills by the teachers themselves, and the respondents thought that it was difficult to see how corpora should be integrated into teaching. Further concerns were that pupils might lack mathematical skills and the approaches were not always logical. One respondent answered that

2. It [i.e. corpus use] is much too involved for every-day teaching activities, it is totally unrealistic to really use them [the corpora].

Correspondingly, it was also feared that the pupils would encounter difficulties with corpus use. The most frequently mentioned concern was that corpus work would be too complicated for the pupils (n = 4) and that they would need to learn basic approaches first (n = 4). Also corpus work was considered too time consuming (n = 3). Further concerns were that the size of the corpora might be intimidating for the students, or that the work might be boring. From these concerns that were voiced on behalf of the pupils, the perceived dangers for the lessons follow logically. Time-constraints were the most-frequently mentioned item (n = 5). It was also feared that pupils would be overcharged, and that small technical problems might stall the teaching process, or that the use of corpora would be difficult to integrate into the lessons. One participant sums of his or her concerns as indicated in example 3.

3. Too time demanding and not enough benefit.

Another concern raised by a respondent was that everyone needed internet access. While this would not be a valid concern in many countries, this is indeed an issue in Germany, where digitalization of schools is only in its infancy, a fact which is likely to further discourage trainee teachers in Germany from focusing on corpus use in the classroom.

2.4.3 Trainee Teachers' Quantification of Advantages and Disadvantages of Corpus Use

For the current study, the respondents were asked to evaluate the benefits and difficulties of working with corpora. Table 2.1 gives an overview of how beneficial corpus work is thought to be.

Table 2.1 What do respondents consider positive about working with corpora

Item	Mean value	Maximum points	Minimum points	Std. deviation	Variance
Varied research topics	6.14	10	0	3.27	10.66
Choosing an appropriate corpus	5.71	10	2	2.41	5.82

Item	Mean value	Maximum points	Minimum points	Std. deviation	Variance
Using a corpus interface	4.53	10	1	2.96	8.78
Using the right query syntax	5.06	9	1	2.53	6.39
Obtaining a lot of data	6.71	10	2	2.49	6.2
Obtaining well-structured data	5.32	10	1	2.88	8.31

0 = neutral; 10 = very positive

The data in Table 2.1 show that Obtaining-a-lot-of-data is the highest-rated benefit of corpus work in the classroom. This is closely followed by the possibility to provide varied research topics. The possibilities to choose an appropriate corpus, to obtain well-structured data and the use of query syntax are somewhat less appreciated. The use of corpus interfaces is least appreciated by the respondents. The overall high standard deviation and the high variance of the answers shows that attitudes on all these issues are very divided. The most divided opinions exist on how varied the research topics are that can be provided.

A comparable picture emerges when asking the respondents about the perceived disadvantages of working with corpora in the classroom. The respondents' assessments are shown in Table 2.2.

Table 2.2 What do the respondents consider negative about working with corpora?

Item	Mean value	Maximum points	Minimum points	Std. deviation	Variance
Varied research topics	4.33	10	1	2.57	6.6
Choosing an appropriate corpus	4.9	9	0	2.28	5.19
Using a corpus interface	5.55	10	1	2.81	7.88
Using the right query syntax	5.55	10	2	2.48	6.16
Obtaining too much data	3.79	10	1	2.73	7.43
Obtaining too little data	4.77	10	0	3.04	9.27

0 = neutral, 10 = very negative

With regard to the negative characteristics of corpus work, the results in Table 2.2 are comparable to those in Table 2.1, which displays the perceived advantages of corpus work, but shows slightly lower mean values. In other words, the expressed attitudes to these questions are slightly more positive than negative, though the number of responses is not high enough to draw any robust conclusion. The most negatively rated characteristics of corpus work were the use of corpus interfaces and the use of query syntax. Choosing an appropriate research topic was considered problematic by some respondents. Obtaining too little data was considered a bigger problem than obtaining too much data and some respondents worried about being able to provide appropriate research topics. Variance and standard deviation were highest in the question of the difficulty of using corpus interfaces and on the danger of obtaining too little data. The opinions on these two points thus were the most divided.

2.4.4 Individual Differences in the Assessments

In general terms, we can observe that while some respondents are enthusiastic about the use of corpus-based approaches in the classroom, others are negative about many or all aspects they experienced, and yet others carefully point out both positive and negative experiences with corpus work. One student observes that while corpora can give them more background knowledge, they are unlikely to ever use corpora because of the amount of effort required. The respondent assumes that their pupils would be entirely overcharged. Correspondingly, this respondent does not answer any of the questions on what they found positive about corpus work, but considers it very problematic to use corpus interfaces, use query syntax and is very concerned about getting too much data.

Taking a more balanced view, another respondent states that they are very likely to use corpora in teaching because corpora provide data at one's fingertips, that corpus-based work is easy to prepare, engages all pupils and will motivate the pupils to do their own research. The same respondent cautions, however, that much instruction will be needed on how to use corpora, that it may be complicated for the pupils and that everyone will need a computer. This respondent considers the

use of query syntax the most problematic aspect of using corpora in teaching, but sees high benefits in the variety of corpora that are available, the breadth of answers, and the amounts of data obtainable, which can then be used to prepare varied teaching materials.

At the most positive end of the scale, another participant explains that they value the comprehensive knowledge they can obtain from corpus work. Such work would enable pupils to learn core elements of the English language, and teaching would become more interesting, although the respondent fears that some pupils might find corpus work difficult. While this respondent is conscious of query syntax being complex, they are most concerned about not being able to obtain enough corpus materials without specifying this statement any further. But the width of possible research questions, use of corpus interfaces and possibility of access to data receive very high marks by this respondent.

Interestingly, attitudes were not determined, at least not exclusively, by the amount of exposure to corpus work. Negative attitudes towards corpus use in the classroom were found both amongst students who had participated in entire courses on using corpora for teaching (23 of 65, 35%) and amongst students who had experienced corpus work as a minor part of other classes (20 of 65, 31%) or those, who stated that they were self-taught (2 or 3%). 20 respondents (31%) chose not to answer this question. Thus, while more and more extensive instruction in the use of corpora and corpus linguistics would be highly desirable, it would be too simplistic to believe that simply providing more instruction on how to use corpora will necessarily result in trainee teachers being more likely to resort to corpus-based teaching.

2.5 Discussion

The respondents to the current study mention advantages and disadvantages of using corpus linguistics in the classroom that have been largely also previously mentioned in existing research. The perceived strength of corpus work is that it allows easy access to, and easy analysis of, data. According to the current survey, most of the trainee teachers are comfortable using corpora themselves after having received some introduction to the methodology. They respondents

value the possibilities corpora can offer inside the classroom, namely the opportunity to engage all pupils, the use of authentic teaching materials for empirical research. The respondents consider that corpora can provide teachers and pupils with varied insights into the use of English vocabulary and syntax. Further, additional details and results can be obtained when more advanced query syntax and statistical approaches are used. However, there also is a trade-off that has to be accepted when using corpora. The corpora themselves may be complicated, and so may the search syntax that is used to obtain the results. A number of informants further comment on their insufficient knowledge of query syntax and tools, or fear that their pupils would be overcharged by using these. Further, the amount of obtainable data is a source of concern. The vast amounts of data can lead to very time-consuming analysis.

In the analysis in Sect. 2.4.4 above it has been shown that some respondents viewed the use of in the language classroom as more problematic than others did. In general, respondents commented favourably on the benefits of corpus use, but considerable reservations exist concerning the difficulties associated with bringing it to the classroom. As shown in Sect. 2.4.2 above, next to concerns about high workloads for teachers in preparing corpus work, there are fears of overcharging students with too much data and with complicated query tools and syntax.

In cases where teachers fear that their pupils will be overcharged by using corpus technology, a clear option to overcome this problem, as already identified by Mukherjee (2004), is to employ indirect corpus methodologies. Teachers themselves can cull concordances from corpora and thus provide their pupils with authentic data which can be adapted to the levels of the students (e.g. Bennett 2010). Such indirect methods still provide many of the positive effects of corpus resources while avoiding the perceived pitfalls. The benefits of such indirect corpus use are also shown by various contributions to this volume (see esp. Harrington, this volume and Schneider 2022, this volume). Indirect corpus approaches can also alleviate problems that have been identified by some of the respondents concerning the availability of technical resources. Some countries seem well prepared for digital teaching, as indicated by Leńko-Szymańska (2014), who considers the

lack of hardware and computing skills to have become less acute. Nevertheless, other countries remain slow to reach the digital age in education contexts, with Germany being a case in point, as computers and internet still remain a desiderate in many schools. In addition to the problem of availability of the necessary hardware and software, the willingness of some, possibly but not necessarily older, teachers to use digital resources has also been questioned (Himmelrath 2019). However, the (young) respondents to the current survey show high willingness to use digital media and to overcome the lack of new media usage in educational contexts. Digitalization and the availability of digital resources must be pressed for by all stakeholders in education to allow pupils and student teachers alike to be able to develop a digital habitus that will be necessary for the digital age.

The additional workload involved in preparing for the use of corpus linguistic approaches in the classroom cannot be avoided by indirect corpus use. If the teachers are to prepare concordances culled from corpora to illustrate specific topics for their pupils, then additional work will be needed for this compared to simply following a textbook or a related resource book. Here, however, universities and teacher training institutions can and should also provide help. Topic areas that constitute particular difficulties in language teaching and acquisition have been identified in various language acquisition contexts (see, for example, Le Foll 2022, this volume), as have gaps in teaching materials. In teacher training, the preparation of such targeted teaching materials could and should be encouraged in order to provide the trainee teachers with relevant precompiled concordances to support their teaching. Alternatively, teacher training students could be taught to compile their own specific materials based on target language corpora. The latter approach would obviously have the advantage that the trainee teachers would obtain the wherewithal to conduct such material preparation themselves in the future.

As corpora offer extensive amounts of data, the teachers, too, would face the problem of choosing the materials or concordance lines which should be presented to their pupils. However, as pointed out by a number of respondents to this study, the added benefit of using corpora is that the pupils can be provided with authentic data. As shown by Le Foll (this volume), if teachers solely rely on textbooks, certain

structures may not be acquired successfully or sufficiently by the pupils and for this the additional use of corpus data (c.f. Le Foll, this volume, Harrington 2022, this volume and Curado Fuentes 2022, this volume)—be they from direct or indirect corpus use—is beneficial.

In order to benefit most from corpus linguistics in classroom teaching, teacher training students on the one hand would benefit from obtaining more insights into the use of corpus search syntax. On the other hand, student teachers should also be provided with easily usable and attractive corpus tools (cf. Bennett 2010; Kavanagh 2021), such as for example Compleat Lextutor (Cobb 2021). A strong focus could be placed on how to manage corpus queries so that these can be carried out straightforwardly with the use of easy search syntax, and can provide clear and easy answers.

A difficulty which has repeatedly been mentioned by the respondents to this survey, the challenge and complexity of corpus tasks for pupils, can be alleviated, as mentioned by Braun (2007: 324), by allowing pupils extra time to get used to and to contextualize and construct knowledge. This is necessary, as Braun notes, because pupils may not yet be able to generalize and contextualize sufficiently from what they see in the corpus data.

As this survey shows, the teacher training students who have responded to this survey are largely both willing and ready in principle to be that new generation of teachers which, according to Braun (2007: 308), is needed to bring corpus linguistics into the classroom. The implications of the results of the current study are that, in order to encourage teachers to use corpus linguistic resources in the classroom, teacher training students would benefit from being provided with corpus tasks that are directly transferable to classroom use, with access options for freely available (online) corpora. A particular emphasis should be put on providing examples of basic and simple query tools which will be readily usable in the classroom setting by interested teachers. These simple steps can help to overcome the teachers' fear of classroom and time management issues that are otherwise to be expected when using corpora in classroom teaching.

2.6 Conclusion

Setting out to describe the perceptions of trainee teachers concerning corpus use in the classroom, this study finds that corpora are considered a valuable resource by many of the trainee teachers who participated in this study. Especially valued are the possibility of providing authentic language data, the chance to use more varied teaching methods and to enable pupils to carry out their own research projects in the classroom. However, much criticism was also levied at the perceived complexity of corpus use, the difficulty of tools needed to query the corpora and—to a certain extent—the availability of technical resources in the classrooms. It is argued here that these latter issues can be overcome by encouraging trainee teachers not only to use the corpora themselves in the classroom, but also by employing indirect corpus query methods, such as providing pupils with teacher-prepared concordances, which will allow the pupils to work on the basis of data that have been adapted for their levels and needs.

However, the current study still suffers from various constraints. On the one hand, this certainly is the small number of respondents to the study. Here, larger cohorts of, especially, trainee teachers should be surveyed in order to obtain more robust results concerning their requirements. Further, this survey is restricted to informants from two universities in two German speaking countries, Germany and Switzerland. While the classroom situations that teachers are likely to experience in these two countries will differ, the use of similar studies in further countries would help to identify other country—or language specific needs that teachers and trainee teachers are likely to experience in their own contexts. Nevertheless, this small-scale study has already identified perceived strengths and weaknesses of the use of corpus-based approaches in the classroom and has suggested some ways of dealing with these.

References

Bennett, Gena R. 2010. *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. Michigan: University of Michigan Press.
[Crossref]

Braun, Sabine. 2007. Integrating Corpus Work into Secondary Education: From Data-Driven Learning to Needs-Driven Corpora. *ReCALL* 19 (3): 307–328.

[Crossref]

Callies, Marcus. 2019. Integrating Corpus Literacy into Language Teacher Education. The Case of Learner Corpora. In *Learner Corpora and Language Teaching*, ed. Sandra Götz and Joybrato Mukherjee, 245–263. Amsterdam: Benjamins.

Cobb, Tom. 2021. *Compleat Lextutor Tutor.* Lextutor 2021. Accessed 1 March 2022. https://www.lextutor.ca/.

Curado Fuentes, Alejandro. 2022. Corpus Affordances in Foreign Language Reading Comprehension. In *Demystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington and Patricia Ronan. Palgrave Macmillan.

Farr, Fíona. 2008. Evaluating the Use of Corpus-based Instruction in a Language Teacher Education Context: Perspectives from the Users. *Language Awareness* 17 (1): 25–43.
[Crossref]

Harrington, Kieran. 2022. Culture in English Language Teaching: Let the Language Do the Talking. In *Demystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington and Patricia Ronan. Palgrave Macmillan.

Himmelrath, Armin. 2019. Lehrer fühlen sich mit digitalen Medien im Stich gelassen. Accessed 1 March 2022. http://www.spiegel.de.

Kavanagh, Barry. 2021. Bridging the Gap from the Other Side: How Corpora Are Used by English Teachers in Norwegian Schools. *Nordic Journal of English Studies* 20 (1): 1–35.

Le Foll, Elen. 2022. Causatives in School EFL Textbooks. In Harrington, Kieran and Patricia Ronan (Eds.).

Leech, Geoffrey. 1997. Teaching and Language Corpora: A Convergence. In *Teaching and Language Corpora*, ed. Anne Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles, 1–23. London: Longman.

Leńko-Szymańska, Agnieszka. 2014. Is This Enough? A Qualitative Evaluation of the Effectiveness of a Teacher-Training Course on the Use of Corpora in Language Education. *ReCALL* 26 (2): 1–19. https://doi.org/10.1017/S095834401400010X.
[Crossref]

McEnery, Tony, and Richard Xiao. 2010. What Corpora Can Offer in Language Teaching and Learning. In *Handbook of Research in Second Language Teaching and Learning*, ed. Eli Hinkel, vol. 2, 364–380. London & New York: Routledge.

Mukherjee, Joybrato. 2004. Bridging the Gap between Applied Corpus Linguistics and the Reality of English Language Teaching in Germany. In *Applied Corpus Linguistics. A Multidimensional Perspective Language and Computers*, ed. Ulla Connor and Thomas A. Upton, vol. 52, 239–250. Amsterdam: Rodopi.

O'Keeffe, Anne, Michael McCarthy, and Ronald Carter. 2007. *From Corpus to Classroom*. Cambridge: Cambridge University Press.
[Crossref]

Reppen, Randi. 2010. *Using Corpora in the Language Classroom*. Cambridge: Cambridge University Press.

Schneider, Gerold. 2022. Detecting and Analysing Learner Difficulties Using a Learner Corpus without Error Tagging. In *Demystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington and Patricia Ronan. Palgrave Macmillan.

Zareva, Alla. 2017. Incorporating Corpus Literacy Skills into TESOL Teacher Training. *ELT Journal* 71 (1): 69–79. https://doi.org/10.1093/elt/ccw045.
[Crossref]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_3

3. A Flexible Framework for Integrating Data-Driven Learning

Jane Templeton[1] ✉ and Ivor Timmis[2] ✉
(1) University of Leeds, Leeds, UK
(2) Leeds Beckett University, Leeds, UK

✉ **Jane Templeton**
Email: j.templeton@leeds.ac.uk

✉ **Ivor Timmis (Corresponding author)**
Email: I.Timmis@leedsbeckett.ac.uk

Keywords Phraseology – Framework – Teacher development – DDL – Autonomy – Writing skills

3.1 Introduction

Data-driven learning [DDL] has been around in ELT at least since 1991, when Johns (1991) coined the term for a technique which brings learners into contact with corpus data. The contact can be either direct, when learners access the corpus data themselves, or indirect, when teachers or materials writers mediate between the corpus data and the learners. Whether the contact with the corpus data is direct or indirect, the main principle of DDL is the same: it encourages learners to make discoveries, lexical, grammatical, or lexico-grammatical, from that data. Since its inception, DDL has acquired a small cohort of keen adherents, a supporting literature, and a body of generally favourable empirical evidence. The empirical evidence suggests that there are concrete learning gains to be had, and that, when teachers and learners try DDL,

they are largely positive towards it. We review this evidence below, before confronting the puzzle that, despite the promising empirical picture in terms of both learning gains and attitudes, application of DDL is not that common and it has found it difficult to escape 'minority sport' status. We then discuss the reasons why uptake of DDL may have remained low before suggesting a flexible framework for integrating DDL which may make DDL more accessible for both teachers and learners.

3.2 The Case for DDL

3.2.1 The Theoretical Rationale for DDL

The theoretical rationale for DDL, which is now quite well established in the literature, has both learning and language perspectives. From a learning perspective, the practice of DDL is consistent with the principles of inductive and discovery learning (Smart 2014; Gilquin and Granger 2010), which hold that learning will be deeper and more durable if the learner has invested effort in achieving it: in DDL, this effort takes place when learners try to adduce generalisations about a language feature from the data presented.

The potential for autonomous learning is another important facet of the learning rationale for DDL (Morgoun et al. 2020). While the degree of autonomy in DDL activities can be carefully regulated by the teacher, the scope for learners to generate and investigate their own language queries encourages learners to develop autonomous learning skills. While not all learners will reach the same level of autonomy, and while some educational contexts do not encourage it, autonomy may be scaffolded by the use of questions which direct the learners towards the answer, by limiting the amount of data, or by selecting data which exemplifies the feature very clearly.

From a language point of view, the main argument in favour of DDL is that corpora present naturally occurring data. If learners are exposed to lexical or grammatical features in their natural habitat, they will ipso facto experience lexical items alongside their frequent collocates or embedded in the lexical phrases in which they typically occur. In the case of grammatical structures, learners will see them with the lexis with which they typically co-occur. How far learners need to be directed

to notice these features of the habitat is an open question, but DDL opens up important opportunities for learners to develop the valuable skills of noticing. We show below how such opportunities can be provided and exploited.

3.2.2 Reservations about DDL

Of course, in language teaching, a complex field full of impassioned debate, no technique, method or approach gets a free pass. DDL is no exception. In terms of learning theory, it can be argued that DDL places a heavy premium on inductive learning, which does not suit all learners. However, as Boulton (2009) has pointed out, detecting patterns from data, and extrapolating from those patterns is a natural human faculty. We would also note that learning styles are not fixed and that no teaching technique is universally suitable, so DDL cannot be disqualified on those grounds alone. DDL certainly requires some modification of teacher role—from teacher-fronted to learner-centred, from directive to facilitative—which is easier to achieve in some contexts than others. There is, however, no need for an overnight volte-face. Indeed, we argue below for an incremental approach to DDL which is integrated into common classroom activities. DDL is not a methodology, but a technique which can be used alongside any methodology which prevails in a particular context.

From the point of view of technical expertise, DDL certainly presents a challenge for both teachers and learners, but in a world where we are constantly using electronic devices of one kind or another, this should not be an insurmountable challenge, particularly for digital natives brought up in the electronic world. At times, however, it may mean teachers adopting the role of co-learner, a role not all teachers will be naturally comfortable with.

DDL can be portrayed as a solitary, non-communicative activity, with a lone student facing computerised output, but it is also possible, as Gavioli and Aston (2001) have pointed out, for learners to carry out certain DDL tasks in groups. Learners, they argue, may notice different things from the same concordance lines and be involved in pooling their knowledge to arrive at conclusions from the data. In this sense, it shares some of the characteristics of task-based learning, and, indeed, of problem-based learning. In this respect, we can draw on the

rationale for problem-based learning, with a language query constituting the problem on which the group works. Although in problem-based learning the problem is usually conceived as a complex issue, if we construe making sense of the data as the problem, then we can see some overlap between group work in DDL and PBL as described by Mishan (2011: 253):

> Problem-based learning is rooted in constructivist philosophy, which holds that knowledge is actively constructed within the mind of the learner and influenced by his/her interactions with peers and with the environment. Furthermore, constructivism holds that learning is spurred by 'the problematic' (i.e., cognitive conflict). In PBL, cognitive conflict is 'concretised', in that a real problem is used to trigger the learning process.

The use of naturally occurring data brings with it its own problems, particularly when presented in the peculiar form of truncated, semi-contextualised concordance lines. In some cases, neither the cultural context, nor the co-text, in which the target feature is embedded, may be clear to the learner. It is possible, however, for the teacher to edit out the more inaccessible concordance lines or, as learners become more autonomous, for them to ignore the lines they do not understand. At a broader level, we would argue that, in an information-rich world, distinguishing what is useful and relevant is a valuable transferable skill.

3.2.3 The Empirical Case for DDL

There is, then, we would argue, a robust theoretical rationale for DDL which stands up to critical scrutiny. This, as we shall see, is matched by a largely positive empirical picture in terms of both attitudes of learners and teachers, and of the learning gains to be had from DDL activities. Research into learners' and teachers' attitudes to DDL reveals a number of common themes, both positive and negative, which we summarise briefly here as they are well-covered in the literature:

Learners can find this mode of learning enjoyable and useful. Götz and Mukherjee (2006: 49), working with university students in Germany, note that overall, learners found DDL "interesting, useful and

fun", while Ädel (2010) reported that the students enjoyed inductive learning and the novelty of using computers in the classroom.

DDL can be particularly useful for lexical learning. Yao (2019: 35) found that learners of Spanish appreciated the KWIC format which 'provided a rich context for the target vocabulary'. The learners in the study by Asik et al. (2016) also highlighted the value of DDL for lexical development (see also Selivan, this volume).

Learners can appreciate the scope for autonomy. The scope for autonomy was highlighted as a positive feature of DDL by Götz and Mukherjee (2006) and Alsolami and Alharbi (2020).

Learners need time to get to grips with DDL. Ädel (2010) found that a serious limitation of her research was that it was an isolated experiment with no follow-up work, while Götz and Mukherjee (2006) also noted that their learners would have preferred a longer introduction.

Learners can find DDL an unattractive way to learn. Chambers (2010) has pointed out that comments about the time-consuming and sometimes tedious and laborious nature of the tasks are quite common.

Individual learners may have mixed attitudes towards DDL. Kennedy and Miceli (2001: 80) make the interesting observation that individual learners can also have mixed attitudes, specifically that they can find DDL "helpful and confidence boosting, but sometimes also discouraging, time-consuming and frustrating."

When they are introduced to DDL, many teachers respond positively. When Lin (2016) introduced early career teachers in Taiwan to DDL she reported that they responded positively and considered that it improved their learners' attitudes to learning grammar. Breyer (2009) also found trainee teachers to be positive about DDL, while Naismith (2016) found them to be very interested in corpus tools, whether they had actually used them or not. Working with graduate teacher trainees in Poland, Leńko-Szymańska (2014) reported a generally favourable response.

Teachers' attitudes are influenced by a number of variables. A study carried out by Chen et al. (2019) in Hong Kong identified the following variables as influencing teachers' attitudes to DDL: prior knowledge of corpora, prior experience in using corpora, motivation for professional

development and teaching experience. Naismith (2016) identified confidence as another key factor in limiting classroom application.

Getting to grips with DDL takes time. In Leńko-Szymańska's (2014) study, teachers reported that they needed more time both to master the mechanics of DDL and to understand the pedagogic rationale, while Lin's (2016) participants experienced technical difficulties in implementing DDL and found that it increased their workload.

In considering the empirical picture in terms of learning gains in DDL, we are fortunate in being able to draw on a number of meta-studies of DDL in action. A meta-analysis of 64 DDL studies carried out by Boulton and Cobb (2017) showed that DDL had a significant effect on learning irrespective of the context, corpus, activity or learner involved in the individual research study. The meta-analysis produced by Lee et al. (2017) collated 29 studies and focused particularly on vocabulary learning: they found "an overall positive medium-sized effect of corpus use on L2 vocabulary learning for both short-term […] and long-term periods" (Lee et al. 2017: 7). Lee et al. identified a number of variables which impacted on learning gains: L2 proficiency; the type of corpus; the nature of the interaction; training in corpus use, and the duration of the experiment. A smaller meta-analysis of 14 studies in a Japanese context was carried out by Mizumoto and Chujo (2015). They report qualified success: while DDL worked well for vocabulary and basic grammar items (see Selivan and Jones, this volume), it seemed to have little effect in terms of overall language proficiency. Luo and Zhou (2017) concentrate in their meta-study on the effect of DDL on learners' writing. They found that DDL had great potential to develop learners' writing skills but qualify their finding by pointing out that some learners preferred traditional reference resources to corpus tools for some purposes (see also Friginal et al., this volume).

We return now to the puzzle outlined in the introduction: if the theoretical and empirical picture is so bright, why does Meunier (2020: 43) speak of "a lack of uptake and sustainable practices in DDL"? We would like to suggest in this chapter that a major factor which limits the application of DDL is a lack of the design principles which mediate between theory and practice, and an inventory of the skills needed to implement DDL. We cannot expect, to borrow Thornbury's (1998)

description of the lexical approach, teachers and learners to embark on 'a journey without maps'. What is needed, we feel, is a sketch map which shows routes, some more scenic, some more direct, to the destination of constructive use of DDL.

3.3 A Way Forward for DDL

Both theory and practice as discussed above seem to suggest that the key benefits of DDL relate to its potential for autonomous language learning. If the ultimate destination of the DDL journey is that learners are able to use corpora for their own language learning or production purposes (or make a principled decision not to), then it follows that learners need to understand the affordances of corpora and be able to exploit them for a range of purposes. We argue that this is a complex skills set in its own right, and should be treated as such. What we propose, therefore, is a DDL syllabus whose core aim is to equip learners with these skills.

In outlining a theoretical rationale for our approach, we can repurpose two theoretical notions we cited above. The principles of discovery learning, which we discussed in relation to learning language in DDL, apply equally to learning the techniques of DDL: the effort invested in learning techniques, sometimes independently of the teacher, will be rewarded with more durable learning of those techniques. The principles of problem-based learning apply to the challenge of finding the right DDL technique or resource as much as they do to the problem of solving a linguistic query.

While the destination of the overall journey is the same, the route may require a shift in focus, i.e. from DDL as a means of learning target language to DDL as a means of developing skills. The aim of a DDL activity thus becomes not to teach relevant language using DDL, but to teach a technique or develop facility with a tool using relevant language. The problem to be solved in each iteration of DDL becomes not merely about language, but also about skills, and comprises two elements:

1. What is the answer to this language query?

2. How could we answer this query about a different language item?

The following learning principles are also consistent with our approach to DDL.

3.3.1 Zone of Proximal Development

We believe that technological and technical development must be incremental and that the teacher, who may only be a step ahead of the learners, plays an important role in scaffolding techniques when s/he judges that learners are ready for them. This reflects the zone of proximal development theory proposed by Vygotsky (1978). The corollary of this is that learners should have experience of interpreting concordance lines before moving on to generating their own concordance lines directly from the corpus.

3.3.2 Motivation

It is important, in our view, that the language used when modelling DDL techniques or resources should be useful or relevant to the learners. We would also argue that DDL techniques introduced at point of need are likely to make a greater impact on the learner. A key principle with respect to motivation is that teachers need to recognise appropriate 'problems' in order to exploit opportunities to present DDL techniques as solutions. From this over-riding principle, a number of entailments can be derived, as outlined below:

- The nature of the language query should determine the corpus tool or function selected.
- New DDL techniques should not be introduced arbitrarily, but with the aim of solving a specific and relevant language query.
- Teachers should be alert to the opportunity to introduce DDL techniques in response to learner errors or an immediate need for an item of language.
- Learners should interact directly with the corpus wherever possible, and as soon as possible in the journey.

3.4 A Flexible Framework for DDL in Practice

In the following section, we present a prospective DDL framework based on these principles, to be integrated into an existing language syllabus. We first summarise the key features and then exemplify how it might be enacted in practice.

3.4.1 Summary of Key Features of the Framework

The aims are simple: learners should understand the affordances of corpora and be able to exploit these affordances for their own purposes. The extent to which these aims are met can be evaluated using typical measures in the literature, i.e., whether learners use the tools independently or make any noticeable positive adjustments to their language having used the tools.

The data from which they will learn comprise an online reference corpus, 'expert' texts, and learners' own writing. Although there are many corpora and corpus tools, the essential purposes for which learners may use them can be divided into two categories: general reference and text-specific reference. Therefore, we suggest that learners should be trained in using tools that allow them to search pre-existing corpora and tools that allow them to search texts of their choice. In some tools these features are combined. The specific choice of tool is not important—what matters is that it works for the teacher, so the teacher can make it work for the students.

The journey comprises three main stages of development, moving from simple to complex in terms of technical skills. The first stage, which we believe is key in terms of scaffolding, introduces learners to the core principles of DDL and the core techniques of inferring meaning from language presented in KWIC format. The key feature of the first stage is that it does not require teachers or learners to interact with a corpus. The second stage introduces learners to the affordances of reference corpora, using a user-friendly online tool. The final stage introduces learners to the affordances of concordancing software. The activities in this framework can be embedded according to the context and the existing skills and language syllabus—essentially, when looking at language about which DDL techniques will yield relevant and useful insights.

The journey is not linear, and one stage does not necessarily end when the next begins—online reference corpora and learner corpora

offer different solutions to different needs. Learners will be ready to begin the next stage when the limitations of the current tools or techniques become apparent, i.e., when they have a need or question that cannot be answered using the resources at hand. This can arise naturally, when learners ask questions about language or techniques, or be contrived by the teacher, in response to learner error or by exploiting opportunities to problematize the need.

In terms of teacher expertise, the teacher should be merely one step ahead of the learners in planned class activities. If a spirit of joint enquiry is cultivated, this does not require comprehensive expertise with the tool in question, as issues can be dealt with by turning the questions over to the learners, or to embarking on a journey of discovery together to find out the answers. For comfort, and to avoid unnecessary demotivation in the initial encounters with each tool, we recommend doing planned searches in advance to ensure they will yield relevant and appropriate results, and being prepared for troubleshooting and questions.

3.4.2 Example DDL Journey in Practice

We believe this method is appropriate for all contexts where teachers and learners have no prior experience of concordancing, where institutional constraints allow, and where learners have access to the internet and digital devices, whether they are learning English for general or specific purposes.

Stage 1: Core Principles and Techniques

The aims of the stage are that:

1. Learners understand the core principles of DDL:

    (a) Language which is frequent in a context is likely to be useful in that context;

    (b) Collocation is important for understanding and expressing meaning accurately;

    (c) Examining language in context enables useful inferences about collocation, colligation and patterning.

2. Learners are able to find key words in a text, use the co-text to identify collocates, and draw relevant conclusions from those collocates.

The pre-reading stage of a reading class is an ideal time to highlight the core principles, using frequency as a basis for brainstorming and predicting content. This can be achieved by methods such as showing a word cloud of the reading text and asking learners to predict the content based on the most frequent words, or using the collocates of a key concept to generate a concept map. For example, in a context where the topic is some kind of problem—let us say climate change, since it is universally relevant and lends itself well to a range of activities and output tasks—the verb collocates of the generic noun *problem* can be used to create a framework for analysing the specific problem in question. The most frequent verb collocates obtained from a word search of *problem* in the Corpus of Contemporary American English (Davies 2008) are: *solve, cause, face, address, fix, resolve, arise***,** and *pose*. Learners could thus consider who faces the problem of climate change, how it can be solved and what causes it, thus forming the basis for predicting the content of the text. The post-reading language focus is an ideal time to introduce manual concordancing. This activity could comprise collecting sentences containing *climate change* for a glossary, or searching the reading text for verb collocates of *climate change* to start building a concept map. Once learners are familiar with these principles and the KWIC format, processing concordance lines should be well within their capabilities (see Curado Fuentes, this volume, for the use of corpora in foreign language reading comprehension).

Stage 2: Reference Corpora

Learners may quickly recognise the limitations of manual concordancing. The problem to be solved might manifest itself in comments on the laborious nature of manual concordancing in longer or multiple texts, or, when working with one text, an item of interest is not represented. The solution would be a way of collecting examples of the item in context very quickly, or having access to a greater range of data. This is the ideal time to introduce an online reference corpus tool.

The aims of this stage are that:

1. Learners understand what the tool can do and how they can use it;

2. Learners are able to:

   (a) Produce concordance lines, which includes searching effectively, i.e. understanding what search terms to use to produce relevant results and filtering those results where necessary;

   (b) Interpret concordance lines, which includes identifying patterns, sentence position, punctuation, collocation, and colligation. Where it is possible to infer meaning from concordance lines, it includes identifying more advanced features of language such as semantic prosody and preference;

   (c) Cope with irrelevant results.

It will be clear that these aims have been achieved when learners notice opportunities to refer to the online corpus, when they are able to check things on the spot when asked to do so, and when they can use the tool autonomously.

As mentioned above, there are many such tools. In choosing one, the key consideration is that it is user-friendly for the teacher. For the purposes of this example, we are using the embedded tool in the COCA corpus (Davies 2008). Like similar tools, there is a host of useful features, and it is important to control focus in the initial encounters in order to avoid getting distracted and thus exhausting learners' motivation or attention capacity. It is also important to control activity. In keeping with the zone of proximal development principle (Vygotsky 1978), the only new technique in the first encounter should be the generating of concordance lines to answer one language query. The simplest way to do this with the COCA tool is to do a word search for a single word and look at its concordances. Table 3.1 below shows two possible initial searches, based on common classroom triggers.

Table 3.1 Possible initial searches on the embedded COCA tool

	Word search -> Concordances	**Word search -> Collocates -> concordances**
Trigger	Question about or error in preposition collocate e.g. learner says ***due of*** when expressing cause and effect	Question about or error in noun-noun collocation e.g. if we can say ***climate change***, can we say ***change _ climate***?
Problem expressed as questions	(a) What is the correct preposition to use here? (b) How can we find out which preposition to use with other words?	(a) How exactly do ***climate*** and ***change*** collocate? (b) How can we find out how two nouns collocate?
Procedure	1. Word search ***due*** 2. Check clusters, see ***due to*** but not ***due of*** 3. Check concordances, see ***due to*** is used in every case 4. Think of another collocation we could check in this way, and repeat the process	1. Word search ***climate*** 2. Check noun collocates 3. Click on ***change*** 4. Check concordances, see ***climate change*** almost always used, and where ***change _ climate*** is used it doesn't mean climate change 5. Choose another relevant noun collocate and repeat the process

This is all that is needed in any iteration of DDL, which is one of the key benefits of this approach in terms of classroom application: it does not need to take a significant amount of time, and does not need to replace existing language activities. It can be used at any point in the class where relevant opportunities arise.

Follow-Up

In keeping with the principle of scaffolding, practice activities should require learners to use the same technique, either with a different aspect of the same language item, or the same aspect of a different item. Thereafter, teachers should exploit opportunities to introduce and train learners in using the various features of the tool. As discussed in our principles, this involves being alert for problems to which DDL is a useful solution. Learners and teachers might express problems or ask questions directly, or learners might make errors which the language reference resource can help correct. Alternatively, teachers can contrive the need to consult one by asking learners to produce language that is not in their text.

The COCA tool can help answer a wide range of linguistic and technical questions. It has recently been significantly upgraded, and as we are still exploring the updated options we do not claim to provide an exhaustive list. See Davies (2020) for a comprehensive guide to the features and affordances of the corpus. However, we list here some of the broad types of question that might arise naturally or be gently contrived:

- How do we generate concordances for a multi-word phrase?
- What is the difference in meaning between these two items?
- What word should we use here?
- Can we use this word with this word?
- How is this item used in a particular genre or domain?
- How useful is this word for me to learn?
- How does this item pattern in a sentence?
- What other words can go in this gap?

It is important that follow-up activities are scaffolded. While independent work should be encouraged, the teacher should be on the lookout for classroom opportunities to test learners' developing skills and understanding, for example by directing learners to use the tool to answer questions.

Stage 3: Learner Corpora

Sooner or later, learners are likely to recognise the limitations of reference corpora, such as the low representation of some technical terms, the inability to search in specific texts, and the mass of irrelevant data. The solution is concordancing software, or web tools into which learners can input texts for analysis. For the purposes of this example, we are using AntConc (Anthony 2020).

The aims of this stage are that learners understand the affordances of AntConc and are able to exploit them to investigate language use in a corpus of their own making, and to improve their own writing (see Gilquin and Granger (2010) on the value and uses of learner corpora and Walsh (2010) on using learners' own writing as a corpus). As above, it will be clear when the learners have achieved these when they are able to notice opportunities to check things in their corpus, be able

to check things when asked to do so on the spot, and be able to use their own corpus autonomously.

In the first classroom encounter, for classroom management purposes, we recommend that teachers demonstrate the tool and encourage learners to download the software and create their corpus outside class. In keeping with the zone of proximal development principle (Vygotsky 1978), the only novelty should be the tool—the learners should already be familiar with the function and techniques). For example, the teacher could ask how *climate change* is used in verb phrases in a specific, long text. To find all the instances manually, even using the find function, would be time-consuming and laborious. The Collocates tab of AntConc can then be presented as a solution. If learners are familiar with the Collocates feature of COCA, they will already have the necessary technical skills and only need to apply them to a new interface.

Follow-Up

A logical next step is to add more texts to the corpus, enabling language use to be compared across them, which is useful for identifying features of 'expert' writing in the genre and for helping learners correct errors in their writing. From here, the learners' context, needs and technical proficiency should determine the sequence of development. The only requirements are that learning is scaffolded and learners are introduced to the other features of the tool incrementally and at point of need. Inspiration can be found in published DDL studies which include details of the sequences used to develop skills with particular tools. The sequences described in Charles (2018), for example, who trained EAP learners to use AntConc to improve draft chapters of their PhD thesis, could be adapted to any context where learners would benefit from the ability to analyse expert texts in order to write more effectively in that genre.

3.5 Towards Systematic Opportunism in DDL

It could be argued that what we propose is experiential learning par excellence (Kolb 1984) as learners go through the prototypical experiential learning: concrete experience of the challenge of using

corpus tools; reflection on that experience; generalisation from that experience, followed by active experimentation. We have chosen to describe our approach to DDL as systematic opportunism for the following reasons: it is systematic in that it is informed by language teaching principles, as outlined below; it is opportunistic in that particular skills and techniques are implemented when circumstances seem favourable for their use, rather than by following a fixed a priori syllabus: the DDL syllabus is embedded in the overall syllabus rather than being seen as a discrete thing. Providing a rationale for systematic opportunism in DDL does not require, then, a new set of principles; we have simply adduced relevant and well-established learning principles to underline its credentials, and to show the specific implications of these principles for DDL practice.

3.6 Conclusion

In conclusion, we have sought to show that an effective way to introduce DDL can be, through systematic opportunism, to integrate it incrementally with other typical classroom activities, especially when dealing with new topics and/or texts. We do not present this approach to introducing DDL as superior to dealing with it on pre-service courses or on in-service teacher development courses. However, to the best of our knowledge, DDL is not a staple on such courses, and where it is, Naismith's (2016) study suggests that even teachers who have been introduced to DDL and are favourably disposed to it can be reluctant to take the plunge in the classroom. A flexible framework for implementation might, then, be very useful in bridging the gap between courses and practice and might be the only option for teachers without access to specific training courses. Such a framework, we hope, will help teachers to apply DDL selectively when it is appropriate for a particular learning objective. We also argue, given that we cannot deal with all the language that learners might conceivably need, that it is important to equip them with the skills and tools to find out for themselves about language which is relevant to them, i.e. to develop the generic, transferable skills of DDL. From the teacher's point of view, a willingness to temporarily relinquish the role of expert, and adopt the role of fellow traveller, is likely to be a useful asset for those taking their

first steps in DDL. For DDL to be sustainable (Meunier 2020), we contend, its practices need to be durable, versatile and effective. One way of achieving that, we have argued, is, through systematic opportunism, to adopt a flexible framework to guide the principled and judicious application of DDL in the classroom. We have stressed that we are offering just one possible way forward for DDL. There are, of course, others: teacher education and teacher development courses have an important role to play by including at least an introduction to DDL principles and practice in their programmes. If teachers realise the potential of DDL, then that potential can be realised. There is a role too for materials writers, who could contribute by including tasks which use open access corpus resources to provide practice on the language points included in the materials. Finally, resource books which offer examples of principled and practical DDL activities with clear aims and rationale (e.g. Viana 2022) would be of great value in bridging the gap between theory and practice. Should these desiderata be achieved, the focus of future research can shift from whether DDL is effective to the conditions in which DDL is most effective, which is a small shift but important in terms of ultimately increasing classroom uptake.

References

Ädel, Annelie. 2010. Using Corpora to Teach Academic Writing: Challenges for the Direct Approach. In *Corpus-Based Approaches to English Language Teaching*, ed. Ma Carmen Campoy-Cubillo, Begoña Belles-Fortuño, and Lluïsa Gea Valor, 39–55. London and New York: Continuum.

Alsolami, Turki, and Assrar Alharbi. 2020. Saudi EFL Learners' Perceptions of The Use of Corpora in Academic Writing Teaching. *Studies in English Language Teaching* 8 (4): 94–111. https://doi.org/10.22158/selt.v8n4p94.
[Crossref]

Anthony, Laurence. 2020. *AntConc (Version 3.5.9) [Computer Software]*. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software.

Asik, Asuman, Arzu Sarlanoglu Vural, and Kadriye Dilek Akpinar. 2016. Lexical Awareness and Development through Data Driven Learning: Attitudes and Beliefs of EFL Learners. *Journal of Education and Training Studies* 4 (3): 87–96.

Boulton, Alex. 2009. Testing the Limits of Data-Driven Learning: Language Proficiency and Training. *ReCALL* 21 (1): 37–54.
[Crossref]

Boulton, Alex, and Tom Cobb. 2017. Corpus Use in Language Learning: A Meta-Analysis. *Language Learning* 67 (2): 348–393. https://doi.org/10.1111/lang.12224.
[Crossref]

Breyer, Yvonne. 2009. Learning and Teaching with Corpora: Reflections by Student Teachers. *Computer Assisted Language Learning* 22 (2): 153–172. https://doi.org/10.1080/09588220902778328.
[Crossref]

Chambers, Angela. 2010. What Is Data-Driven Learning? In *The Routledge Handbook of Corpus Linguistics*, ed. Anne O'Keeffe and M. Michael McCarthy, 345–358. Abingdon: Routledge.
[Crossref]

Charles, Maggie. 2018. Corpus-Assisted Editing for Doctoral Students: More Than Just Concordancing. *Journal of English for Academic Purposes* 36: 15–25. https://doi.org/10.1016/j.jeap.2018.08.003.
[Crossref]

Chen, Meilin, John Flowerdew, and Laurence Anthony. 2019. Introducing In-Service English Language Teachers to Data-Driven Learning for Academic Writing. *System* 87. https://doi.org/10.1016/j.system.2019.102148.

Davies, Mark. 2008–. *The Corpus of Contemporary American English (COCA).* https://www.english-corpora.org/coca/.

———. 2020. English-Corpora.org: A Guided Tour. *English Corpora.* Accessed 1 November 2021. https://www.english-corpora.org/pdf/english-corpora.pdf.

Gavioli, Laura, and Guy Aston. 2001. Enriching Reality: Language Corpora in Language Pedagogy. *ELT Journal* 55 (3): 238–246.
[Crossref]

Gilquin, Gaëtanelle, and Sylviane Granger. 2010. How Can Data-Driven Learning Be Used in Language Teaching? In *The Routledge Handbook of Corpus Linguistics*, ed. Anne O'Keeffe and Michael J. McCarthy, 359–371. Routledge. https://doi.org/10.4324/9780203856949.
[Crossref]

Götz, Sandra, and Joybrato Mukherjee. 2006. Evaluation of Data-Driven Learning in University Teaching: A Project Report. In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, ed. Sabine Braun, Kurt Kohn, and Joybrato Mukherjee. Peter Lang.

Johns, Tim. 1991. Should You Be Persuaded: Two Samples of Data-Driven Learning Materials. In *Classroom Concordancing*, ed. Tim Johns and Philip Ling, 1–16. Centre for English Language Studies, University of Birmingham.

Kennedy, Clarie, and Tiziana Miceli. 2001. An Evaluation of Intermediate Students' Approach to Corpus Investigation. *Language Learning and Technology* 5 (3): 77–90.

Kolb, David. 1984. *Experiental Learning: Experience as The Source of Learning and Development*. Englewood Cliffs, NJ: Prentice-Hall.

Lee, Hansol, Mark Warschauer, and Jang Ho Lee. 2017. The Effects of Corpus Use on Second Language Vocabulary: A Multilevel Meta-analysis. *Applied Linguistics* 40 (5): 721–753. https://doi.org/10.1093/APPLIN/AMY012.
[Crossref]

Leńko-Szymańska, Agnieszka. 2014. Is This Enough? A Qualitative Evaluation of the Effectiveness of a Teacher-Training Course on the Use of Corpora in Language Education. *ReCALL* 18 (1): 83–104. https://doi.org/10.1017/S095834401400010X.
[Crossref]

Lin, Ming Huei. 2016. Effects of Corpus-Aided Language Learning in the EFL Grammar Classroom: A Case Study of Students' Learning Attitudes and Teachers' Perceptions in Taiwan. *TESOL Quarterly* 50 (4): 871–893. https://doi.org/10.1002/tesq.250.
[Crossref]

Luo, Qinqin, and Jie Zhou. 2017. Data-driven Learning in Second Language Writing Class: A Survey of Empirical Studies. *iJET* 12 (3): 182–196. https://doi.org/10.3991/ijetv12/2i03.6523.
[Crossref]

Meunier, Fanny. 2020. Data-Driven Learning: From Classroom Scaffolding to Sustainable Practices. *ELLE* 8 (2): 423–434. https://doi.org/10.30687/ELLE/2280-6792/2019/02/010.
[Crossref]

Mishan, Freda. 2011. Whose Problem Is It Anyway? Problem-Based Learning in Language Teacher Development. *Innovation in Language Learning and Teaching* 5 (3): 253–272.
[Crossref]

Mizumoto, Atushi, and Kiyomi Chujo. 2015. A Meta-analysis of Data-driven Learning Approach in the Japanese EFL Classroom. *English Corpus Studies* 22: 1–18.

Morgoun, Natalia, Nataliya Mekeko, Margarita Kozhevnikova, and Nadezhda Arupova. 2020. *Enhancing Learner Autonomy with DDL: A Case Study of Learners Perspective*. Paper presented at the 4th International Conference on Education and Multimedia Technology. https://doi.org/10.1145/3416797.3416840.

Naismith, Ben. 2016. Integrating Corpus Tools on Intensive CELTA Courses. *ELT Journal* 71 (3): 273–283. https://doi.org/10.1093/elt/ccw076.
[Crossref]

Smart, Jonathan. 2014. The Role of Guided Induction in Paper-Based Data-Driven Learning. *ReCALL* 26 (2): 184–201.
[Crossref]

Thornbury, S. 1998. The Lexical Approach: A Journey without Maps? *Modern English Teacher* 7 (4): 7–13.

Viana, Vander. 2022. *Teaching English with Corpora.* Abingdon: Routledge.

Vygotsky, Lev. 1978. *Mind and Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Walsh, Steve. 2010. What Features of Spoken and Written Corpora Can Be Used in Creating Language Teaching Materials and Syllabuses? In *The Routledge Handbook of Corpus Linguistics*, ed. Anne O'Keeffe and Michael J. McCarthy, 333–344. Routledge.
[Crossref]

Yao, Gang. 2019. Vocabulary Learning through Data-Driven Learning in the Context of Spanish as a Foreign Language. *Research in Corpus Linguistics* 7: 18–46. https://doi.org/10.32714/ricl.07.02.
[Crossref]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_4

4. Speaking and Listening: Two Sides of the Same Coin

Michael McCarthy[1] ✉ and Jeanne McCarten[2]
(1) Nottingham University, Nottingham, UK
(2) Cambridge, UK

✉ **Michael McCarthy**
Email: mactoft@cantab.net

Keywords Speaking skills – Listening skills – Conversation – Corpus – Naturalness – Response – Turn taking

4.1 Introduction

This chapter takes as its starting point the assumption that naturalness in conversation is a positive aspiration in second and foreign language learning. It is important to point out from the start that naturalness is not to be equated with native-like performance in the target language (Warren 2006: 11–14). We hold it to be a normal reflex of human social behaviour rather than the behaviour of any individual linguistic community. We do, however, accept that there may be nuances in different realisations of linguistic behaviour which require caveats to guard against negative evaluation and intercultural stereotyping (Stubbe 1998). We define naturalness as behaviour aligned with expected social norms, and its opposite, unnaturalness, with behaviour that lacks or falls short of those norms. Our particular concern here is with the typical behaviour of those in the role of 'listener' in social conversation, albeit the listener's role and that of the speaker are bound together in interdependent activity. The focus on listeners comes

about because of the relative historical imbalance towards what speakers do in the 'speaking skills' components of language syllabuses. The focus on the interdependence of speaker and listener behaviour aims also to counteract the tendency to see oral production as the product of the single mind and the single actor. Corpus evidence suggests otherwise, as we shall see.

Our aim is to propose a closer relationship between the teaching of listening and the teaching of speaking in English as a second or foreign language. Our target student audience is learners of CEFR B1 level upwards, though we shall argue that practice in exercising good listener skills can begin earlier, given the relatively uncomplicated linguistic repertoire available to listeners to behave actively and naturally in social conversation. The present authors, as teachers and users of materials, had their training and careers rooted in the traditional notion of 'listening comprehension', that is to say, learners engaging in listening activities in the classroom in order to demonstrate their understanding of audio or audio-visual input in the target language. Typically, occupying a separate strand of the syllabus, came 'speaking skills', based on the skills of constructing communicatively efficient utterances in the L2 on specified topics, or else the 'conversation class', where freer practice was the aim. Separating the listening skills and the speaking skills in this way flies in the face of corpus evidence and, for that reason, we embarked on writing projects which aimed to bring the insights of corpus research to materials which would be more closely aligned with the unitary nature of speaking and listening and which would aspire to achieving naturalness. We report on the methodological implications and the types of activities our research led us to and in what ways we believe them to be faithful to the characteristics of natural social conversation.

4.2 Small Words, Big Meanings: Listener Responses in Conversation

O'Keeffe et al. (2007: Chapter 7) noted that certain lexical items in English were many times more common in conversational language than in formal writing. O'Keeffe et al. took as their example the

occurrences of *right* per million words across spoken and written datasets. They concluded that the disparity in frequency was because *right* performed basic discourse-marking functions in speaking and, most significantly for the present chapter, as markers of listener response, as in Extract 1, where *right* serves to acknowledge the incoming talk:

Extract 1.
[discussing the work of a poet]
<S510> no what's his set form tend to be then?
<S509> oh he has like different forms for different poems
<S510> **oh I see**
<S509> but they tend to like er like a lot of poets will like break form for effect
<S510> **right**
<S509> whereas he is a more traditional poet
<S510> **right right yeah** (Spoken BNC2014 S7VD)

Speaker 510 additionally acknowledges the information supplied by speaker 509 with *oh I see* and *yeah*. These banal, small tokens, extremely frequent in everyday conversation, raise a number of intriguing questions. Not least of these is why anyone who is in the role of 'listener' or 'recipient' in face-to-face encounters needs to say anything at all. Why not just give a series of head-nods, cup a hand around one ear, smile or grimace, or wait till the whole package of information is delivered and then just say *okay* or something similar? Furthermore, the grammatical status of such linguistic expressions when used to acknowledge incoming talk poses questions about word-classes. Is *see* a transitive verb in *oh I see*? If so, why does it have no object in its clause in the extract? Is *right* an adjective, an adverb, or is it *sui generis*, a one-off item? Do other items operate in the same way? If so, how shall we account for them? All these questions are raised by the behaviour of listener-recipients but can only effectively be tackled analytically in conjunction with the behaviour of the speaker-sender, and vice-versa. Unless we can see conversation as a jointly produced artefact, the single coin with two sides, it will be difficult to account for the phenomena associated with listener behaviour, or *listenership* as we shall refer to it hereafter.

4.3 Listeners in the Literature

Linguists have long observed and commented on the occurrence of listener responses which manifest in short utterances and non-verbal actions (e.g., head nods, laughter). These include studies by Fries (1952), Yngve (1970), Duncan (1974), Oreström (1983), Tottie (1991), Drummond and Hopper (1993a, 1993b), Gardner (1997, 1998, 2001), Norrick (2012) and Tolins and Fox Tree (2014). Key to the role of 'listener' is that they are participants who do not bid to take over the floor, i.e., to become 'current speaker'. Tottie (1991: 225) refers to actions which 'grease the wheels of the conversation but constitute no claim to take over the turn', a neat metaphor for the positive contribution of, on the face of it, minor conversational increments: grease on the wheels enables the machine to run more smoothly.

However, where one role (listener) crosses over to the other (speaker) is by no means straightforward, insomuch as the actions of listeners can range all the way from non-verbal behaviour (head-nods, hand positions, eye movements, etc.), through vocal but non-verbal responses(*uh-huh*, *erm*, *aha*, etc.), through single verbal items (*yeah, right, no, well*), or more complex items such as *That's true*, *How nice*, and pseudo-interrogative items such as *Oh really? You don't say? Is that so?* etc. This last group are different from the other reactive tokens in that they demand a response from the current speaker, so they may be said to have crossed from listenership into the territory of 'new speaker' or are balanced on the cusp between the two. Duncan and Niederehe (1974), while acknowledging that it is often not easy to identify the boundary between listener responses and full-blown, floor-grabbing turns, suggest that response tokens of the types we have considered so far encode a compact between speaker and listener that the turn has not been yielded. Equally, Mott and Petrie (1995) argue that short response tokens should not be seen as interruptions (see also Peters and Wong 2014), so there is an acknowledgement that there exists a territory between silence and full-blown, floor-grabbing content in which speaking turns function to feed back to the speaker for a variety of purposes.

Schegloff (1982: 73) suggested that the turn-taking system has built into it a tendency to 'minimize turn size'; speakers typically say no

more than what is necessary for efficient communication. Communicative economy certainly seems to be characteristic of the brief response tokens discussed so far. However, McCarthy (2002) argued that listeners attend equally to the interactional/relational aspects of the conversation as to the informational content and that 'economy' must take both transactional and relational exigencies into account. A year later, McCarthy (2003: 40) concluded: "Speakers do not, it seems, economize when it comes to sociability, unless there are the most urgent circumstances demanding a purely transactional response".

The contributions of listeners must be seen as far from constituting mere acknowledgement; listener responses are a part of the woven fabric of the conversation as an artefact. Schegloff (1982: 74) further underlines this point, asserting that to regard discourse as 'a single speaker's, and a single mind's, product' is a mistaken approach. Examples of studies which do take listeners more fully into account include the papers in McGregor (1986), Bublitz (1988), McGregor and White (1990), Rühlemann (2007, 2019) and Peters and Wong (2014). Rühlemann (2019: 42), for instance, offers the illuminating terms 'supporting speaker' and 'responsive recipient' to capture the activity of non-floor-grabbing, active and natural listenership.

Short, responsive items have gone by different names in the literature. Yngve (1970) used the term *backchannel* to describe the 'short messages' that the speaker attends to while holding the floor (1970: 568). Fellegy (1995) chooses *minimal response* (see also Coates 1986; Zimmerman 1993). Roger et al. (1988) refer to *listener response*. In this chapter, we use the term *response tokens* to refer to the verbal non-floor-grabbing items that listeners use to respond to the floor-holding speaker in social conversation. Response tokens are often divided into *minimal* and *non-minimal* tokens, though the distinction is ultimately no more clear-cut than that between listener and speaker. Minimal responses are typically defined as brief tokens (*no*, *yeah*) or non-verbal vocalisations (*mm, aha*). Non-minimal response tokens are mostly what would be classes in sentence grammar as adverbs or adjectives (e.g. *good, fine, really, absolutely*) or short phrases/clauses (e.g. *You're joking!, Is that so? Fair enough, That's right, No way!*). In this chapter, we confine our interest to verbal phenomena in short,

responsive turns, and set a boundary between them and floor-grabbing turns or overt and purposeful interruptions, all the while acknowledging the grey area in which listenership merges seamlessly into speakership.

4.4 Corpus Evidence

We examine corpus evidence for a repertoire of items which project non-floor-grabbing listenership and other evidence for the joint activity which natural listenership betokens in social conversation. We argue that the corpus evidence suggests that the pedagogy of speaking and listening needs to come together in a symbiotic partnership which takes into account the active role good listeners take on. We hold that good, natural listenership is a central element of efficient communication and that, in addition, active listenership is relevant to the modelling of fluency in speaking and relevant to the teaching of fluency. In the present chapter, we exemplify listener responses using the 11.5 million-word Spoken British National Corpus of 2014 (Love et al. 2017).

4.4.1 Forms of Listenership: Freestanding Response Tokens

McCarthy (2002) used British and North American conversational data to catalogue a range of non-minimal response tokens (in contrast with minimal responses such as *yeah, no*, or *mm*). The list included items such as *right, exactly, fine, true, great, definitely* and *good* for the British data and a similar list for the American data with some differences in frequency ranking—for example *wow* and *gosh* appeared higher in rank on the American list. Alongside these results, O'Keeffe et al. (2007) noted the occurrence of response items with religious associations in Irish English conversational data, such as *oh my God, Jesus Christ* and *God help us*, underscoring the aspect of cultural nuance referred to earlier. McCarthy (2002) further noted the relative frequencies of occurrence of the key items as response tokens compared with their other uses and noted that items such as *right, gosh, true, wow* and *absolutely* functioned as response tokens in more than 50% of all their occurrences.

The dual-or multi-functionality of words such as *right, fine, great, exactly, absolutely*, etc. underlines the difficulty of categorisation referred to earlier. If a high percentage of a word's occurrences function in a way that traditional word class labels such as adjective or adverb fail to capture, then this further strengthens the argument for establishing in a grammar of speaking a class of items whose primary function has become, through use, the encoding of listenership. This process of grammaticalization or pragmatic specialism is a characteristic of the emergent nature of language use in everyday speaking. The corpus evidence is powerful, and the frequency data enable us to circumscribe a useful repertoire of items for the teaching of listenership, which we explore later in this chapter.

When we consider the responsive repertoire in context, we see that it shows great flexibility, ranging from single-word turns consisting only of a response token, via repetitions of tokens and various combinations of them, as exemplified in extracts (2) to (4).

Extract 2.
<S?* > what you are is more like to do with your environment and
<S0475> **absolutely**
<S?*> the effects it had on you
[* unidentified female speaker] (Spoken BNC2014 SGWU)

In Extract 2, the original transcript indicates an overlap between the first three turns, further supporting the interpretation that speaker 0475's *absolutely* is received neither as an interruption nor as floor-grabbing, yet it is clearly fully lexical and encodes more than mere acknowledgement, indicating enthusiastic acceptance of the proposition about environmental influence. Extract 3 shows multiple repetition of *right*, again projecting more than minimal engagement with the incoming talk; we also see reactive *oh*, both alone and in combination in *oh nice*:

Extract 3.
[Swindon is a southern English town]
<S0162> Swindon are in the, some football final
<S0073> yeah Swindon yeah

<S0129> playing Wembley
<S0162> and you know that [anonymised names] are Swindon fans
<S0073> **oh**
<S0162> and they've got a ticket for [anonymised name]
<S0073> **ah right right right right right right**
<S0129>they've got a spare ticket and I just happen to be going to Wembley the next day on Monday by sheer coincidence to go for a [unclear] at Wembley arena
<S0073> **oh nice**
<S0129> so if I can change it to go the day before oh hello excellent (Spoken BNC2014 S5PF)

Some response tokens, notably *absolutely, definitely* and *certainly* can be negated with *not*. These most typically occur in responses agreeing with a negative proposition by the current speaker, as in Extract 4:

Extract 4.
[speakers are discussing child adoption]
<S0196> funnily enough it came up Saturday night <pause> and er you know cos they said oh you know when do you think you'll start trying? and <pause> I just said you know I'm very hesitant I'm very content and very happy and I'm
<S0192> mm
<S0196> not ready for anybody else to come into that right now
<S0192> no no <pause > **definitely not** (Spoken BNC 2014 S5XD)

These are all examples of active listeners doing what comes naturally—attending to, engaging and converging with the current speaker, briefly and efficiently, yet with sufficient lexical content and personal involvement to enable the current speaker to monitor the reception of their own contribution moment-by-moment and to adapt it to the feedback offered (see Tolins and Fox Tree 2014 for an interesting discussion of how listener feedback can shape the course of oral story-telling). In this sense, the talk is jointly constructed, proceeding and flowing naturally, thanks to the back and forth of each party attending successively to each other's utterances—the two sides of the metaphorical coin.

4.4.2 Response Tokens as Turn-Openers

One of the features of the response token repertoire is that they may also operate as the initial word(s) in longer turns, whereby the listener transitions to become the speaker. In Extract 5, we see two occurrences of *really*; the first occupies a single-word turn (marked as an overlap in the original transcript) and is simply received as an acknowledgement token or at best a request for further elaboration, while the second is followed by a question to the previous speaker. The previous speaker then cuts off their contribution mid-flow and attends to the question, thus acknowledging that speakership has passed to the questioner:

Extract 5.
[speakers are discussing a device for extending and boosting the internet signal]
<S0520> I think sometimes there's just like a little bug in the system with that
<S0519> I do too er because I've had it that with my iPad even before we had the extender that when I arrived it didn't recognise
<S0520> **really**?
<S0519> or maybe it was since the extender but we do have an extender in [place name anonymised] and we never had any problem with it so I'd be even tempted to
<S0521> **really? okay is that the same make**?
<S0519> it's the same one
<S0520> mm (Spoken BNC2014 S24E)

Response tokens in the turn-initial slot align well with Tao's (2003) research on turn-openings in conversation. Tao's American conversational data demonstrated that turn-initial items in English are generally syntactically independent lexical items. Tao's list of turn-openers includes *yes, well, right, okay*, pronouns (occurring in fixed expressions such as *I think, You know, I mean*) and *that's* + adjective expressions (*that's right, that's true*), etc. The turn-initial position displays items which show interpersonal engagement and convergence, just as the same items do when used alone without further content. In short, while response items may occur as the opening word(s) of a floor-grabbing turn, they still perform the duties of good listenership.

The natural, conventional behaviour can be summed up in the adage, "Before you say what you want to say, say something about what you've just heard". Listenership therefore permeates transitions to speakership, summed up in the description of the three internal components of a speaking turn proffered by Sacks et al. (1974: 36), the first of which "addresses the relation of the turn to a prior".

4.4.3 Joint Construction and Confluence

Another aspect of listenership is the way listeners pick up the syntactic thread of a speaker's turn and continue it as they assume speakership. Clancy and McCarthy (2014) looked at how listeners latch utterances onto those of current speakers using various clausal devices. An example of this is seem in Extract 6, where the current speaker uses a *which* comment clause, and the listener immediately assumes speakership with an echoic, matching *which* clause:

Extract 6.
[speakers are discussing the need to train puppies]
<S0337> it's not a erm you have to train them they just they're they're familiar with it
<S0338> yeah brilliant
<S0337> **which makes sense**
<S0336> **which is easier said than done like**
<S0337> yeah <laugh> (Spoken BNC2014 S2K7)

The original transcript shows considerable overlapping in Extract 6, demonstrating how quickly, smoothly and efficiently speakers and listeners align their turns to create a seamless syntactic flow. Extract 6 is typical of the way *which*-clauses are used to express an evaluation or opinion; they may also be tagged on by listeners when they wish to provide what they consider to be further relevant information or elaboration, as in Extract 7:

Extract 7.
[speakers have visited a big antiques shop and later, by chance, find another one]
<S0198> yeah found another massive one

<S0230> yeah
<S0198> and had it not been for the signpost
<S0230> yeah **which I spotted**
<S0198> **which you spotted** yeah
<S0192> nicely done
<S0230> < laugh>
<S0198> we wouldn't've wouldn't've stumbled across it
<S0230> no (Spoken BNC 2014 S2KP)

These joint constructions demonstrate the ease with which interlocutors create and maintain conversational flow. The term *flow* is not otiose: it points to the joint responsibility of participants to create a smooth conversational progression. We saw how the response tokens, when used at the beginning of a speaker's turn, fulfilled the need to comment on the previous discourse before embarking upon new content, thus linking speaker turns into chains of jointly produced, jointly negotiated meanings. Schiffrin (1994: 351) sums this up as 'the emerging set of understandings that participants gain through the give and take of interaction—through the process of orienting towards the other person'. Each participant orienting towards the other produces what McCarthy (2010) calls *confluence*, an approach to fluency that embeds it in joint construction rather than seeing it as the burden of the individual speaker.

4.5 The Pedagogy of Good Listenership

4.5.1 How Many Skills?

Where does this lead us in terms of pedagogy? If listenership plays a central role in conversation, as we have argued so far, then it behoves us to incorporate listenership skills in our pedagogy. We owe students the right of access to the keys to successful, fluent and natural conversational exchanges, as exemplified in response tokens and other joint construction principles. But how to achieve this? Classrooms are indeed a place where response tokens are liberally used but these are often in evaluations by teachers (*okay, right, good*) rather than by students in pair work conversational exchanges, an observation that was evidenced as far back as the data in Sinclair and Coulthard's (1975) classroom study which led to their original description of the structure

of spoken discourse. And such an imbalance in speaker-listener roles is likely to continue unless active listenership is incorporated into teaching materials.

The term 'the four skills' is deeply rooted in both the commercial and academic branches of language teaching, not just ELT. Course books are often described as 'four-skills courses' in publishers' promotional literature. Amongst researchers, 'the four skills' remain a focus of interest as evidenced by the number of article and book titles containing this phrase (Clenton and Booth 2020 is a recent example). This chapter questions the notion of *the* four skills as a finite set and proposes the addition of at least one further skill: that of interacting, for present purposes in conversational settings. As we have suggested, interacting involves an inseparable integration of comprehension and production and is the basis for the co-creation of successful talk, requiring participants to draw on a number of skills. Materials should ideally, therefore, set out to enable students to acquire these skills and achieve interactional competence. We describe below our own experience of attempting this in our co-authorship of a six-level English Language Teaching coursebook series.

4.5.2 Listenership in the Syllabus

The starting point for the pedagogical application of the language research as described above is to design a principled, graded syllabus. McCarthy et al. (2012, 2014a-e) devised a syllabus of the many subskills involved in interaction, referred to as *Conversation strategies*, of which good listenership is one of the four main elements. A fuller description of the four parts of this conversation syllabus can be found in McCarthy and McCarten (2018), but here we will focus on the listenership element. This is distinct from the listening *comprehension* syllabus strand, which requires students to listen to spoken texts of various genres in the role of third-party eavesdropper. The aim of listening comprehension is to develop the skills of extracting meaning of a general or detailed nature—usually factual—where students show their comprehension of the text in activity types such as chart filling, multiple choice items, True/False statements and so on. The listenership syllabus, on the other hand, places students in the role of potential or actual participant within a conversational setting. The goal

here is for learners to demonstrate their understanding by reacting, saying something appropriate to the context or which expresses their own personal attitude to what another speaker has said.

Listenership skills and the language items which realise them are included in the syllabus, for both receptive and productive use and from elementary to more advanced levels. The inclusion of such skills even for relatively low-proficiency learners aligns with Harrington's (2018) corpus-based illustrations of the survival communicative instincts of discourse communities operating with limited linguistic resources. Items are graded according to factors including frequency, which is an indicator of their currency and usefulness, ease of explanation in student-friendly terms and the complexity of the language students are required to manipulate in employing them. By way of examples, at elementary level there are simple one or two-word response tokens such as *Really? Right, That's great*, etc.; strategies with more complex linguistic realisations, such as co-construction with *which* clauses, feature at a more advanced level.

4.5.3 Methodological Issues

The methodology to deliver this and other parts of the Conversation strategies syllabus was inspired by the approach proposed in Carter and McCarthy (1995): the three 'I's of Illustration—Interaction—Induction, which are referred to below. The most suitable vehicle for teaching conversational behaviour, we would argue, is extracts of natural conversation to show or *illustrate* typical or representative uses of the target items or strategies. The issues surrounding the choice of appropriate extracts from a spoken corpus and reasons for editing them are many and will not be elaborated here (for a fuller discussion of these issues in relation to published materials, see McCarten 2022). However, it is our view that the judicious editing of conversations can still result in natural or naturalistic conversations which are accessible and contribute to achieving the material's pedagogical aims. For example, it would be counter-productive to edit out from a transcript listener responses in a conversation aimed at illustrating active listenership.

Comprehension of the conversation's propositional content—its 'aboutness'—is a necessary first step before students focus in on

aspects of language use. As part of their pedagogy, McCarthy et al. (2012, 2014a-3) adopt a noticing methodology to encourage students to *interact* with the conversational text and identify examples of the target strategy and associated items. At lower levels these might include showing interest or surprise by repeating a word or using the token *Really*? (see McCarthy et al. 2014a: 28); a more advanced strategy would be to paraphrase another speaker's contribution in drawing a conclusion (e.g. *So what you're saying is* … see McCarthy et al. 2012).

The process of enabling students to incorporate a target strategy or item such as a response token into their own conversational language repertoire—*induction*—is through repeated encounters with it, in natural conversational settings. To facilitate learning, activity sequences may be staged through phases where the student is placed in that third party observer role, before they have the opportunity to use it for themselves. Such steps may include reading and writing conversational exchanges, for example matching comments and responses, or choosing the best response. However, these activity items serve as further illustrations of how the strategy functions and are, wherever possible, the springboard for the students' own personalised use, which should be the final goal of any sequence of activities.

4.5.4 Listenership Activities

Let us take as an example responses with *That's* + adjective (e.g. *That's great*) or simply an adjective (G*reat!*) to express a positive or negative attitude to what has been said. This is one example of a listenership skill that can be taught at elementary level once students have learned the typical meanings of basic adjectives. Activity items may include choosing a best-fit (rather than 'correct') response in short exchanges, either as a written activity without audio, or as a listening and speaking activity with audio. In the first example, A's part can be read or heard. The item is set within the conversational theme of catching up with friends.

(1) **Choose the best expression to respond.**

A: My grandmother's staying with us this week.
B: That's interesting. / That's nice. / That's right.

This can be followed by students suggesting their own responses to pieces of information, again as a written or spoken activity, as in the example item below.

(2) **Respond to A's comments using an expression with *That's.***

A: I just started an art class.
B: _____ .

Such activities can also be extended by asking students to suggest appropriate follow-up questions or comments, modelling the principle of 'saying something about what you've just heard before you say or ask more', as discussed in the first part of this chapter. Finally, and arguably the most important part of the sequence, is that students then try out for themselves such responses and follow-ups in their own real, personalised conversations within the context of the theme of the lesson. (For further examples, see McCarthy et al. 2014a: 71 and 2014b: 26–27). In designing such activities, however, it is important that the content reflects natural, realistic conversational behaviour and content so that it supports and scaffolds students' own production (see McCarthy and McCarten 2018 for further discussion of practice).

Unlike traditional grammar and vocabulary activities, where accuracy and correctness of the language item are paramount, listenership activities often have no right and wrong answers and can be unpredictable. It is a matter of personal choice whether the student responds in Example 2 above with e.g. *That's nice. I love art.* or *That's weird. Me too. Who's your teacher?* Therefore, such activities need to be seen as a means of improving students' general fluency and not simply a vehicle to reinforce the target item of the lesson. Unpredictability, in this case, is to be seen as an opportunity rather than a risk.

4.6 Conclusion

In this chapter we have argued, based on research evidence, that there is more to listening than just comprehending, and that the roles of speakers and listeners in natural, successful conversational interaction are harmoniously intertwined in what we have referred to as

confluence. Confluence is more than just an accidental after-effect of a conversational encounter. The fact that we can observe it after the event as an artefact of talk hides its emergent character, its moment-by-moment, purposeful adjustments on the part of all parties in the conversation, whether it be listeners feeding back to speakers or speakers adjusting their message to take into account whatever listeners feed back to them. Speakers listen to listeners and listeners listen to speakers. They are two sides of the same coin. To split the coin into two separate parts is to debase the currency.

References

Bublitz, Wolfram. 1988. *Supportive Fellow-Speakers and Cooperative Conversations*. Amsterdam: John Benjamins.

Carter, Ronald, and Michael J. McCarthy. 1995. Grammar and the Spoken Language. *Applied Linguistics.* 16 (2): 141–158.

Clancy, Brian, and Michael J. McCarthy. 2014. Co-constructed Turn-Taking. In *Corpus Pragmatics: A Handbook*, ed. Karin Aijmer and Christoph Rühlemann, 430–453. Cambridge: Cambridge University Press.

Clenton, Jon, and Paul Booth, eds. 2020. *Vocabulary and the Four Skills: Pedagogy, Practice, and Implications for Teaching Vocabulary*. Abingdon: Routledge.

Coates, Jennifer. 1986. *Women, Men and Language: A Sociolinguistic Account of Sex Differences in Language*. London: Longman.

Drummond, Kent, and Robert Hopper. 1993a. Back Channels Revisited: Acknowledgment Tokens and Speakership Incipiency. *Research on Language and Social Interaction.* 26 (2): 157–177.

———. 1993b. Some Uses of Yeah. *Research on Language and Social Interaction.* 26 (2): 203–212.

Duncan, Starkey. 1974. On the Structure of Speaker-Auditor Interaction During Speaking Turns. *Language in Society.* 3 (2): 161–180.

Duncan, Starkey, and George Niederehe. 1974. On Signalling That It's Your Turn to Speak. *Journal of Experimental Social Psychology.* 10 (3): 234–247.

Fellegy, Anna M. 1995. Patterns and Functions of Minimal Response. *American Speech.* 70 (2): 186–199.

Fries, Charles C. 1952. *The Structure of English*. New York: Harcourt, Brace.

Gardner, Rod. 1997. The Listener and Minimal Responses in Conversational Interaction. *Prospect.* 12 (2): 12–32.

———. 1998. Between Speaking and Listening: The Vocalisation of Understandings. *Applied Linguistics.* 19 (2): 204–224.

———. 2001. *When Listeners Talk: Response Tokens and Listener Stance*. Amsterdam: John Benjamins.

Harrington, Kieran. 2018. *The Role of Corpus Linguistics in the Ethnography of a Closed Community: Survival Communication*. Abingdon, Oxon: Routledge.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. The Spoken BNC2014: Designing and Building a Spoken Corpus of Everyday Conversations. *International Journal of Corpus Linguistics* 22 (3): 319–344.

McCarten, Jeanne. 2022. Corpus-informed Course Design. In *Using Corpora to Explore Linguistic Variation*, ed. Anne O'Keeffe and Michael J. McCarthy, 49–71. Amsterdam: John Benjamins.

McCarthy, Michael. 2002. Good Listenership Made Plain: British and American Non-minimal Response Tokens in Everyday Conversation. In *Using Corpora to Explore Linguistic Variation*, ed. Randi Reppen, Susan Fitzmaurice, and Doug Biber, 49–71. Amsterdam: John Benjamins.

McCarthy, Michael J. 2003. Talking Back: "Small" Interactional Response Tokes in Everyday Conversation. In: Coupland, Justine. (Ed.). Special issue of *Research on Language and Social Interaction* on 'Small Talk'. 36(1), 33–63.

———. 2010. Spoken Fluency Revisited. *English Profile Journal.* 1. http://journals.cambridge.org/action/display/Journal?jid=EPJ. Accessed 29 November 2021.

McCarthy, Michael J., and Jeanne McCarten. 2018. Practising Conversation in Second Language Learning. In *Practice in Second Language Learning*, ed. Christian Jones, 7–29. Cambridge: Cambridge University Press.

McCarthy, Michael J., Jeanne McCarten, and Helen Sandiford. 2012. *Viewpoint Student's Book 1*. Cambridge: Cambridge University Press.

McCarthy, Michael J., McCarten, Jeanne and Helen Sandiford. 2014a–d. *Touchstone Student's Book 1–4, Second Edition.* Cambridge: Cambridge University Press.

McCarthy, Michael J., Jeanne McCarten, and Helen Sandiford. 2014e. *Viewpoint Student's Book 2*. Cambridge: Cambridge University Press.

McGregor, Graham, ed. 1986. *Language for Hearers*. Oxford: Pergamon Press.

McGregor, Graham, and R.S. White. 1990. *Reception and Response: Hearer Creativity and the Analysis of Spoken and Written Texts*. London: Routledge.

Mott, Helen, and Helen Petrie. 1995. Workplace Interactions: Women's Linguistic Behavior. *Journal of Social Psychology.* 14 (3): 324–336.

Norrick, Neal R. 2012. Listening Practices in English Conversation: The Responses Responses Elicit. *Journal of Pragmatics.* 44 (5): 556–576.

O'Keeffe, Anne, Michael J. McCarthy, and Ronald Carter. 2007. *From Corpus to Classroom*. Cambridge: Cambridge University Press.

Oreström, Bengt. 1983. *Turn-taking in English Conversation*. Lund: Gleerup.

Peters, Pam, and Deanna Wong. 2014. Turn Management and Backchannels. In *Corpus Pragmatics: A Handbook*, ed. Karin Aijmer and Christoph Rühlemann, 408–429. Cambridge: Cambridge University Press.

Roger, Derek, Peter Bull, and Sally Smyth. 1988. The Development of a Comprehensive System for Classifying Interruptions. *Journal of Language and Social Psychology.* 7 (1): 27–34.

Rühlemann, Christoph. 2007. *Conversation in Context: A Corpus-driven Approach*. London: Bloomsbury.

———. 2019. *Corpus Linguistics for Pragmatics*. Abingdon, Oxon: Routledge.

Sacks H., Schegloff E. A., and Jefferson G. 1974. A simplest systematics for the organisation of turn-taking for conversation. *Language*. 50 (4): 696–735.

Schegloff, Emanuel A. 1982. Discourse as Interactional Achievement: Some Uses of 'uh huh' and Other Things That Come Between Sentences. In *Analyzing Discourse: Text and Talk*, ed. Deborah Tannen, 71–93. Washington, DC: Georgetown University Press.

Schiffrin, Deborah. 1994. *Approaches to Discourse*. Oxford: Blackwell.

Sinclair, John, and Malcolm Coulthard. 1975. *Towards an Analysis of Discourse*. Oxford: Oxford University Press.

Stubbe, Maria. 1998. Are You Listening? Cultural Influences on the Use of Supportive Verbal Feedback in Conversation. *Journal of Pragmatics.* 29 (3): 257–289.

Tao, Hongyin. 2003. Turn Initiators in Spoken English: A Corpus-based Approach to Interaction and Grammar. In *Corpus Analysis, Language Structure and Language Use*, ed. Charles Meyer and Pepi Leistyna, 187–207. Amsterdam: Rodopi.

Tolins, Jackson, and Jean E. Fox Tree. 2014. Addressee Backchannels Steer Narrative Development. *Journal of Pragmatics.* 70: 152–164.

Tottie, Gunnel. 1991. Conversational Style in British and American English: The Case of Backchannels. In *English Corpus Linguistics*, ed. Karin Aijmer and Bengt Altenberg, 254–271. London: Longman.

Warren, Martin. 2006. *Features of Naturalness in Conversation*. Amsterdam: John Benjamins.

Yngve, Victor H. 1970. On Getting a Word in Edgewise. In *Papers from the 6th Regional Meeting, Chicago Linguistic Society*. Chicago: Chicago Linguistic Society.

Zimmerman, Don H. 1993. Acknowledgement Tokes and Speakership Incipiency Revisited. *Research on Language and Social Interaction.* 26 (2): 179–194.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_5

5. Corpus Linguistics and Writing Instruction

Eric Friginal[1] ✉, Ashleigh Cox[2] ✉ and Rachelle Udell[2] ✉
(1) The Hong Kong Polytechnic University, Hung Hom, Hong Kong
(2) Georgia State University, Atlanta, GA, USA

✉ **Eric Friginal (Corresponding author)**
Email: eric.friginal@polyu.edu.hk

✉ **Ashleigh Cox**
Email: acox27@gsu.edu

✉ **Rachelle Udell**
Email: rudell1@gsu.edu

Keywords Corpus linguistics – Pedagogy – L2 writing – Data-driven learning

5.1 Introduction

Corpus linguistics (CL) is a research approach to the study and exploration of spoken and written discourse patterns, structure, and use (Biber et al. 2010). A corpus is a large and principled collection of computer-readable, authentic texts that are sampled to be representative of a particular language or discourse variety (Biber et al. 1998). Corpora, therefore, may serve as datasets of actual language analyzed and utilized for a variety of purposes by researchers and, especially, teachers, as well as language learners themselves, when introduced effectively in the classroom (Friginal and Cox 2022).

Over the years, the use of corpora has become very popular in the analysis of the linguistic characteristics of written discourse (Durrant et al. 2021). This approach has resulted in the development of more authentic teaching materials in academic writing classrooms that represent actual language-in-use across settings and contexts, i.e., *registers* (Friginal 2018; Friginal and Hardy 2014). Direct applications of corpora and corpus tools in writing classrooms support

various language acquisition theories and approaches, especially related to use of realia and authentic texts, positive motivation through learner-computer and learner-learner interactions in writing and editing, explicit teaching of written language features and patterns, and learner autonomy in developing effective writing skills (Römer et al. 2020; Friginal et al. 2020).

There has been an exponentially increasing number of writing instructors who have utilized corpora and corpus-based materials in their classrooms from the mid-1990s to the present. Online databases and corpus tools such as concordancers are now easily accessible, and several CL for teaching textbooks have been published in the past two decades. Considering their applications, it is easy to envision the positive contribution of corpus-based approaches to a variety of learning contexts in writing instruction. From as early as 2005, Teubert noted that CL has been held to be the default resource in linguistic research since it consists of real language data. Many learners, therefore, will benefit from the practical and pragmatic applications of corpus data as they learn about the acquisition of sound writing skills in their classrooms. As Römer (2009: 140) observes, "language is highly patterned", and often, these patterns are important to highlight and teach explicitly in the writing classroom, covering a range of features from phraseological patterns to more complicated clausal, sentence, and discourse-level structures (Friginal et al. 2020; Friginal and Cox 2022).

This chapter is an overview of recent corpus-based research focusing on written registers primarily in academic contexts and explores the use and applications of corpora in the writing classroom. Gray (2011) argues that academic writing has distinct characteristics that make it different from other types of registers, and several studies (e.g., Hiltunen 2016; Zhang et al. 2021) from the field of applied linguistics that used a range of different research methodologies have proven this statement to be accurate. Over the past two decades, corpus-based methodologies have been favored in the analysis of a wide variety of written texts frequently produced in academia because these methodologies provide reliable and generalizable findings that successfully describe these registers and, in many cases, have direct pedagogical applications (Römer et al. 2020).

5.2 Exploring Written Registers

One of the ways that corpora can be used to enhance the teaching of writing across various learner levels and language backgrounds is by deriving practical implications from corpus studies that explore the characteristic features of written registers (Reppen 2016). Teachers can use the findings of corpus studies on writing to inform pedagogical choices. Some examples of types of

corpus studies that could be helpful for teachers are comparative corpus studies, studies on writing development, and studies of English for Specific Purposes (ESP) and English for Academic Purposes (EAP).

5.2.1 Comparative Corpus Studies

A popular approach in corpus-based research has been to compare two corpora of different sub-registers, such as professional and novice academic writing or L1 and L2 writing, to look for ways that learners can improve their writing. Studies comparing novice and expert writing, for example, can provide teachers with an idea of what 'good' writing entails and what students are lacking in comparison to professional writers, which can inform teachers' curriculum design. Teachers could examine studies exploring specific characteristics of interest to them such as differences in frequent phrases (Römer 2009), structures related to the target genre (Flowerdew 2003; Hartig and Lu Hartig and Lu 2014), lexical bundles (Zhang et al. 2021), or noun modifiers (Ansarifar et al. 2018), with the goal of establishing an idea of how students can work towards writing with greater expertise.

Regarding L2 writing research, Granger (2015) recommends comparing corpora of learner and native speaker writing to draw pedagogical implications, and there are many studies that have taken this approach (i.e., utilizing learner corpora) generally addressing the needs of writing instructors. For example, in terms of grammatical studies, teachers can read descriptions of complement clauses (Biber and Reppen 1998), verb complementation patterns (Deshors 2015), and syntactic complexity in L1 and L2 writing (Lu and Ai 2015). Another feature of writing that has been explored in L1/L2 corpus comparisons that may interest writing instructors is the use of words and structures that are traditionally seen as *informal* in formal academic writing (Lee et al. 2019), such as the use of the singular first-person pronoun in learner and native speaker writing (Chang 2015). Studies on common features in writing from learners from specific L1 backgrounds (e.g., Leńko-Szymańska 2015) may be useful for EFL teachers whose students share the same first language. There have been many other published studies comparing L1 and L2 writing, and teachers can search for studies related to the features that are most relevant for their learner population.

5.2.2 Writing Development

Another popular area of research that may be helpful for teachers is the concept of writing development. Durrant et al. (2021) conduct a meta-analysis of 248 studies using corpora to describe L1 or L2 writing syntax, vocabulary, formulaic language, or cohesion development. Their findings on features associated with high quality writing may be particularly helpful for teachers

because they provide concrete ideas about what characteristics are valued in L1 and L2 writing. In L1 writing, perceived quality is associated with the length of t-units, or independent clauses with any dependent clauses attached to them, increased use of adjectives and prepositional phrases, increased type counts, fewer high frequency words and more low frequency words, and in some genres, noun phrase complexity, passive voice, and lexical cohesion. In L2 writing, Durrant et al. (2021) find perceived quality to be associated with sentence, t-unit, and clause length; noun phrase complexity; frequency of relative clauses and prepositional phrases—possibly depending on the genre; passive voice in non-narrative writing; increased type counts; academic vocabulary; large proportions of commonly used lexical bundles; and collocations with higher mutual information statistic values—meaning words that have a closer association with each other. While there are overlaps between L1 and L2 writing development, Durrant et al. (2021) note that there are differences in the features underlying perceived quality in L1 and L2 writing. Instructors interested in these differences could look into studies that reflect their students' age group and learning context to learn about how to help their students develop high quality writing (Friginal et al. 2014).

5.2.3 EAP-ESP Research and Technical Writing

For English for Academic Purposes (EAP) and English for Specific Purposes (ESP) writing instructors, corpus studies in writing genres similar to the students' domain are likely to have helpful implications. Swales (1990, 2004) is universally considered to be the lodestar in using corpus data to inform and guide pedagogical best practice, especially when the goal is to improve EAP student writing. He has used corpora to investigate rhetorical moves and steps that characterize academic research articles—a study motivated by the desire to improve and develop the academic writing skills of his own students. Swales' instructionally-focused contributions contrast notably with Biber's (2006) corpus study of university language across registers and academic disciplines and Hyland's (2012) inclusion of corpus data in explorations of variations in language use across academic disciplines that signal disciplinary identity. Nonetheless, the results of Biber's and Hyland's research have proven useful for teachers of academic writing and have inspired other corpus researchers to investigate the linguistic forms and functions of written language across a variety of EAP/ESP fields (see Römer 2010, for an exploration of the pedagogical applications of utilizing the Michigan Corpus of Upper-Level Student Papers or MICUSP to identify lexico-grammatical and phraseological patterns in advanced student writing). Building off of these foundations, the EAP community has come to wholeheartedly embrace the use of corpora and corpus tools for pedagogical research (e.g., Flowerdew 2015, 2017), materials

development (e.g., Friginal 2013, 2018), and even teacher training (e.g., Chen et al. 2019; Friginal et al. 2020).

Some specific examples of other linguistic features and contexts that have been studied in EAP corpora are complement clauses (e.g., Biber et al. 2002), academic vocabulary in medical research articles (e.g., Chen and Ge 2007), passives in research articles and student essays (e.g., Hiltunen 2016) and the co-occurrence of multiple linguistic features in research articles (e.g., Gray 2011). Exploring findings relevant to students' proficiency level and target genre could be very informative for teachers.

There is also a wealth of ESP writing research exploring a range of features in writing from different fields such as stance expressions in dentistry texts (Crosthwaite et al. 2017), passive voice and nominalizations in legal texts (Hartig and Lu Hartig and Lu 2014), politeness in job application letters (Upton and Connor 2001) and problem-solution structures in recommendation-based technical reports (Flowerdew 2003). Writing instructors can search for ESP-based corpus studies that involve texts that reflect their target domain and purpose to learn about research that may be relevant and useful for their group of learners. Pedagogical Example 1 below illustrates how teachers can use findings from corpus research to develop classroom activities.

Pedagogical Example 1 Using Corpus Research to Inform Pedagogical Decisions

Below is an example of how teachers can use research centered on frequently used verb forms that characterize academic discourse to develop pedagogical activities.

Step 1: Alangari et al. (2020) point out that the use of phrasal verbs is often frowned upon in academic writing. However, their research and that of Liu (2012) utilizes corpus analysis to pinpoint frequently occurring phrasal verbs and other multi-word constructions used in professional academic texts. English teachers in the upper-secondary levels as well as those with undergraduate learners can use the results of these studies to identify key multi-word structures that characterize academic writing and the important discourse functions they perform.

Step 2: Instructors can develop activities to help students notice the subtle shifts in register that occur when phrasal verbs are replaced with single, lexical verbs that carry the same meaning. Likewise, students can be directed to notice when and where professional writers choose to use phrasal verbs. Further, teachers and students may find it helpful to discuss why such strategic uses are both necessary and effective.

Step 3: Teachers can develop writing revision activities to help students make strategic verb and phrase choices to better align their writing with

reader expectations for academic registers. A sample activity follows below:

Writing Revision Activity (Handout instructions for students). The following selection is an excerpt of a draft paper written by a student. Read through the excerpt and, with a partner, complete the following tasks:

1. Identify and highlight all of the **main verbs**, taking care to differentiate between *phrasal verbs* and *single, lexical verbs*.

2. Talk to your partner about what the ***phrasal verbs*** mean. For each one, identify whether it aligns or does not align with reader expectations for academic register.

3. For phrasal verbs that do not align with expectations for an academic register, replace them with single, lexical verbs that match the student author's intended meaning.

4. For phrasal verbs that do align with expectations for an academic register, write 1–2 sentences explaining what function these phrasal verbs serve and why the student author may wish to keep them in place (rather than replace them with a single, lexical verb).

Note: Teachers who are familiar with their student population should also consider information from their needs analyses when developing pedagogical materials. Before using this sample activity, teachers can make adaptations. An important consideration, for example, would be whether students know the difference between phrasal verbs and single lexical verbs.

5.2.4 Resources for Corpus-Based Materials in Writing Classrooms

There are many available resources containing corpus-informed materials and ideas for teachers. In an article aimed at writing instructors, Cortes (2018) provides a varied list of accessible corpora that academic writing instructors can use for curriculum and materials development as well as a series of corpus tools that can be incorporated into more direct instructional activities. Charles (2018a) also targets an audience of writing teachers, but instead delivers an overview of tools and software available for student use. Resources include a list of online corpora (from the small and specialized to the large and general), concordancing tools, and websites designed to aid students in collecting and analyzing their own corpora. Two books written for academic writing instructors are worthy of mention here. Aull's (2015) book presents a corpus analysis of academic writing produced by first-year, undergraduate students.

The author defines the linguistic features of first-year writing and compares them with those present in professional academic writing with the goal of developing an instructional plan to modify the features of first-year writing to better align with the established expectations of professional academic discourse. In the final chapters, Aull offers numerous pedagogical applications for writing instructors including a thorough discussion of four elements of academic argumentation that are frequently missing in first-year papers: hedging language, phrase structures that indicate the argument's scope, transitions and other cohesive devices, and reformulations of key ideas related to the central claim.

Another notable book, written by Salazar (2014: 4), utilizes corpus research to generate "an inventory of the most frequent and pedagogically useful lexical bundles in scientific prose". Salazar's Health Sciences Corpus (HSC) contains peer-reviewed biology, biochemistry, biomedicine, and medicine research articles written by both L1 and L2 English users. Through analysis of the HSC data, Salazar identifies a list of frequently occurring 'target bundles' that can be taught in terms of their structural and functional characteristics. In addition, Salazar makes comparisons between the types of bundles that frequently appear in both native English speakers' (NES) and non-native English speakers' (NNES) writing, noting areas of divergence and overlap. The book provides writing instructors with 14 activities designed to aid Health Sciences students in noticing the most frequent lexical bundles (words in a corpus that frequently appear together in strings), retrieving data related to lexical bundles from field-specific corpora, and generating appropriate bundles in their own writing.

5.3 Directly Engaging Learners in Corpus Use

There is a preponderance of evidence that corpus-informed pedagogy, especially the use of corpora and corpus tools as a component of direct classroom instruction, produces improved writing competence across drafts (Gilmore 2009), increases register-appropriate word choice and collocation use (Jafarpour et al. 2013), results in higher scores on linguistic element assessments (Birhan et al. 2021), and enhances development of linguistic and genre awareness (Yoon 2011). Students express positive responses to corpus-based instructional materials and demonstrate high levels of engagement in tasks involving direct participation in corpus building and analysis (Nasution 2018; Smith 2020). Furthermore, students also report increased feelings of confidence in their abilities as writers and ownership of their learning process (Lee and Swales 2006).

There are several studies exploring the benefits of direct student engagement with publicly available online corpora. Walker (2011) observes

that teachers of business English sometimes lack sufficient content knowledge related to shades of meaning associated with field-specific terminology and frequently occurring collocations. However, lessons that direct students to use corpora to research key vocabulary and phraseology may allow students to resolve challenging questions, especially those related to positive/negative word connotations. Mansour (2017) also focuses upon student-led corpus analysis to improve English collocation use in academic writing, providing instructors with a series of specific search strings that produce more accurate and targeted results when using corpora.

A few noteworthy studies examine the use of online corpora as a means of encouraging error correction across multiple drafts of academic papers. Gilmore (2009) observes that utilizing the British National Corpus (BNC) and Collins Birmingham University International Language Database (COBUILD) allowed Japanese university students to more effectively respond to corrective feedback resulting in improved drafts of their academic essays. Liou (2019) also demonstrates the effectiveness of using corpora and concordancing software to facilitate EFL university student uptake of written corrective feedback during a semester of classroom-based instruction. Students completed three multi-draft writing assignments in varying genres while using the Corpus of Contemporary American English (COCA) (Davies 2008–) and a number of Chinese-English concordancing programs as they progressed.

A more extensive longitudinal study conducted by Dolgova and Mueller (2019) analyzes learner approaches to correcting four types of errors commonly encountered in academic writing produced by NNES students. Essays written by 175 graduate science students enrolled in EAP courses were collected and analyzed over the course of four semesters. Students were trained how to use COCA to investigate feedback regarding word choice related to meaning, grammatical constructions, punctuation/spelling, and verb/adverb use to signal academic and non-academic registers. Then participants were instructed to use the results of their corpus searches to revise their errors in subsequent drafts. Dolgova and Mueller find that successful revisions were dependent upon error type and noted that word choice and inflection errors that were easily discerned from the immediate results of a corpus search were most likely to be corrected. However, errors requiring a more careful examination of key words in context were far less likely to be addressed. Students also experienced challenges with the COCA user interface and with selection of search criteria. Nevertheless, the participating students deemed corpus tools to be more useful than dictionaries during the revision process and an overwhelming majority of students (97%) indicated that they would definitely use COCA again in the future.

By far, the most popular application of corpus-based data in EAP, as well as other writing contexts, is the use of do-it-yourself (DIY) corpora (i.e. self-compiled corpora of texts in a specific domain) as a tool for students in general courses where the instructor may not be able to focus on discipline-specific features and variations within frequently encountered genres (Charles 2018b). Lee and Swales (2006) were amongst the first to provide evidence that DIY corpus use in academic writing classrooms was not only effective but also well received by students. Their groundbreaking paper documents the results of an innovative corpus-informed EAP writing course for NNES doctoral students, with steps in guiding students through corpus analysis that could be adapted by other teachers: (1) brief training on utilizing corpus building and analysis tools, (2) students' creation of two specialized corpora, one comprising their own writing and another composed of field-specific research articles, (3) manual discourse analysis methods alongside concordancing software to analyze linguistic and rhetorical features as well as genre-specific discourse elements present in each corpus, (4) comparison of the results, highlighting the areas in their own writing that were in need of development, and (5) written reflections on their findings and how the results might benefit their development as writers. Students expressed favorable attitudes regarding their corpus building experience, noting their increased sense of power and agency as writers and their sense that their self-constructed comparison corpora were of more value to them than traditional instructional materials that are focused on vocabulary and grammatical structure. Building upon Swales' (1990, 2004) use of corpora to identify the rhetorical moves and steps in general research article introductions, others have employed the DIY corpus approach to help students discover the features specific to discussion sections in science and engineering thesis papers (Flowerdew 2015), developing research report writing skills in forestry (Friginal 2013), the introductions of engineering research articles (Dong and Lu 2020), and the refutation section of argumentative essays (Charles 2018a). Lu et al. (2021) observe that taking such a blended genre and corpus-based approach to academic writing pedagogy allows instructors to assist novice writers as they develop rhetorical and linguistic knowledge in a variety of discipline-specific genres. Smith (2020) also notes that the use of DIY corpora allows for the targeted, specific selection of articles and texts featuring critical variations in vocabulary use that most closely align with students' chosen sub-fields and interests. Writing teachers could consider trying similar DIY corpus activities in their own classrooms.

Pedagogical Example 2 illustrates an activity focusing on how a DIY corpus approach could be used in a general ESL writing class where students may work in various different settings, and a one-size fits all approach may not be ideal.

Pedagogical Example 2 Text Collection Homework for DIY Corpus Activities

One of the first steps in DIY corpus activities is helping students collect texts for their self-compiled corpora. The sample homework assignment below instructs students to start collecting texts that could be used for in-class activities later.

Homework Directions: In the next couple of weeks, you will be collecting samples of exemplar writing from your work place. You will bring these samples to class to analyze. For this assignment, you will start considering what kinds of texts would be most helpful for you to analyze and collecting samples.

Questions

1. What types of writing do you need to do at work?
   _____

2. Among the genres you listed in question 1, which genres are most important to your job performance?
   _____

3. Do you feel like you need to improve your writing in one of the genres you listed (you could ask your supervisors if you are not sure)?
   _____

4. Based on your answers to the previous questions, which work-related writing genre would you like to focus on during the next unit in class?
   _____

Collecting Texts

After deciding what genre of workplace writing you will focus on, start collecting quality samples to analyze in class. **Find ten examples of good writing in that genre and bring the files on your computer to class** (you will find more later to have a larger collection of examples). In class, we will convert them to text files together and learn how to analyze them. Use the following criteria when choosing samples:

- Choose examples of good writing (consider asking your supervisors for guidance).
- Make sure the writing belongs to your focus genre.
- Pick examples that have topics that you often need to write about.

5.4 Conclusion and Future Directions

There are, of course, challenges inherent in the use of corpora and corpus tools for direct instruction. Chief amongst these are corpus availability, the effort required to decipher search results, the difficulty involved in formulating effective search parameters, data overwhelm, and the often daunting task of interpreting, evaluating, and contextualizing corpus data (Ädel 2010; Anthony 2019). Flowerdew (2017) cites Swales' admonition not to discount the merits of 'old-fashioned', manual discourse analysis methods. Römer (2010) likewise cautions that not all corpora and corpus tools will be useful to developing writers and adds that careful and repeated training is required to render the maximum value from corpus-based instruction. As Dolgova and Mueller (2019) discovered, without proper guidance, students may come to believe that a corpus will immediately provide them with ready answers to their questions when, in fact, students must contextualize, synthesize and apply their findings to their own writing. Furthermore, even if students eagerly embrace the use of corpora and corpus tools as a means of developing their academic writing and successfully navigate the challenges described above, there is no guarantee that they will continue to utilize corpus-based approaches beyond the classroom.

Despite the potential challenges, there is cause to believe that highly motivated, especially advanced students who have taken the time and effort to collect field-specific DIY corpora will carry the use of their corpora forward into future academic and professional writing (Zhang et al. 2017). Even a drop-off in corpus use may not indicate an outright rejection of the tool and its value. Instead, students may simply be at a point in their work where using or adding to their corpora may not be feasible. When the time comes to write their papers, most will recall the sense of empowerment and accomplishment that using corpora gave them, and they will once again employ corpus strategies to produce effective and informative contributions to their discourse communities.

References

Ädel, Annelie. 2010. Using Corpora to Teach Academic Writing: Challenges for the Direct Approach. In *Corpus-Based Approaches to English Language Teaching*, ed. Mari Carmen Compoy-Cubillo, Begoña Belles-Fortuño, and Maria Lluïsa Gea-Valor, 39–55. London and New York: Continuum International Publishing Group.

Alangari, Manal, Sylvia Jaworska, and Jacqueline Laws. 2020. Who's Afraid of Phrasal Verbs? The Use of Phrasal Verbs in Expert Academic Writing in the Discipline of Linguistics. *Journal of English for Academic Purposes* 43: 100814.

Ansarifar, Ahmad, Hesamoddin Shahriari, and Reza Pishghadam. 2018. Phrasal Complexity in Academic Writing: A Comparison of Abstracts Written by Graduate Students and Expert Writers in Applied Linguistics. *Journal of English for Academic Purposes* 31: 58–71.

Anthony, Laurence. 2019. Tools and Strategies for Data-Driven Learning (DDL) in the EAP Writing Classroom. In *Specialised English*, ed. Ken Hyland and Lillian L.C. Wong, 179–194. Abingdon: Routledge.

Aull, Laura. 2015. *First-Year University Writing: A Corpus-Based Study with Implications for Pedagogy*. New York: Palgrave Macmillan.

Biber, Douglas. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam and Philadelphia: John Benjamins.

Biber, Douglas, and Randi Reppen. 1998. Comparing Native and Learner Perspectives on English Grammar: A Study of Complement Clauses. In *Learner English on Computer*, ed. Sylviane Granger, 145–158. London and New York: Routledge.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, and Marie Helt. 2002. Speaking and Writing in the University: A Multidimensional Comparison. *TESOL Quarterly* 36 (1): 9–48.

Biber, Douglas, Randi Reppen, and Eric Friginal. 2010. Research in Corpus Linguistics. In *The Oxford Handbook of Applied Linguistics*, ed. Robert B. Kaplan, 2nd ed., 548–570. Oxford: Oxford University Press.

Birhan, Amare Tesfie, Mulugeta Teka, and Nibret Asrade. 2021. Effects of Using Corpus-Based Instructional Mediation on EFL Students' Academic Writing Skills Improvement. *Theory and Practice of Second Language Acquisition.* 7 (2): 133–153.

Chang, Ji-Yeon. 2015. A Comparison of the First-Person Pronoun *I* in NS and Korean NNS Corpora of English Argumentative Writing. *English Teaching* 70 (2): 83–106.

Charles, Maggie. 2018a. Corpus Tools for Writing Students. In *The TESOL Encyclopedia of English Language Teaching. Teaching Writing*, ed. Diane Belcher and Alan Hirvela. Hoboken: Wiley. https://doi.org/10.1002/9781118784235.eelt0554

———. 2018b. Using Do-It-Yourself Corpora in EAP: A Tailor-Made Resource for Teachers and Students. *Journal of Teaching English for Specific and Academic Purposes* 6 (2): 217–224.

Chen, Qi, and Guang-chun Ge. 2007. A Corpus-Based Lexical Study on Frequency and Distribution of Coxhead's Academic Word List Word Families in Medical Research Articles. *English for Specific Purposes* 26 (4): 502–514.

Chen, Meilin, John Flowerdew, and Laurence Anthony. 2019. Introducing In-Service English Language Teachers to Data-Driven Learning for Academic Writing. *System* 87. https://doi.org/10.1016/j.system.2019.102148.

Cortes, Viviana. 2018. Corpus Tools for Writing Teachers. In *The TESOL Encyclopedia of English Language Teaching. Teaching Writing*, ed. Diane Belcher and Alan Hirvela. Hoboken: Wiley. https://doi.org/10.1002/9781118784235.eelt0553

Crosthwaite, Peter, Lisa Cheung, and Feng Jiang. 2017. Writing with Attitude: Stance Expression in Learner and Professional Dentistry Research Reports. *English for Specific Purposes* 46: 107–123. https://doi.org/10.1016/j.esp.2017.02.001.
[Crossref]

Davies, Mark. 2008–. *The Corpus of Contemporary American English (COCA): 520 Million Words, 1990–Present.* https://corpus.byu.edu/coca/.

Deshors, Sandra C. 2015. A Multifactorial Approach to Gerundial and to-Infinitival Verb-Complementation Patterns in Native and Non-Native English. *English Text Construction* 8 (2): 207–235. https://doi.org/10.1075/etc.8.2.04des.
[Crossref]

Dolgova, Natalia, and Charles Mueller. 2019. How Useful are Corpus Tools for Error Correction? Insights from Learner Data. *Journal of English for Academic Purposes* 39: 97–108.

Dong, Jihua, and Xiaofei Lu. 2020. Promoting Discipline-Specific Genre Competence with Corpus-Based Genre Analysis Activities. *English for Specific Purposes* 58: 138–154.

Durrant, Philip, Mark Brenchley, and Lee McCallum. 2021. *Understanding Development and Proficiency in Writing. Quantitative Corpus Linguistic Approaches*. Cambridge and New York: Cambridge University Press.

Flowerdew, Lynne. 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): 489–511.

———. 2015. Using C-Based Research and Online Academic Corpora to Inform Writing of the Discussion Section of a Thesis. *Journal of English for Academic Purposes* 20: 58–68.

Flowerdew, John. 2017. Corpus-Based Approaches to Language Description for Specialized Academic Writing. *Language Teaching* 50 (1): 90–106.

Friginal, Eric. 2013. Developing Research Report Writing Skills Using Corpora. *English for Specific Purposes* 32 (4): 208–220.

———. 2018. *Corpus Linguistics for English Teachers: New Tools, Online Resources, and Classroom Activities*. New York and London: Routledge.

Friginal, Eric, and Ashleigh Cox. 2022. Corpus Uses in Language Teaching. In *Handbook of Second Language Teaching and Learning*, ed. Eli Hinkel. New York: Routledge.

Friginal, Eric, and Jack A. Hardy. 2014. *Corpus-Based Sociolinguistics: A Guide for Students*. New York: Routledge.

Friginal, Eric, Man Li, and Sara Weigle. 2014. Revisiting Multiple Profiles of Learner Compositions: A Comparison of Highly Rated NS and NNS Essays. *Journal of Second Language Writing* 23 (2): 1–16.

Friginal, Eric, Peter Dye, and Matthew Nolen. 2020. Corpus-Based Approaches in Language Teaching: Outcomes, Observations, and Teacher Perspectives. *Boğaziçi University Journal of Education* 37 (1): 43–68.

Gilmore, Alex. 2009. Using Online Corpora to Develop Students' Writing Skills. *ELT Journal* 63 (4): 363–372.

Granger, Sylviane. 2015. Contrastive Interlanguage Analysis: A Reappraisal. *International Journal of Learner Corpus Research* 1 (1): 7–24. https://doi.org/10.1075/ijlcr.1.1.01gra.
[Crossref]

Gray, Bethany E. 2011. *Exploring Academic Writing through Corpus Linguistics: When Discipline Tells Only Part of the Story*. PhD Dissertation, English Department, Northern Arizona University. https://www.proquest.com/dissertations-theses/exploring-academic-writing-through-corpus/docview/918227058/se-2?accountid=11226.

Hartig, Alissa J., and Xiaofei Lu. 2014. Plain English and Legal Writing: Comparing Expert and Novice Writers. *English for Specific Purposes* 33 (1): 87–96.

Hiltunen, Turo. 2016. Passives in Academic Writing: Comparing Research Articles and Student Essays Across Four Disciplines. In *Corpus Linguistics on the Move*, ed. María José López-Couso, Paloma Núñez-Pertejo, Bélen Méndez-Naya, and Ignacio Palacios-Martínez, 132–157. Amsterdam: Rodopi.

Hyland, Ken. 2012. *Disciplinary Identities: Individuality and Community in Academic Discourse*. Cambridge: Cambridge University Press.

Jafarpour, Ali Akbar, Mahmood Hashemian, and Sepideh Alipour. 2013. A Corpus-Based Approach toward Teaching Collocation of Synonyms. *Theory & Practice in Language Studies* 3 (1): 51–60.

Lee, David, and John Swales. 2006. A Corpus-Based EAP Course for NNS Doctoral Students: Moving from Available Specialized Corpora to Self-Compiled Corpora. *English for Specific Purposes* 25 (1): 56–75.

Lee, Joseph, Tetyana Bychkovska, and James D. Maxwell. 2019. Breaking the Rules? A Corpus-Based Comparison of Informal Features in L1 and L2 Undergraduate Student Writing. *System* 80: 143–153. https://doi.org/10.1016/j.system.2018.11.010.
[Crossref]

Leńko-Szymańska, Agnieszka. 2015. Abundance, However, Is Not Always Desirable: Connectors in Polish EFL Learners' Texts. In *Productive Foreign Language Skills for an Intercultural World. A Guide (not only) for Teachers*, ed. Michal B. Paradowski, 237–253. Frankfurt: Peter Lang.

Liou, Hsien-Chin. 2019. Learner Concordancing for EFL College Writing Accuracy. *English Teaching & Learning* 43 (2): 165–188.

Liu, Dilin. 2012. The Most Frequently-Used Multi-Word Constructions in Academic Written English: A Multi-corpus Study. *English for Specific Purposes* 31 (1): 25–35.

Lu, Xiaofei, and Haiyang Ai. 2015. Syntactic Complexity in College-Level English Writing: Differences among Writers with Diverse L1 Backgrounds. *Journal of Second Language Writing* 29: 16–27. https://doi.org/10.1016/j.jslw.2015.06.003.
[Crossref]

Lu, Xiaofei, J. Elliot Casal, and Yingying Liu. 2021. Towards the Synergy of Genre-and Corpus-based Approaches to Academic Writing Research and Pedagogy. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)* 11 (1): 59–71.

Mansour, Deena. 2017. Using COCA to Foster Students' Use of English Collocations in Academic Writing. In *Proceedings of the 3rd International Conference on Higher Education Advances*, ed. Domenech i de Soria, Josep, Vincent Vela, Maria Cinta, de la Poza, Elena and Desamparados Blazquez. Valencia: Editorial Universitat Politècnica de València, 600–607.

Nasution, Dewi Kesuma. 2018. Corpus Based-Approach in Enhancing Students' Academic Writing Skill: Its Efficacy and Students' Perspectives. *International Journal* 6 (2): 210–217.

Reppen, Randi. 2016. Enhancing Language Teaching: How Corpus Linguistics Can Help. *Corpus Linguistics Research* 2: 25–32.

Römer, Ute. 2009. English in Academic: Does Nativeness Matter? *Anglistik: International Journal of English Studies* 20 (2): 89–100.

———. 2010. Using General and Specialized Corpora in English Language Teaching: Past, Present and Future. In *Corpus-Based Approaches to English Language Teaching*, ed. Mari Carmen Compoy-Cubillo,

Begoña Belles-Fortuño, and Maria Lluïsa Gea-Valor, 18–35. London and New York: Continuum International Publishing Group.

Römer, Ute, Viviana Cortes, and Eric Friginal. 2020. Introduction: Advances in Corpus-Based Research on Academic Writing. In *Advances in Corpus-Based Research on Academic Writing: Effects of Discipline, Register, and Writer Experience*, ed. Ute Römer, Vivianna Cortes, and Eric Friginal, 1–6. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Salazar, Danica. 2014. *Lexical Bundles in Native and Non-Native Scientific Writing: Applying a Corpus-Based Study to Language Teaching*. Vol. 65. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Smith, Simon. 2020. DIY Corpora for Accounting & Finance Vocabulary Learning. *English for Specific Purposes* 57: 1–12.

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

———. 2004. *Research Genres: Exploration and Applications*. Cambridge and New York: Cambridge University Press.

Teubert, Wolfgang. 2005. My Version of Corpus Linguistics. *International Journal of Corpus Linguistics* 10 (1): 1–13.

Upton, Thomas A., and Ulla Connor. 2001. Using Computerized Corpus Analysis to Investigate the Textlinguistic Discourse Moves of a Genre. *English for Specific Purposes* 20 (4): 313–329.

Walker, Crayton. 2011. How a Corpus-Based Study of the Factors which Influence Collocation Can Help in the Teaching of Business English. *English for Specific Purposes* 30 (2): 101–112.

Yoon, Choongil. 2011. Concordancing in L2 Writing Class: An Overview of Research and Issues. *Journal of English for Academic Purposes* 10 (3): 130–139. https://doi.org/10.1016/j.jeap.2011.03.003.
[Crossref]

Zhang, Feng, Yuanhua Zheng, and Li Li. 2017. Using Medical Academic English Corpus for Postgraduates Students Academic Writing Training. *Theory and Practice in Language Studies* 7 (10): 868–873. https://doi.org/10.17507/tpls.0710.07.
[Crossref]

Zhang, Shaojie, Hui Yu, and Lawrence Jun Zhang. 2021. Understanding the Sustainable Growth of EFL Students' Writing Skills: Differences between Novice and Expert Writers in Their Use of Lexical Bundles in Academic Writing. *Sustainability* 13 (10): 1–7.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_6

6. Corpus Affordances in Foreign Language Reading Comprehension

Alejandro Curado Fuentes[1] ✉
(1) Universidad de Extremadura, Cáceres, Spain

✉ **Alejandro Curado Fuentes**
Email: acurado@unex.es

Keywords DDL – Reading comprehension – Vocabulary – ESP – Digital tools – Concordances – Intermediate levels

6.1 Introduction

Data-Driven Learning (DDL) developments in the Foreign Language classroom result from awareness of corpora applicability in educational settings (see Timmis and Templeton, this volume). DDL challenges learners' abilities to deduce and/or induce linguistic usage and meaning from corpus data. Boulton and Cobb (2017: 348) have analysed 64 different research studies whose data indicate "that DDL approaches result in large overall effects for both control/experimental group comparisons (d = 0.95) and for pre/posttest designs (d = 1.50)". Overall, their meta-analysis demonstrates that DDL applications lead to positive learning outcomes in FL settings. This claim, however, is skewed by, among other observations, the fact that the main linguistic skill exploited in most DDL approaches is writing while "the other skills (…) remain largely underresearched" (Boulton and Cobb 2017: 379). In fact, only three studies surveyed in Boulton and Cobb (2017) deal with reading comprehension, and only one of them (Curado Fuentes 2007)

juxtaposes control and experimental groups with pre- and post-test measurements.

By comparison, the use of corpus tools and information in FL reading comprehension is rare in an already atypical applied corpus linguistics scenario for FL learning: actual practices of corpus-based methods in English as a Foreign Language (EFL) teaching, syllabus design, and teacher training are, indeed, noticeably limited (see Aijmer 2009; Pérez-Paredes 2019). The overall situation in English for Specific Purposes (ESP) is also one where the "application of corpora (…) is a rarity" (Boulton et al. 2012: 1). According to extensive surveys across European schools, there is minimum teaching familiarity with open educational resources, natural language processing tools, and corpus exploitation in language learning at all educational levels (Pérez-Paredes et al. 2018).

The main goal of this chapter is to provide an update of corpus affordances and DDL methods for FL reading comprehension. The first aim is to examine pedagogical issues and procedures showcased at the crossroads of reading skills, ESP, EFL, and corpus advancements. The second aim is to analyse two ESP courses where corpus and non-corpus technologies (e.g., online glossary-making and interactive quizzes) were deployed for vocabulary and reading comprehension (see Selivan, this volume). The informants were business administration and tourism students at University of Extremadura, Spain. Reading comprehension and vocabulary exercises were combined during the sessions, as "the size of the reader's vocabulary knowledge is (…) a consistently strong predictor of the level of text comprehension" (Paribakht and Webb 2016: 124), and "small increments of vocabulary knowledge contribute to reading comprehension" (Laufer and Ravenhorst-Kalovski 2010: 15). The students' linguistic gain was tested via pre- and post-tests, and their perceptions collected by short questionnaires and interviews. This mixed-method approach afforded a more learner-centred coverage in reading assessment.

After this introduction, previous research on reading skills and corpora is described; then, the rationale for the design of our case study and the study implementation are explained in the Methodology section. This is followed by the analysis of the students' performance and their perceptions about DDL in the Results section. Finally, the

Conclusions (Sect. 6.5) specify key learning benefits of DDL for reading in our context, and we suggest future needed lines of work on DDL and FL reading comprehension skills.

6.2 Reading Skills and Corpora

In this section, FL written reception is reviewed in connection with corpora and DDL. Text and task authenticity, lexical knowledge, the role of text types and registers, and the fact that reading is treated as a secondary skill in the corpus literature are salient facets. Authenticity is generally described in terms of real texts and purposeful learning approaches (e.g., to gain content and linguistic knowledge from real text material). In ESP/EAP (English for Academic Purposes), the exploitation of authentic texts from representative academic corpora with DDL methods is considered a useful didactic focus (e.g., Aston 2001; Gavioli 2005). Corpus-based analyses of authentic written material are also important for reading comprehension assessment and test design (e.g., Gilmore 2009; Webb and Paribakht 2015). By contrast, other authors (e.g., Allan 2008; Hadley and Charles 2017) argue that task authenticity can be achieved with modified/graded readers, best suited for lower proficiency learners in EFL contexts which may profit more from teacher-guided linguistic explorations with simplified reading corpora (Hadley and Hadley 2021).

The focus on vocabulary is a major direction in EFL/ESP corpus-based explorations of texts. Lexical knowledge requirements for text comprehension are often observed in terms of corpus-based word frequency bands (e.g., Nation 2006; Laufer and Ravenhorst-Kalovski 2010). Academic/professional reading complexity is evaluated by exploring normalized lexical item/family frequencies in texts (e.g., Kubát and Milička 2013; Coxhead and Boutorwick 2018). Vocabulary profilers, featured in various corpus tools (e.g., Compleat Lexical Tutor, Versatext, and Skell, among others), can be used to describe vocabulary according to proficiency levels and text coverage. Furthermore, vocabulary can be explored with DDL approaches in different ESP/EAP scenarios (e.g., Cobb 1999, 2018). These approaches to students' lexical competence include concordance-based activities and pre-/post-tests to evaluate whether substantial lexical gains are obtained after

examining concordances. Delayed post-testing (Cobb 1999) also evinces lexical knowledge improvement after concordancing subject content texts (e.g., course textbooks, lab reports, and so on).

In ESP/EAP, reading capabilities involve the recognition and understanding of text type and register features. This view points to a sensible interaction of lexico-grammatical elements with macro-structural features in texts (e.g., Carrell 1987; Carrell and Carson 1997; Gavioli 1997). The dynamics of micro-/macro-structural interaction tap into the exploitation of bottom-up and top-down text decoding abilities in DDL (Johns 1994; Cobb 2018). Learning developments indeed occur without the learners' direct realisation that they are reading a text, as 'sneaked-in formulas' (Cobb 2018: 200) are set up, linking textual features with concordance data that students require to complement meaning in those texts inductively. In contrast, more deductively, "in the context of a reading activity, a teacher can use a corpus tool to determine which repeated strings in a text they should draw learners' attention to" (Cobb 2018: 201). Both inductive and deductive styles are cardinal in DDL and beneficial in language education and teaching instruction (e.g., Boulton and Tyne 2014; Lee and Lin 2019). Related to inductive skills is the so-called hard DDL approach, which involves more explicit and direct practice with hands-on concordances. In turn, semi-hard and soft DDL strategies combine teacher explanations that point learners to specific linguistic chunks in the concordances, shown on screen or in printed form (e.g., Hadley and Charles 2017; Hadley and Hadley 2021).

A last but not least important observation in the corpus literature is that reading comprehension is treated as a secondary/minor linguistic skill. As Tribble (2002), Gilmore (2009), Cobb and Boulton (2015), and Boulton and Cobb (2017), among others, observe, most teacher-researchers value writing as the primary linguistic skill to exploit with corpora, and consider reading an intermediate endeavour towards production abilities. In fact, DDL studies focusing on academic writing get published more in Computer-Assisted Language Learning (CALL)-related journals (Pérez-Paredes 2019). In highly ranked ESP-related journals, a similar phenomenon occurs. For instance, in *English for Specific Purposes*, no studies have explored DDL for reading comprehension, and only a handful of articles examine reading

strategies with experimental groups. The most recent (to date) Teaching and Language Corpora (TALC) conference (2020) corroborates this trend. There, out of 74 contributions, only eight address reading issues, although not as their primary concern; other skills/competences are (e.g., three focus on writing/translating, two on vocabulary, and three others on listening and project writing).

By contrast, some authors focus on reading skills as a major linguistic competence and examine the potential benefit of DDL for text comprehension; for example, Ballance (2021) considers that concordances provide distinct narrow reading conditions, and Hadley and Hadley (2021) conclude that DDL can assist extensive reading. Nonetheless, corpus studies on FL reading are scarce. Figure 6.1 shows their scattered distribution over the past 15 years. The sources have been searched for in journals such as *English for Specific Purposes*, *System*, *Journal of English for Academic Purposes*, *Language Learning & Technology*, *Computer-Assisted Language Learning*, and *ReCALL*. In addition, some other works have been retrieved from the first 50 pages returned by Google Scholar after using terms such as *corpus*, *DDL*, *EFL*, *ESP, reading*, *text authenticity*, and *texts*. Publications in conference proceedings, minor journals, and non-international outlets are excluded from the tally.

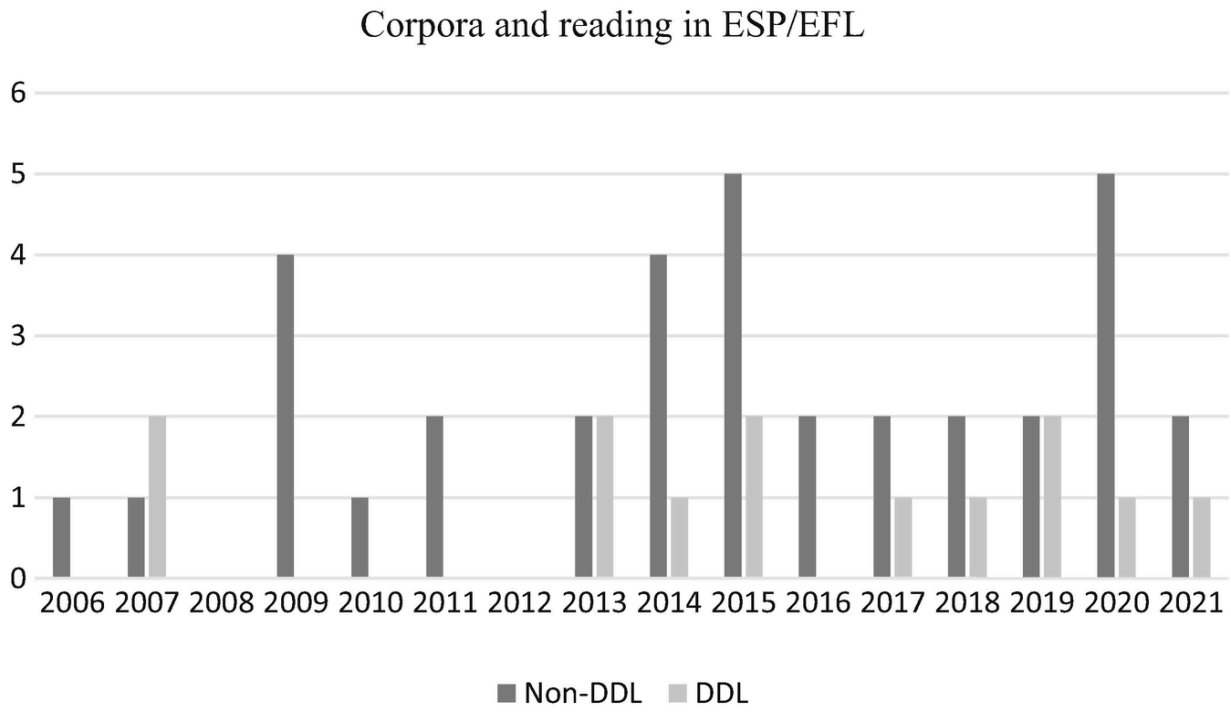

Fig. 6.1 Number of works dealing with corpora and reading including and excluding DDL

The studies embracing DDL are even more reduced than general corpus-based approaches (27% of all corpus studies found). They mostly belong to Asian and Middle Eastern academic contexts. This regional distribution coincides with a recent meta-analysis on DDL studies conducted by Boulton and Vyatkina (2021). In general, in these works, the authors are EFL teachers who find that their learners are insufficiently prepared for written text comprehension, and thus, they embark on DDL experimentation in their teaching situations. A probable cause for reading deficiency is the increased use of out-of-class digital audio-visual content to the detriment of text reading (EF English Proficiency Index Report 2019: 6). This reading competence weakness is observed by Curado Fuentes and Edwards-Rokowski (2007), Curado Fuentes (2015), Wang et al. (2015), and Xu et al. (2019), whose DDL approaches involve learners' decoding of specific lexical, grammatical, discourse, and thematic features in texts. Vocabulary is also a key focus in most text comprehension studies using concordance data (e.g., Chen and Huang 2014; Lee et al. 2019), and other works target learners' attitudes and impressions with DDL by comparing this method with other learning procedures during reading modules/courses (e.g., Gordani 2013; Hadley and Charles 2017).

6.3 Methodology

6.3.1 Participants

This analysis is based on a survey of 40 ESP students in the double degree of business administration and tourism at our university. Learners belong to two courses/groups: 22 to first-year English (Year 1: Y1), and 18 to fourth year (Year 4: Y4). Their average age is 18 in Y1, and 20.2 in Y4. Most learners are female (15 and 11 respectively). At the beginning of the Spring semester (2021), Y1 students answered a standardised level test on the Cambridge Assessment English website (https://www.cambridgeenglish.org/test-your-english/general-english/), and Y4 participants took the British Council online test (https://learnenglish.britishcouncil.org/online-english-level-test). Their scores indicate that intermediate (B1) levels prevail (in 67.5% of the cases); other levels are B2 (five students in Y4 and two in Y1), and

A2 (six participants in Y1). This mixed proficiency scenario, common in our courses, is not suited for hard DDL practice (i.e., autonomous inductive work with concordances); therefore, semi-hard DDL (more teacher-centred guidance, explanations, and examples with concordances) has been adapted, especially in Year 1.

Our learners generally favour a focus on listening and speaking practice, group work, and digital tools in EFL according to their answers to preliminary questionnaires at the beginning of the semester. Few informants see the role of specific vocabulary knowledge and reading skills as a priority, although the percentage of students who value written texts increases by 17.7% in Y4. Y1 students' last EFL experience was secondary education, where 65% of learners perceived that their primary focus had been on grammar, vocabulary, and reading/writing.

In Y1, a significant section of our course is devoted to specific linguistic-communicative aspects of ESP, e.g., polite discourse in academic/professional situations, specific vocabulary and phrases to convey formality in written and oral texts, and so on. By contrast, Y4 students have already had three ESP courses at university, including the Y1 course. The Y4 subject (Communicative skills in English) differs from the previous courses in terms of a greater focus on PBL (Project-Based Learning) as opposed to more teacher-directed interactions in Y1. Y4 students are more accustomed to independent work throughout the academic term. Therefore, there is no final written exam, and their final marks derive from continuous autonomous/collaborative project developments.

6.3.2 Corpus and Non-corpus Resources Used

Corpus tools and other digital technologies/applications can lead to learning motivation and productive developments in the EFL classroom (e.g., Golonka et al. 2014). An adequate selection of language learning tools requires a clear view and understanding of the learning goals that can be accomplished with those resources (Loucky 2009; Chen and Huang 2014). It is also important to choose technologically updated tools that can adapt more flexibly to users' different device types (IDEAL Desktop Project 2020).

The IDEAL project, financed under the Erasmus+ programme ((http://platform.ideal-project.eu/open-resources), provides a wide range of tools for language learning. IDEAL operates under the Digital Competence Framework for Educators (DigCompEdu, cf. Punie and Redecker 2017) to examine adult language teachers' knowledge of and competences with digital resources. It is important, in view of the results and input provided by focus and expert groups in IDEAL, to raise "awareness on the potential impact which ICT can have on teachers' competences and the overall quality of teaching, the pedagogical benefits of the use of digital technology for language students and teachers" (IDEAL Desktop Project 2020: 15). Up to nine different digital skills are described by IDEAL for FL teaching and learning. In our case, 'reading reception', 'mediating a text', and 'written and online interaction' have been selected as parameters to search for likely candidate tools. It is found that KAHOOT, Playposit, ACADLY, and Quizlet (non-corpus tools), and FLAX, Versatext, and Compleat Lexical Tutor (Compleat) (corpus resources) are suitable for our teaching aims.[1] All the sessions with the tools took place in the computer lab to ensure that all learners accessed computers individually, since not all students own/can afford laptops. Furthermore, all students regularly bring smartphones to class, which facilitates access to some tools.

A teacher-made 10,329-word electronic corpus was used in the two courses for learner engagement with DDL and reading material. It consists of 20 web-retrieved texts on product innovation, creativity, and marketing. Its vocabulary corresponds to three frequency bands, identified with Vocabulary Profiler (VP) in Compleat, one of the corpus tools implemented in our Y1 class (examined below). 88.4% of the tokens from our corpus appear in VP's top two frequency bands of 1000 and 2000 word-families, 6.35% of tokens in the Academic Word List (AWL), and 5.3% in off-lists (e.g., proper names, acronyms, and so on). These figures were not considered a yardstick of reading proficiency (e.g., Laufer and Ravenhorst-Kalovski 2010) but a reference for lexical complexity in the texts. Thus, this lexical range was taken as suitable for our reading comprehension aims.

6.3.3 Implementation of the Study

The topics exploited in both years for reading comprehension are innovation, creativity, and advertising/marketing. Y4 students read longer texts, especially on marketing, whereas Y1 learners approached more introductory texts on innovation. The sessions lasted four weeks (four hours per course and week) in March and April of 2021. Table 6.1 displays the scheduled sessions, activities, and tools. In terms of the corpus tools used, all three are web-based: Compleat (only included in Y1), and FLAX and Versatext (in Y4). Three non-corpus tools were deployed in the two courses: KAHOOT, Quizlet, and Playposit, and one resource (ACADLY) was integrated in Y1 only.

Table 6.1 Scheduled activities and tasks in each course (* activities carried out in pairs or/and groups)

Time scale	Specific English I / Y1	Communicative skills in English / Y4
Week 1 (includes pre-test)	1.—Pre-class: Poll / discussion (ACADLY) 2.—Introducing DDL (Compleat) 3.—Vocabulary and reading quiz (ACADLY)	1.—Pre-class: Poll / discussion (KAHOOT) 2.—Introducing DDL (FLAX) 3.—Listening quiz (KAHOOT)
Week 2	1.—Pre-class: Vocabulary quiz (ACADLY) 2.—Reading and listening (Compleat)* 3.—Concordances (Compleat)* 4.—Listening quiz (KAHOOT)	1.—Pre-class: Vocabulary (FLAX) 2.—Lexical activities (FLAX)* 3.—Glossary-making (FLAX and Moodle) 4.—Poll (KAHOOT)
Week 3	1.—Pre-class: Noun combinations and reading (ACADLY) 2.—Listening / reading (Playposit) 3.—Poll (KAHOOT) 4.—Vocabulary quizzes (KAHOOT) 5.—Concordance quizzes (Compleat)*	1.—Pre-class: Glossary making (continued on FLAX) 2.—Keywords in texts (Versatext)* 3.—Vocabulary quizzes (KAHOOT) 4.—Explanations of key words and collocates (Quizlet)*
Week 4 (includes post-test)	1.—Pre-class: Reading quiz (ACADLY) 2.—Vocabulary (Quizlet)* 3.—Poll and discussion (ACADLY)	1.—Pre-class: Key words continued (Quizlet)* 2.—Listening and reading (Playposit) 3.—Poll (KAHOOT) 4.—Vocabulary review (Versatext)*

The pre- and post-tests (see Appendix 1 and 2) present the same outline: ten questions on vocabulary (five on collocations, three on definitions, and two on synonyms), and five comprehension questions on a short text. The main reading goal was the understanding of "the main ideas of complex text on both concrete and abstract topics", characteristic of B2 (CEFR 2018: 24). The individual pre-class activities during the sessions focused on learners' lexical pre-knowledge. During classes, the main developments made with the tools were interactive, e.g., concordancing (cf. Fig. 6.2), guessing from context, competing in games, task completion, and participation in quizzes, polls, and discussions.

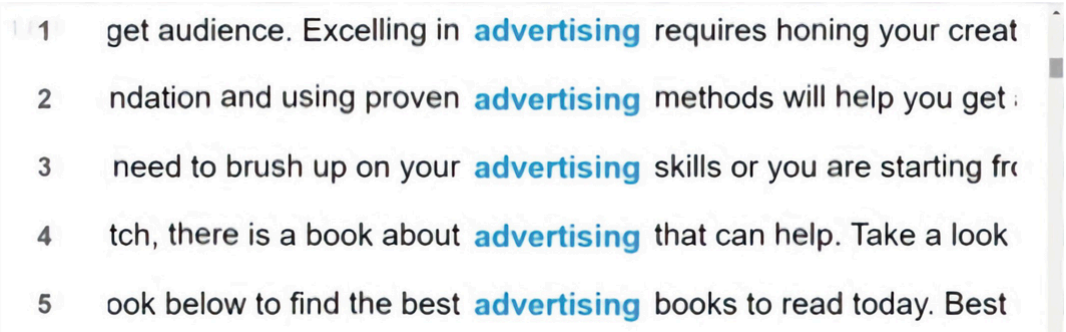

Fig. 6.2 Concordance reading of "advertising" to discern it as either an adjective or a noun

In the first Y1 class, the lecturer explained the corpus tool (Compleat). The main aim was to check some words and phrases in the texts to explore lexical-grammatical meaning and usage. Then, two sections in Compleat were shown: Hypertext to read and listen to texts online, and Web/Text Concordance to explore key word behaviour in context. All the steps, procedures, and examples were explicitly shown and explained on screen: in Hypertext, we chose Hypertext II and listened to a short text (from our ad hoc corpus) about product innovation strategies, read by the text-to-speech plug. Then, we checked some key content words, and learners recognized and deduced lexical meaning and usage from the collocations that the teacher explicitly highlighted on screen. At the end of Week 3, we used Compleat again to recall the use of key words through the Brown corpus-based MultiConc utility (in Hypertext). This application generates concordance lines that display gaps where the targeted node words go. However, this guessing exercise barely motivated

participation, and in most cases, the teacher ended up solving and explaining the answers. The main cause was the linguistic difficulty found in the Brown corpus output, with many words and expressions unknown to most Y1 learners.

Y4 students' exposure to and interaction with corpus tools were less teacher-dependent, as the students had already used FLAX before. FLAX is a corpus tool mainly used to find academic key words and collocates, classified according to word categories. Y4 participants selected 10 content words from our corpus-based frequency list (provided by the instructor), and then explored collocates within word categories in Social Sciences (a subject area included in FLAX). The activity was then reviewed in class, where students worked in groups to translate the meaning of the collocations and phrases found. Each student created lexical entries in a glossary-making task. Then, they uploaded the results on the Moodle platform. During Week 2, they also used an exercise-creation section in FLAX to design Matching and Gapping activities with word sets. The procedures to create collocation exercises were shown and explained by the instructor. Students, working in groups, were told to select the information from corpus-driven collocations in FLAX. Once finished, other groups attempted to solve their classmates' activities. Students generally enjoyed this section in FLAX and it took them almost one whole two-hour class to complete and share their lexical quizzes.

Y4 students also used Versatext, another web-based corpus tool. They uploaded three texts from our home-made corpus to generate word clouds and concordances in Versatext. The concordance lines were few and clear, and so, key word meanings and usage were easily deduced. Some learners also induced lexico-grammatical aspects, e.g., *advertising* as either an adjective or a noun (Fig. 6.2). Versatext was explored again to review key concepts and vocabulary in Week 4. Three drag-and-drop quizzes were assigned on the Moodle platform, where the students built sets of collocations after checking concordances with the link provided in the activity (Fig. 6.3).

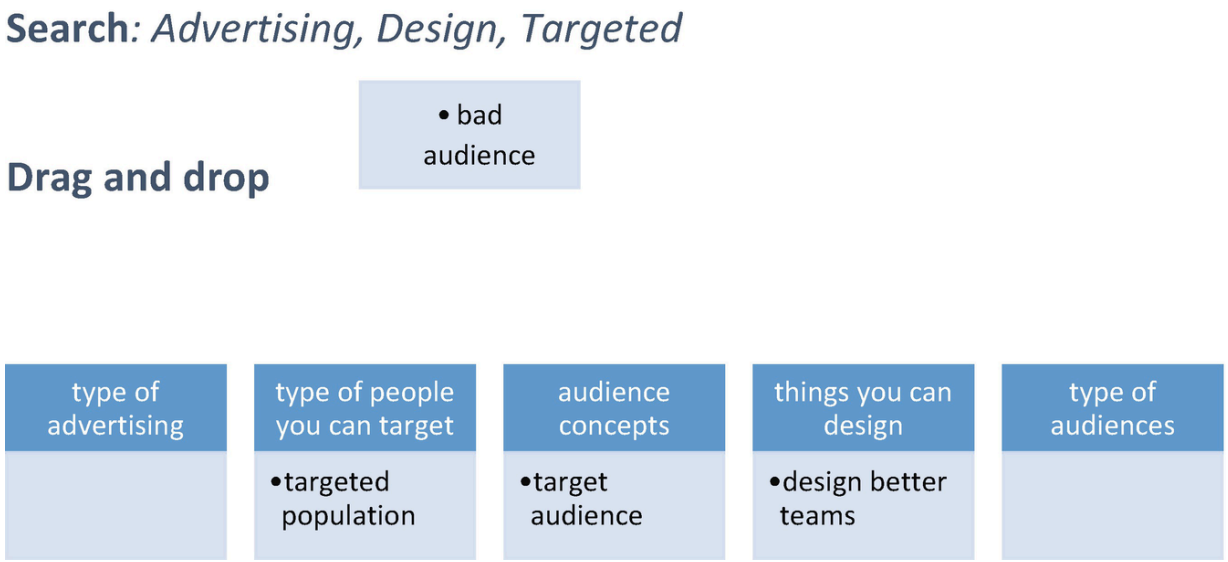

Fig. 6.3 Drag-and-drop activity with collocations derived from the concordances

The non-corpus tools were deployed intermittently throughout the sessions, as shown in Table 6.1 above. In Y1, ACADLY was used as a class organizing utility with which students could access pre- and in-class assignments as a complement to reading material. Learners also answered comprehension questions on texts and vocabulary quizzes. In Y1 and Y4, KAHOOT was exploited to introduce and review thematic concepts, key words, and collocations (with polls and quizzes). Most students in the two courses accessed ACADLY and KAHOOT with their smartphones.

Both groups used Quizlet to build word sets: Y1 learners picked five phrasal verbs from the ad hoc corpus wordlist and translated their meanings into Spanish; Y4 participants elaborated collocation lists in Quizlet, based on previous DDL activities, and later, the students were divided into groups to guess those collocations in oral contests. Finally, Playposit allowed learners to interact with home-made videos using their smartphones. The videos featured questions on key vocabulary examined in previous DDL activities.

6.4 Results

Students' pre-test and post-test scores were compared. The computation of scores was separated into vocabulary and reading comprehension. As a first inquiry, we wondered whether there was a significant difference between pre-test and post-test data for each

section, and applied a Wilcoxon signed-rank test to verify this (Table 6.2). This type of test was selected because it works well to compare mean values when the dependent variable is not normally distributed, and to find out if groups of scores lower than N=30 (minimum number for parametric tests) were overtly different.

Table 6.2 Statistical significance in test comparisons

Test sections being compared	Wilcoxon signed-rank test (intra-group scores)
Vocabulary	1st year (pre / post-tests): 0.00001 at $p \leq 0.05$ 4.0145 at $z \geq 1.96$ 4th year (pre/post-tests): 0.0003 at $p \leq 0.05$ 3.6214 at $z \geq 1.96$
Reading comprehension	1st year (pre / post-tests): 0.00094 at $p \leq 0.05$ 3.3137 at $z \geq 1.96$ 4th year (pre/post-tests): 0.00064 at $p \leq 0.05$ 3.4078 at $z \geq 1.96$

Each student's points were compared (e.g., for vocabulary, S1 obtained 3 in the pre-test and 6 in the post-test, and for reading comprehension, 1 versus 3, and so forth). Except for one Y1 student in the reading comprehension section, the scores increased in the post-tests. One Y1 and one Y4 student had the same scores in both vocabulary tests, and four Y1 and three Y4 participants achieved the same points in the two reading sections. According to the statistical values, the two groups' post-test achievements were significantly better with both sections (vocabulary and reading). Y4 students attained a less meaningful difference with the post-test vocabulary because they had initially obtained higher pre-test scores than Y1.

Additionally, most learners passed a final test on specific vocabulary and reading comprehension (86.5%t in Y1 and 94.5% in Y4), which confirms their learning outcome at the end of the semester.

The second data sets used to evaluate the sessions and tools came from post-session questionnaires and interview responses. The participants' numerical answers to questionnaires are shown in

Appendix 3 according to a 1–5 Likert scale. Learners, especially Y4, manifested their overall satisfaction with the sessions and resources. They judged the use of digital elements and platforms as positive. Most participants named speaking and listening as favourite practice, similar to their pre-course questionnaire inclinations, despite having more text-based activities in the sessions. They also praised working in groups and minimized pre-class work as important.

Regarding the tools, learners preferred ACADLY and KAHOOT. More Y4 students appraised Playposit as interesting, probably because they also used video presentations in their projects. They also rated Versatext as appealing. The reasons put forward by some students about why they liked a tool more were: (1) Fun, attractive (KAHOOT—Playposit); (2) Dynamic, competitive (KAHOOT); (3) Online progress (ACADLY); and (4) Useful for vocabulary (ACADLY—Playposit). In their discernment of less favoured options, three students alleged two particulars: (1) Difficult, messy (Compleat—FLAX); and (2) unattractive interface (FLAX—Quizlet). Y1 learners assigned lower scores to the corpus tool perhaps due to the students' use of a more complex tool and because the class time devoted to it was insufficient. Nonetheless, the text reading/listening activities and explicit teacher-directed concordance exploration with specific vocabulary drew moderate attention and some approving reactions.

In a series of short semi-structured interviews as part of their final oral exams (see Appendix 4), Y4 students approved of the sessions, e-platform, and tools. Even though their face-to-face oral encounters in English with the lecturer likely affected such a positive stance and comments, the interviewer obtained valuable information. First, he asked why they thought the tools were beneficial for learning, and most referred to interactive quizzes and videos as entertaining and effective for vocabulary gain. Most informants reiterated their choices of KAHOOT and video applications. Two students stated that they preferred these tools because they could use them with their mobile phones.

When asked about the corpus tools, five students put forward the usefulness of specific corpora that contain terminology. Two interviewees referred to FLAX as handy for terminology, and three others recalled a business letter corpus as a noteworthy tool to identify

important expressions for writing (outside of this case study). Two students also remarked that the pre-class work with corpora (e.g., first FLAX activity) should be moved to the lab because the activities would be carried out better in groups/pairs. The same two participants' and one other student's answers pointed to learning benefits from creating wordlists and analysing vocabulary in specific texts. When asked about Versatext, 66.6% of students said it was easy to use and good for learning vocabulary. Around 25% appreciated the use of the small ad hoc corpus. The Wordcloud feature, text uploading option, and concordance managing tool in Versatext were also liked by students. In this regard, the processing of lexical items was perceived as productive because it was done for a task purpose, in agreement with Xu et al. (2019).

6.5 Conclusions

In general, it has been shown that corpus and non-corpus tools can be accommodated in ESP/EFL reading comprehension situations. According to our findings, specific text comprehension improves after web-based DDL and digital tools have been deployed with those texts and vocabulary. The commensurate progress found in Y1's and Y4's post-test achievements with both test sections indicates that each group's text comprehension performance parallels their corresponding advancement with vocabulary. Most students also passed final reading comprehension and vocabulary tests, corroborating this learning progress.

We find that exploiting lexical behaviour activities with concordances can lead to learning improvement when this approach is made with authentic texts (in our case, on a specific topic related to learners' studies). Given the positive impressions and comments on the sessions, we conclude that both corpus and non-corpus tools contribute to productive developments with authentic texts/tasks and language. We also observe that learners feel relatively comfortable accessing multiple resources to cope with texts and vocabulary. The full adaptation of some tools to mobile/smartphone devices accentuates these positive reactions. Technologies exploiting DDL should focus on

this, since using difficult interfaces and decontextualized computer-based concordances can put learners off.

As this study has been limited to four weeks, it lacks a longitudinal scope. Exploring DDL affordances over longer terms, comparing learning contexts with experimental/control groups and mixed methods, or including delayed post-testing, are aspects that need further scrutiny (Boulton and Vyatkina 2021: 75). Pre-training learners in inductive skills or "higher-level academic skills" (Hadley and Charles 2017: 145) is also found necessary to enable learners to engage more easily with hard DDL strategies. Corpus-aided approaches to reading comprehension and other linguistic skills should be explored further in homogeneous linguistic settings (i.e., where all learners have similar lexical ranges, grammatical knowledge, etc.). In reading comprehension, this uniformity is important so that most words in texts (98% according to Nation 2006) are recognized. Teachers/researchers should adapt DDL practices to their teaching/learning conditions (e.g., FL proficiencies, learning styles, preferred course materials, digital resources, methods, institutional constraints, and so on.). In this way, learners could feel more motivated and adept at using corpus-driven linguistic feedback in their learning process inside and outside the classroom.

Appendix

Access to pre- / post-tests, questionnaires, and interview questions:
https://drive.google.com/file/d/1sFqDX3EB_KNrhQHCbefHT1W_mk-EJl5m/view?usp=sharing

---

References

Aijmer, Karin, ed. 2009. *Corpora and Language Teaching*. Amsterdam: John Benjamins.

Allan, Rachel. 2008. Can a Graded Reader Corpus Provide 'Authentic' Input? *ELT Journal* 63 (1): 23–32. https://doi.org/10.1093/let/ccn011.
[Crossref]

Aston, Guy, ed. 2001. *Learning with Corpora*. Houston, TX: Athelstan.

Ballance, Oliver James. 2021. Narrow Reading, Vocabulary Load and Collocations in Context: Exploring Lexical Repetition in Concordances from a Pedagogical Perspective. *ReCALL* 33 (1): 4–17. https://doi.org/10.1017/S0958344020000117.
[Crossref]

Boulton, Alex, and Tom Cobb. 2017. Corpus Use in Language Learning: A Meta-analysis. *Language Learning* 67 (2): 1–46.
[Crossref]

Boulton, Alex, and Henry Tyne. 2014. Corpus-based Study of Language and Teacher Education. In *The Routledge Handbook of Educational Linguistics*, ed. Martha Bigelow and Johanna Ennser-Kananen, 301–312. New York: Routledge.

Boulton, Alex, and Nina Vyatkina. 2021. Thirty Years of Data-Driven Learning: Taking Stock and Charting New Directions Over Time. *Language Learning & Technology* 25 (3): 66–89. https://hdl.handle.net/10125/73450.

Boulton, Alex, Shirley Carter-Thomas, and Elizabeth Rowley-Jolivet, eds. 2012. *Corpus-Informed Research and Learning in ESP. Issues and Applications*. Amsterdam: John Benjamins.

Carrell, Patricia L. 1987. ESP in Applied Linguistics: Refining Research Agenda Implications and Future Directions of Research on Second Language Reading. *English for Specific Purposes* 6 (3): 233–244. https://doi.org/10.1016/0889-4906(87)90006-8.
[Crossref]

Carrell, Patricia L., and Joan G. Carson. 1997. Extensive and Intensive Reading in an EAP Setting. *English for Specific Purposes* 16 (1): 47–60.
[Crossref]

CEFR. 2018. *Common European Framework of References for Languages: Learning, Teaching, Assessment*. Strasbourg: Cambridge University Press.

Chen, Pi-Ching, and Chun-Han Huang. 2014. Effects of Integrating Online Concordancer and Online Dictionary on EFL Learners' English Vocabulary Retention. *International Journal of Arts and Commerce* 3 (8): 103–114.

Cobb, Tom. 1999. Breadth and Depth of Lexical Acquisition with Hands-on Concordancing. *Computer Assisted Language Learning* 12 (4): 345–360.
[Crossref]

———. 2018. From Corpus to CALL. The Use of Technology in Teaching and Learning Formulaic Language. In *Understanding Formulaic Language: A Second Language Acquisition Perspective*, ed. Anna Sivanova-Chanturia and Ana Pellicer-Sánchez, 192–211. New York: Taylor & Francis.
[Crossref]

Cobb, Tom, and Alex Boulton. 2015. Classroom Applications of Corpus Analysis. In *The Cambridge Handbook of English Corpus Linguistics*, ed. Douglas Biber and Randi Reppen, 478–497. Cambridge: Cambridge University Press.
[Crossref]

Coxhead, Averli, and T.J. Boutorwick. 2018. Longitudinal Vocabulary Development in an EMI International School Context: Learners and Text in EAL, Maths, and Science. *TEOSL Quaterly* 52 (3): 588–610. https://doi.org/10.1002/tesq.450.
[Crossref]

Curado Fuentes, Alejandro. 2007. A Corpus-based Assessment of Reading Comprehension in English for Tourism Studies. In *Corpora in the Foreign Language Classroom*, ed. Encarnación Hidalgo, Luis Quereda, and Juan Santana, 290–328. Amsterdam: Rodopi.

———. 2015. Exploiting Keywords in a DDL Approach to the Comprehension of News Texts by Lower-Level Students. In *Multiple Affordances of Language Corpora for Data-Driven Learning*, ed. Agnieszka Lenko-Szymanska and Alex Boulton, 177–198. Amsterdam: John Benjamins.

Curado Fuentes, Alejandro, and Patricia Edwards-Rokowski. 2007. Reading Comprehension as a Text / Context Focus in Tourism Discourse. In *Approaches to Specialised Discourse in Higher Education and Professional Contexts*, ed. Alejandro Curado Fuentes, Patricia Edwards-Rokowsi, and Mercedes Rico García, 79–101. Newcastle upon Tyne: Cambridge Scholars Publishing.

EF / English Proficiency Index Report. 2019. Ìndice del EF English Proficiency. https://www.ef.con.co/epi/reports/epi-s/. Accessed 14 August 2021.

Gavioli, Laura. 1997. Exploring Texts through the Concordancer: Guiding the Learner. In *Teaching and Language Corpora*, ed. Anne Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles, 126–143. London: Routledge.

———. 2005. *Exploring Corpora for ESP Learning*. Amsterdam: John Benjamins.
[Crossref]

Gilmore, Alex. 2009. Using On-line Corpora to Develop Students' Writing Skills. *English Language Teaching Journal* 63 (4): 363–372. https://doi.org/10.1093/elt/ccn056.
[Crossref]

Golonka, Ewa M., Anita R. Bowles, Victor M. Frank, Dorna L. Richardson, and Suzanne Freynik. 2014. Technologies for Foreign Language Learning: A Review of Technology Types and Their Effectiveness. *Computer Assisted Language Learning* 27 (1): 70–105. https://doi.org/10.1080/09588221.2012.700315.
[Crossref]

Gordani, Yahya. 2013. The Effect of the Integration of Corpora in Reading Comprehension Classrooms on English as a Foreign Language Learners' Vocabulary Development. *Computer Assisted Language Learning* 26 (5): 430–445. https://doi.org/10.1080/09588221.2012.685078.
[Crossref]

Hadley, Gregory, and Maggie Charles. 2017. Enhancing Extensive Reading with Data-Driven Learning. *Language Learning & Technology* 21 (3): 131–152. https://doi.org/10125/44624.

Hadley, Gregory, and Hiromi Hadley. 2021. Exploring the Impact of Data-Driven Learning in Extensive Reading. In *Beyond Concordance Lines: Corpora in Language Education*, ed. Pascual Pérez-Paredes and Geraldine Mark, 149–176. Amsterdam: John Benjamins.
[Crossref]

IDEAL Desktop Project. 2020. Desktop Research and Needs Analysis on the Mapping of Content for Digitally Competent Language Teachers. https://ideal-project.eu/. Accessed 11 August 2021.

Johns, Tim. 1994. From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning. In *Perspectives on Pedagogical Grammar*, ed. Terence Odlin, 293–313. Cambridge: Cambridge University Press.
[Crossref]

Kubát, Miroslav, and Jîri Milička. 2013. Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics* 20 (4): 339–349.
[Crossref]

Laufer, Batia, and Geke C. Ravenhorst-Kalovski. 2010. Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a Foreign Language* 22 (1): 15–30.

Lee, Pinshuan, and Huifen Lin. 2019. The Effect of the Inductive and Deductive Data-Driven Learning (DDL) on Vocabulary Acquisition and Retention. *System* 81: 14–25. https://doi.org/10.1016/j.system.2018.12.011.
[Crossref]

Lee, Sanghee, Sunghee Kim, and Yangsoo Jung. 2019. Effects of DDL Approach on English Productive and Receptive Vocabulary Knowledge and Reading Ability. *The Journal of Studies in Language* 35 (3): 341–360.

Loucky, John Paul. 2009. Constructing a Roadmap to More Systematic and Successful Online Reading and Vocabulary Acquisition. *Literary and Linguistic Computing* 25 (2): 225–241. https://doi.org/10.1093/llc/fqp039.
[Crossref]

Nation, Ian Stephen Paul. 2006. How Large a Vocabulary is Needed for Reading and Listening? *The Canadian Modern Language Review* 63 (1): 59–82.
[Crossref]

Paribakht, Tahere Sima, and Stuart Webb. 2016. The Relationship between Academic Vocabulary Coverage and Scores on a Standardized English Proficiency Test. *Journal of English for Academic Purposes* 21: 121–132. https://doi.org/10.1016/j.jeap.2015.05.009.
[Crossref]

Pérez-Paredes, Pascual. 2019. A Systematic Review of the Uses and Spread of Corpora and Data-Driven Learning in CALL Research during 2011–2015. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2019.1667832.

Pérez-Paredes, Pascual, Carlos Ordoñana Guillamón, and Pilar Aguado Jiménez. 2018. Language Teachers' Perceptions on the Use of OER Language Processing Technologies in MALL. *Computer Assisted Language Learning* 31: 522–545. https://doi.org/10.1080/09588221.2017.1418754.
[Crossref]

Punie, Yves, and Christine Redecker, eds. 2017. European Framework for the Digital Competence of Educators: DigCompEdu. https://ec.europa.eu/eusurvey/runner/DigCompEdu-

H-EN?startQuiz=true&surveylanguage=EN. Accessed 11 August 2021.

TaLC. 2020. Teaching and Language Corpora Conference: Abstracts. https://langident.hypotheses.org/files/2020/07/Abstracts140720b.pdf. Accessed 10 August 2021.

Tribble, Christopher. 2002. Corpora and Corpus Analysis: New Windows on Academic Writing. In *Academic Discourse*, ed. John Flowerdew, 131–149. Harlow: Longman.

Wang, Haiping, Yuanyuan Zheng, and Yiyan Cai. 2015. Application of Corpus Analysis Methods to the Teaching of Advanced English Reading and Students' Textual Analysis Skills. In *Corpus Linguistics in Chinese Contexts*, ed. Bin Zou and Michael Hoey, 158–177. London: Palgrave Macmillan.
[Crossref]

Webb, Stuart, and Tahere Sima Paribakht. 2015. What is the Relationship between the Lexical Profile of Test Items and Performance on a Standardized English Proficiency Test? *English for Specific Purposes* 38 (1): 34–43. https://doi.org/10.1016/j.esp.2014.11.001.
[Crossref]

Xu, Manfei, Xiao Chen, Xiaobin Liu, Xiaoyue Lin, and Qiauxin Zhou. 2019. Using Corpus-Aided Data Driven Learning to Improve Chinese EL Learners' Analytical Reading Ability. In *Technology in Education. Pedagogical Innovations*, ed. Simon K.S. Cheung, Jianli Jiao, Lap-Kei Lee, Xuebo Zhang, Kam Cheong Li, and Zehui Zhan, 15–26. Singapore: Springer.
[Crossref]

Footnotes

1 All the tools used were the free online versions (with limited features). They can be found at: KAHOOT: kahoot.it / ACADLY: acadly.com / Playposit: Playpos.it / Quizlet: quizlet.com / FLAX: http://flax.nzdl.org/greenstone3/flax / Versatext: https://versatext.versatile.pub/ / Compleat: lextutor.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_7

7. Corpus Linguistics and Grammar Teaching

Christian Jones[1] ✉
(1) University of Liverpool, Liverpool, UK

✉ **Christian Jones**
Email: Christian.Jones2@liverpool.ac.uk

Keywords Grammar – Lexico-grammar – Frequency – Form – Function – Data-driven learning – Text-based approach

7.1 Introduction

A chapter on grammar needs to provide a working definition of the term, necessary because it means different things to different people within second language teaching. For example, some may define it via the lens of a structural syllabus, where the verb phrase, tense and aspect tend to dominate the 'canon' (Burton 2019), while some may view it simply as a way to make correct sentences, and others may simply see it as a series of rules to be learnt. This chapter begins with a definition. It then moves on to examples showing how we can use various corpora to analyse grammar for teaching purposes, showing how teachers can explore corpus data to better understand form and function.

Finally, this chapter will give examples of how we can use corpora to inform grammar teaching. This may be in the preparation of materials, or by adopting methodologies such as a text-based approach (Timmis 2018) and data-driven learning (Johns 1991; Boulton and Cobb 2017; Timmis and Templeton, this volume). The chapter will also look at more

simple uses of corpora when teaching grammar and how it can be used in simple classroom practices such as explanation. All examples in this chapter will be from English corpora, which includes the Corpus of Contemporary American English (COCA, Davies 2008) amongst a number of others. For these corpora, registration is free and all corpora can be searched in the same ways.

7.2 What Do We Mean by Grammar?

Defining grammar is never a simple task, as Jones and Waller (2015a) note. Dictionary definitions can help but they often leave us slightly short. The Cambridge Dictionary (2021), for example, suggests that grammar is "(the study or use of) the rules about how words change their form and combine with other words to make sentences." This is of course partly true—grammar does help us to make sentences and to put words in the correct order. This aspect of the explanation fits the idea that grammar can be explained as a description of the internal structure of words and phrases (morphology) and the manner in which these words are arranged into sentences (syntax) (Carter and McCarthy 2006). To give an example from Jones and Waller (2015a: 23), the sentence *My brother goes shopping every weekend*, would be accepted as correct, whereas *Goes weekend every my brother shopping* would not be. This is because it complies with the standard syntax of English, where sentences normally consist of a subject (*My brother*) and predicator (*goes shopping*), normally followed by either an object, adjunct (as in this case *every weekend*) or complement. The morphology of this example also shows us how word forms help to make meaning. Here *goes* indicates general present time and the final *-s* shows that the subject is related to a third person—in this case *he*.

At the same time, there are several things missing from the dictionary definition given above. Firstly, using grammar entails more than making sentences. We can better describe speech (particularly dialogue) as consisting of utterances and turns rather than sentences. Speakers make utterances when they speak and in dialogue those utterances make up turns. A complete turn can be perfectly acceptable without being what we recognise as a sentence. An exchange such as A: *Are you going tomorrow?* B: *Definitely*, for example, consists of a one

word reply which is not a complete sentence but is both a complete utterance and a correct and appropriate turn. At the same time, spoken language has a grammar which often differs in clear ways from written forms. McCarthy and McCarten (2018) give the example of non-defining relative clauses, used as subordinate clauses in most written forms but able to stand alone to comment upon the ongoing discourse in conversational exchanges. Examples 1 and 2 from COCA show this contrast.

***Which* used in a non-defining relative clause in writing and speech**

Written sample

1. In this research, productivity is represented by TFP—the residuals from growth regression equation. It measures the growth of income per worker, which is not due to factor accumulation (physical capital, labour) (Lipsey and Carlaw 2004).
(COCA 2019, Academic, Business and Economic Horizons)

Spoken sample

2. Ms-LONG: All right, here we go. In five, four, three, two, one. It's going to be awesome!
ROKER: OK.
VIEIRA: It's even more complicated than I realize, you know.
ROKER: I don't even think we had this much—we didn't do this much for the Olympics.
LAUER: No. We had a very easy part…
VIEIRA: Yeah, we did.
LAUER: **which was nice.**
HAGER: The main part. (COCA 2010, Spoken, Today).

There are a number of other ways in which spoken grammar differs from written grammar aside from this example and these have been detailed in several corpus-informed descriptive grammars (e.g., Carter and McCarthy 2006).

Two other additions can be made to the aforementioned dictionary definition of grammar. The first is that grammar is also used, in Thornbury's (2005) words 'beyond the sentence' and beyond the individual turn or utterance. It is used across spoken and written texts

to make meaning. An academic article, for example, may make use of a large number of complex sentences because there is a desire to include as much detail as possible when explaining its ideas. Finally, the forms which we can include under the umbrella term *grammar* need to include those which can be labelled 'lexico-grammar' (Halliday and Matthiessen 2004) because they are somewhere between grammar and vocabulary. An example of this is *take pride in* + *-ing* (Halliday and Matthiessen 2004: 45). This pattern includes the collocation *take pride in* but will commonly be followed by an *-ing* form. When we search for the five most common of these forms in a corpus, we can also list these. A search of COCA, for example, shows that these are *being, knowing, having, providing* and *making*.

Taking all these elements together, we can define grammar in the following ways for the purposes of this chapter:

1. Grammar is the internal structure of words and phrases (morphology) and the manner in which these words are arranged into sentences (syntax), turns and utterances.

2. Grammar is used to make meaning established through rules of use.

3. Grammar makes meaning via form and function in context.

4. Grammar is used in spoken and written discourse in sometimes distinctly different ways.

5. Grammar is used to make meaning across texts as well as in sentences and utterances.

6. Grammar includes lexico-grammar, where patterns lie somewhere between what we might view as grammar and vocabulary.

These definitions are developed from ideas in Jones and Waller (2015a), Biber et al. (1999), Carter and McCarthy (2006), Halliday and Matthiessen (2004), Hoey (2005), Hymes (1972) and Sinclair (1991), each of whom give a more detailed discussion of the subject than space allows for here.

7.3 What Can a Corpus Tell Us About Grammar?

7.3.1 Frequency

One obvious thing that any corpus can tell us is how frequent a grammar pattern is. A number of useful corpus-informed references are now available to provide this information without having to search for it ourselves. Most dictionaries, for example, now have their own grammar section and searching for grammatical patterns will produce examples and explanation based on corpus data. Tools such as The English Grammar Profile (EGP) (O'Keeffe and Mark 2017) based on a corpus of learner exam data, also allow us to understand how different patterns are used across levels and when we can typically expect certain patterns to become established in a learner's repertoire. The EGP also includes examples. Both dictionary tools and those such as the EGP tend to work best when searching for the established 'canon' (Burton 2019) of grammar as found in many structural syllabi and coursebooks, which will include what Timmis (2018: 79) calls the 'big beasts' such as conditionals, modal verbs, relative clauses and so on. For example, a sample search for *conditional* in The EGP (2021) tells us that at CEFR A2 level, the following conditional patterns are often used:

PRESENT SIMPLE *IF* CLAUSE, REAL CONDITIONS
Can use *if* + present simple with *can* or imperative in the main clause to refer to things that are true now or very likely to happen (example 3).
3. *If I have spare time, I always read a book.*

Data such as these can be helpful when planning lessons or trying to decide which aspects we might focus on in class, as they occur in questions or texts. This particular example suggests that at this level at least, real uses of conditionals are common and are frequently formed with *if* + present simple, present simple *can* or imperative forms. Such information may, at times, contradict what is suggested in published classroom materials, which are not always informed by corpus data (Burton 2019) and it shines a light on the fact that they really should be. A number of corpus-based studies of commonly taught grammar

patterns show clear differences with what is often presented in published materials, examples of which include conditionals (Jones and Waller 2011) and relative clauses (Tao and McCarthy 2001).

We can also search for frequency across a whole corpus or in sub-corpora. At the most basic level, we can simply input an example of a pattern we wish to check into the search bar of a corpus and find out how many times it occurs within the whole corpus or in different sub-corpora. Such a search might be based on an example from teaching materials or in response to a student question. To take an example from the first textbook I ever used, *Streamline Departures* (Hartley and Viney 1978), we could search for *What are you going to do?* taken from a unit focused on teaching *going to*. A basic search in COCA shows an overall frequency of this pattern of 1550 occurrences in the corpus. On its own, such data does not tell us that much but it does at least tell us that this is a pattern in use. A more useful search can be undertaken by using the 'chart' function in the same corpus. This compares frequency across sub-corpora, in this case these are made up of blogs, the internet, TV and movies, spoken language, fiction, magazines, news and academic texts. This search shows us that *What are you going to do?* occurs most often in the TV and movies sub-corpus. This same 'chart' function also shows us the occurrences per million words and such normalised frequency gives us a clearer idea of how a form is distributed across the whole corpus. In this case it occurs 5.5 times per million words. It is also possible to use corpora to track how the frequency of a form has changed over time. This can help to tell us the extent to which a form is currently in common use or not. For example, we can search the TV corpus (Davies 2019-), consisting of 325 million words of informal TV shows from the 1950s to the 2010s, from the UK and Ireland, the USA, Australia and New Zealand. This search is produced using the 'chart' button available with this corpus. Table 7.1 shows us these results.

Table 7.1 Occurrences per million words of *What are you going to do?* in a TV corpus

1950s	1960s	1970s	1980s	1990s	2000s	2010s
16.23	15.84	11.84	7.79	6.28	5.78	4.32

The data in Table 7.1 suggest that the frequency of this example has steadily decreased over time, at least in this particular set of data.

Such searches are still somewhat limited, but they do allow us to understand in which contexts items are used, which in turn allow us to answer questions from learners or questions teachers may themselves have about items in materials. These might include whether people say X or Y or in which contexts they might use a particular example.

Slightly more complex frequency searches allow us to explore examples and are particularly helpful when trying to understand lexico-grammar. We can, for example, take a pattern such as *What are you going to …* and use the search function to explore which verbs most often come after *to.* To do this in a corpus such as COCA, we need to search for the pattern *What are you going to* in the search bar and follow this with a search for verbs in the infinitive form. This can be found by locating the POS (part of speech) bar to the right of the search box. In COCA, this would be realised as *What are you going to _v?i* (this simply means verbs in the infinitive form). Results can be displayed as a list and in this case, the ten most common verbs in COCA are: *do, say, tell, be, wear, sing, call, get, have* and *name*. If we wish to, we can also explore the same pattern in particular contexts of use, which may be of most relevance to a particular group of learners. To give an example, the most common five infinitive forms in the spoken section of COCA are *do, say, be, tell*, and *sing*. Finally, we can simply start with examples of a pattern and explore the frequent forms which are used with it. This can be done using the wildcard search function, in which an * is put either before or after the form you are focusing on. This gives you the most frequent items of any type which come before and after this form. For example, a search for *going to* * in the spoken section of COCA reveals that *going to be* is the most common pattern. Further searches can then be conducted, such as *going to be* * and in this case, it shows that *going to be + a lot* is frequent. When we look at the examples, we can see that *there is/was going to be a lot of + noun* is very common in these data.

7.3.2 Function

Frequency searches are useful, but they do not really tell us how a particular grammatical pattern is used or how it functions in spoken or written texts. In order to understand this, we need to explore language in contexts of use. We can do this in a corpus by searching for examples

and then clicking on concordance lines to view the examples within texts. Examples 4–11 below show lines with *going to be a lot*, taken from the spoken section of COCA, which consists of data from unscripted TV shows.

Concordance lines for going *to be a lot*

4. But what you're preparing for, for a lot of folks inland is **going to be a lot of** rain and, unfortunately, probably a lot of power * … (COCA 2019, Spoken, PBS NewsHour)
5. Be fought out both in paid media and free media. **And it's going to be a lot.** JUDY-WOODRUFF): But you're—so you're saying … (COCA 2019, Spoken, PBS NewsHour)
6. End up running for these seats?: Well, I think there's **going to be a lot** of pressure on Bullock, especially if he doesn't make …. (COCA 2019, Spoken, PBS NewsHour)
7. This is a formal criminal probe we're told. And so there are **going to be a lot** of former Obama officials who are going to be under the … (COCA 2019, Spoken, Fox News)
8. Terms of opposition research that Stone and others got from foreign governments There's **going to be a lot** to look into. And I couldn't agree more with … (COCA 2019, Spoken, NBC News)
9. And if you're not doing that self-work, having the conversations with kids is **going to be a lot** harder because these are definitely parallel tracks of work that need … (COCA 2019, Spoken, NPR: The TED Radio Hour)
10. The larger structural issue, which is that there were students who knew there was **going to be a lot** of debt. That debt was crippling and didn't even … (COCA 2019, Spoken, PBS NewsHour)
11. People. So we weren't even counted. Other than that, it's **going to be a lot** of oral history. And how many people do we have … (COCA 2019, Spoken, NPR: Weekend Edition)

In these examples, inspection of the concordance lines shows that this particular grammar item seems to function most obviously to make a prediction about something which is going to happen soon and is the current focus of the TV shows, which, in these examples, often discuss current news and make predictions about what will happen in

upcoming events. If we click on the context of one example, we can see *going to be a lot* used in this way, as example 12 shows.

Context for *going to be a lot*

12. The challenge with these kind of storms, is they're so far out. They have slowed down. We have got lots of time. People, get your supplies and stuff. But if you have got everything and you're set, there's no reason why you can't at least salvage some of these July—this Labor Day holiday. But if you're not ready, you still got time to get—you get your supplies. But what you're preparing for, for a lot of folks inland is **going to be a lot of rain** and, unfortunately, probably a lot of power outages if the storm does come over the state.
(COCA 2019, Spoken, PBS NewsHour)

Careful searches of concordance lines and texts can show us how particular patterns function in context and this can also reveal differences in function between use in different types of texts. Jones and Waller (2015a), for example, show how *married* tends to figure in the British National Corpus (BNC) newspaper sub-corpus within a non-defining relative clause when reporting news, normally as part of the background detail on people in stories. They contrast this with *married* in the COCA spoken sub-corpus, where it tends to be used with *going to* in order to report plans and news.

It is also perfectly possible to analyse a single text with a focus on the grammar within it, and then check the form and function of that in a larger corpus. We might wish to do this to understand how certain texts use grammar to help make meaning, and this can be particularly useful if learners need to study, or are interested in, certain genres. To take one possible example, if learners need or wish to watch soap operas, we can look at one script and analyse the way spoken grammar is used within it. We can then look at a bigger corpus of similar texts and check how common these features are and how they function. A teacher could start by examining the features of one script used in class and then use a soap opera corpus such as Davies (2011-) and check how common these features are. Example 13 shows a sample section of script from the American Soap *Bold and Beautiful*.

Dialogue from a soap opera corpus
13.**Liam**: Hey.
Bill: Why are you calling me? You're at a resort. Put the phone away. Swim up to a bar.
Liam: Yeah, well, it'd be a lot more fun if my wife were swimming with me. She keeps on disappearing, Dad I-I don't know what's going on.
Bill: I wouldn't worry too much about it. Knowing Steffy, she's off planning a romantic dinner for you.
Liam: I guess. Maybe I should go looking for her.
Bill: No, why don't you just let her be? I'm sure whatever she's doing, it's for your benefit. (Corpus of American Soap Operas, 2012, Bold and Beautiful)

There are several noteworthy features in this sample, including lexico-grammatical patterns (*keeps on disappearing. I don't know what's going on*), the use of nouns and pronouns to refer to speakers within and outside of the text (e.g. *Steffy, my wife, her, she*), and ellipsis in examples such as *I guess* (*I guess she might be doing that*), *I wouldn't worry too much about it* (if I were you), and *knowing Steffy* (as I know her). These examples tally with results in Jones (2017), who investigated the spoken grammatical forms commonly used in a corpus of the UK soap opera *EastEnders* and found that some common grammatical features such as ellipsis were present in the soap data. Scripted dialogues frequently featured omission of words or clauses where the situation made the meaning clear and it seems that ellipsis functions in a conversation such as this to add to the naturalness and to make the dialogue faster moving. We can search for any or all of these patterns in the sample text in the larger corpus and check how common they are. For example, *I guess* has 45,412 occurrences in total, with 456.45 occurrences per million words, and a search for *keeps on + ing* shows that common verbs which follow this pattern are *giving*, *going, telling, saying* and *getting*. Following this, we can then view such examples in context using concordance lines or simply as they arise and explore their form and function.

7.4 Teaching Applications

7.4.1 Materials Development

Using corpus data can help in simple ways when we focus upon materials development or adaptation. To take the *What are you going to do?* example shown previously, if such an example comes up in materials we are using, we can easily use a corpus to check how this sample is realised in terms of form and function. As mentioned, in the spoken section of COCA, the most common five verbs are *do, say, be, tell*, and *sing* are the most frequent by far. We can use such information to ensure that these verbs have coverage in teaching materials, classroom explanations and examples which a teacher may use. We can also examine particular instances in COCA. For example, when we click on samples of *What are you going to be*, one common pattern is for this to be followed by *-ing* in examples such as *What are you going to be doing?* This seems to function in a similar way to *What are you going to do?* when used to ask about plans, at least in these data. Example 14 shows this in context.

***What are you going to be doing?* in context**

14. Spain is out of the recession, no doubt. We are growing, last year 1.4 percent; this year probably around 2 percent to 2.5 percent. And Spain and Germany is going to be—are going to be very agents of growth in Europe and probably this can be solved. No, the situation is far better than a couple of years ago. QUEST# And in terms of your bank, what's the future for your bank? **What are you going to be doing in the future?** RODRIGUEZ# We are an international bank and we are in Spain and Mexico…

(COCA, Davies 2015, Spoken, Quest Means Business)

While many coursebooks focus on *going to* patterns, it is rare, in my experience, to offer practice of examples such as *What are you going to be doing?* as a way to ask about future intentions. This kind of corpus data can help us to supplement or amend materials to reflect such common uses.

7.4.2 Methodology

In terms of how we could use corpus data in classrooms when the focus is on grammar, there are a number of options. Due to space constraints,

I will focus on three here: helping with questions, a text-based approach (Timmis 2018) and data-driven learning.

7.4.2.1 Helping with Questions

It is common for learners at all levels to ask, 'Can I say X?'. In the middle of a class, it can be hard to answer such questions clearly. Intuition might suggest something is possible, but the more important consideration is whether it is likely that a grammatical form or forms would be used and if so, in which contexts. We can also determine the relative frequency of different patterns. For example, should a learner ask, 'Can I say *you must be worrying*?' we could check by simply inputting the example into a corpus. A search in COCA reveals only one example of this, suggesting it is not common. If we wish to we can then search for the most common -*ing* forms that follow *you must be* by searching for 'v?g' in the search. This shows that the five most frequent forms are *joking, kidding, feeling, thinking* and *doing*. These searches give us quick and simple answers and there is no reason why we cannot make them a regular feature of classes, provided there is access to the internet. Should learners wish to check patterns in their own time or in class, sites such as Netspeak (2021) allow a quick and easy method of doing this. This site uses the worldwide web as a corpus and enables for simple frequency searches. For example, a search for *you must be thinking* shows there are eighteen thousand occurrences in this data, each of which can be viewed in its context. It is also possible to search for common patterns by use of the question mark symbol. A search for *you should be?*, for example, shows the most common pattern in this corpus is *you should be able to* and that this is often used to give advice /instructions in this data, as shown in example 15:

15. *Unless you are embarking on a massive employment campaign, you* ***should be able to*** *conduct interviews and make hiring decisions without using employment agencies or executive recruiters.* (Netspeak 2021, Google Books)

Learners could use such searches to inform themselves when working outside of class.

7.4.2.2 Text-based Approach

A more comprehensive methodology for use when focusing on grammar is what Timmis (2018: 81–82) terms a text-based approach. His suggestions for such an approach can be summarised as follows:

1. A course or series of lessons can be designed based on a collection of spoken or written texts. These texts are chosen largely for their potential to interest and engage learners, rather than because they contain certain language areas.

2. Such texts are processed for meaning first, before there is a focus on the form(s) within them.

3. Language work [in Timmis' example, this was grammar] is chosen based on the language in the texts, rather than choosing the language to focus upon first and then finding texts which have examples of such language.

Timmis argues that this approach has several advantages, the most obvious of which is that choosing potentially engaging texts can motivate learners to read or listen to them, making any following work on language more meaningful. He also suggests that such a methodology could include a number of different aspects of grammar, rather than focusing on one pre-selected area and that all areas are contextualised, thus better aiding understanding of form and meaning. See Timmis (2018: 81–82) for a more detailed summary.

While Timmis does not specifically suggest we use corpus tools to analyse texts such as those chosen in this approach, there is clearly no reason why we cannot do so. A teacher might, for example, choose a number of texts and then use a tool such as Lextutor (2021) to analyse the grammar patterns within them. We might start by looking at the keywords (the words which are significantly more frequently in our text compared to a reference corpus) and then look to see which grammar patterns these occupy. To do this, you access Lextutor and click on 'keywords'. You then upload (or paste in) your texts and choose a reference corpus. The general reference corpus is the default, and this

mixes the British National Corpus and the Corpus of Contemporary American English.

Jones and Waller (2015b) give one practical example of this. They show one genre which learners on pre-sessional courses in Preston, UK may be interested in reading—factual texts about the city, giving information about its size, history and so on. For students who have just arrived in the UK, they suggest that this kind of text may be engaging as it allows them to find out something about the place they have recently arrived in. Their analysis shows that the word *city* is one keyword in three sample texts, and a look at the concordance lines shows that it tends to form a part of quite heavily pre-and post-modified noun phrases such as *the city's green spaces, one of six city museums.* These noun phrases often occur in complex sentences such as *Preston is a mid-sized **CITY**, located a short distance from the coastline*, something which is clearly a feature of factual texts, where there is a need to pack in lot of information. We can use such information to help us to choose different areas of grammar to focus on when taking a text-based approach, with information from keywords helping to inform our decisions about what we might highlight in the texts we choose. The example above focuses on one pattern with one keyword but there is no reason why several patterns cannot be focused on in any given text or series of texts.

7.4.2.3 Data-Driven Learning

One methodology which makes corpus evidence central, whether used inside or outside the classroom, is data-driven learning (DDL) (Johns 1991). This methodology is based on the idea of using corpus data in classes particularly in the form of exercises and activities which employ concordance lines from corpora, on paper, on a PC or any kind of mobile device. Johns (1991: 5) suggests that a useful framework for learning from concordances is 'identify-classify-generalise' and that this is a guided, inductive approach to learning, whereby teachers help students to make discoveries about language for themselves. Learners can be given some concordance lines and then undertake exercises and discuss questions which help them identify something about the pattern or patterns, connected to form, meaning or both. These initial findings are then grouped or classified in some way and learners can then discuss if

the patterns can be generalised. Although this approach is not a new one, there is good evidence that this can have a positive effect on learners' acquisition of language. Two recent meta-analyses (Boulton and Cobb 2017; Lee et al. 2019) synthesised the results of 64 studies and 29 studies respectively to explore the use of DDL. The first meta-analysis examined DDL in general and shows a strong effect across a range of learning areas and the second a medium effect for vocabulary learning specifically.

As mentioned, a common way in which DDL can be used is to create exercises based on concordance lines. A teacher might identify a problematic area for learners, perhaps based on repeated errors that s/he has observed. A relevant corpus is then found and examples chosen to illustrate the language area being used in context; questions are then created to help learners to identify the grammatical area, classify it in some ways and then generalize about it. The sample exercise in Textbox 7.1 below shows one example of this, focusing on the determiner *some*. The samples are taken from the TV Corpus, which uses informal TV shows from the 1950s to the 2010s as its data (Davies 2019-).

Textbox 7.1 Sample DDL Exercise Focused on *some*

1. Look at the examples below. Can we follow *some* with (a) countable nouns?' (b) uncountable nouns? Or (c) both?
*a. What job? Why? What difference does it make? Thought we might have **some** friends in common*
*b. Hansel. I'm a professor. I'm, uh, involved in **some** crucial research on the, uh, Ca-pe Canaveral thing.*
*c. I've made **some** iced coffee. Do you want **some**?*
*d. Thanks, it is very comforting.—Will you give me **some** chips?*
*e. We found it! Thank you, Mr Bond! Da… Can I get **some** water?*
*f. Oh, please, I have **some** money, too. It's in my purse in the car.*
*g. And I got you **some** diapers and stuff just to get you started.*

2. Look again at the examples above. Which is correct?
   A. We use *some* for an exact amount
   B. We use *some* for an amount which is not made clear.

3. Find the example where there is no noun after the word *some.* Why is there is no noun? When can we drop the noun?
4. Make a list of uncountable nouns we can use after *some* and a list of countable nouns. Start with the sentences above and then add your own ideas.

Some

COUNTABLE NOUNS	UNCOUNTABLE NOUNS
	Money

5. Correct the errors below, taken from the last homework task
I went shopping and met some friend.
I wanted to have some glass of water but I could not find a shop.
I made some tea and I asked him, 'would you like?' and he said 'yes'.

In summary, the following principles may be helpful to guide teachers who may choose to work with DDL when focusing on grammar:

1. Choose an area to focus on. DDL works well if used to help with areas your learners are struggling with.

2. Design activities based on this grammatical area with a limited and manageable amount of corpus data.

3. Design exercises which guide learners to attend to the form(s) and meanings(s) you wish to focus upon.

4. Use the basic framework of 'identify-classify-generalise' (Johns 1991: 5) when designing exercises.

7.5 Conclusion

In this chapter, I have tried to show how corpora can be used to answer classroom questions, to develop materials and in two sample methodologies. Corpora obviously do not teach for us, but they are helpful aids when we are focusing on grammar. They can help us to improve on the answers we give to common questions such as *Can I say*

X? Intuition often suggests that a grammatical form is possible, but corpus data gives us clearer evidence as to whether it is probable and likely. Evidence from a corpus can also tell us whether a particular form is more common in one context or another. When designing classroom materials, consulting a corpus can help us to ensure these materials contain the most frequent forms and reflect common usage. To repeat an example given previously, if we look at the *going to* form, we can use a corpus to tell us the most common infinitives which follow this in written and/or spoken texts. We can then use this information and ensure we include examples in exercises with this most frequent language. Finally, I have shown how corpora can be used in two sample methodologies, a text-based approach and data-driven learning. Both methodologies use corpus data to inform teaching. In all cases, I have tried to show that corpora can contribute to the teaching of grammar, ensuring that such teaching is corpus-informed Overall, it is clear that corpora are a useful tool for teachers when focussing on grammar and one which I would encourage any teacher to explore.

References

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Boulton, Alex, and Tom Cobb. 2017. Corpus Use in Language Learning: A Meta-analysis. *Language Learning* 67 (2): 348–393.
[Crossref]

Burton, Graham. 2019. *The Canon of Pedagogical Grammar for ELT: A Mixed Methods Study of Its Evolution, Development and Comparison with Evidence on Learner Output*. Unpublished PhD thesis, Marry Immaculate College, Limerick.

Cambridge Dictionary. 2021. https://dictionary.cambridge.org/. Accessed 1 June 2021.

Carter, Ronald, and Michael McCarthy. 2006. *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Davies, Mark. 2008-. *The Corpus of Contemporary American English (COCA): 600 Million Words, 1990-Present*. https://www.english-corpora.org/coca/. Accessed 1 February 2021.

———. 2011-. *Corpus of American Soap Operas: 100 Million Words*. https://www.english-corpora.org/soap/. Accessed 1 February 2021.

———. 2015-. *Corpus of Contemporary American English (Spoken Section) 127 Million Words*. https://www.english-corpora.org/coca/. Accessed 1 February 2021.

———. 2019. *The TV Corpus: 325 Million Words, 1950–2018*. https://www.english-corpora.org/tv/. Accessed 1 February 2021.

English Grammar Profile Online. 2021. http://www.englishprofile.org/english-grammar-profile/egp-online. Accessed 22 March 2021.

Halliday, Michael, and Christian Matthiessen. 2004. *An Introduction to Functional Grammar*. 3rd ed. Abingdon: Routledge.

Hartley, Bernard, and Peter Viney. 1978. *Streamline English Departures*. Oxford: Oxford University Press.

Hoey, Michael. 2005. *Lexical Priming: A New Theory of Words and Language*. Abingdon: Routledge.

Hymes, Dell. 1972. On Communicative Competence. In *Sociolinguistics: Selected Readings*, ed. John B. Pride and Janet Holmes, 269–293. Harmondsworth: Penguin.

Johns, Tim. 1991. Should You Be Persuaded: Two Examples of Data-Driven Learning. In *Classroom Concordancing*, ed. Tim Johns and Philip King. *English Language Research Journal* 4: 1–16.

Jones, Christian. 2017. Soap Opera as Models of Authentic Conversations: Implications for Materials Design. In *Authenticity in Materials Development for Language Learning*, ed. Alan Maley and Brian Tomlinson, 158–175. Newcastle: Cambridge Scholars Publishing.

Jones, Christian, and Daniel Waller. 2011. If Only It Were True: The Problem with the Four Conditionals. *ELT Journal* 65 (1): 24–32.
[Crossref]

———. 2015a. *Corpus Linguistics for Grammar: A Guide for Research*. Abingdon: Routledge.
[Crossref]

———. 2015b. Using a Pedagogic Corpus to Develop Language Awareness. *Humanising Language Teaching* 17 (4). http://old.hltmag.co.uk/aug15/idea.htm.

Lee, Hansol, Mark Warschauer, and Jang Ho Lee. 2019. The Effects of Corpus Use on Second Language Vocabulary Learning: A Multilevel Meta-analysis. *Applied Linguistics* 40 (5): 721–753.
[Crossref]

Lextutor. 2021. https://www.lextutor.ca/. Accessed 1 July 2021.

McCarthy, Michael, and Jeanne McCarten. 2018. Now you're talking! Practising Conversation in Second Language Learning. In *Practice in Second Language Learning*, ed. Christian Jones. Cambridge: Cambridge University Press.

Netspeak. 2021. https://netspeak.org/. Accessed 1 July 2021.

O'Keeffe, Anne, and Geraldine Mark. 2017. The English Grammar Profile of Learner Competence: Methodology and Key Findings. *International Journal of Corpus Linguistics* 22 (4): 457–489.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tao, Hongyin, and Michael McCarthy. 2001. Understanding Non-restrictive Which-Clauses in Spoken English, Which is Not an Easy Thing. *Language Sciences* 23 (6): 651–677. [Crossref]

Thornbury, Scott. 2005. *Beyond the Sentence: Introducing Discourse Analysis*. Oxford: Macmillan.

Timmis, Ivor. 2018. A Text-based Approach to Grammar Practice. In *Practice in Second Language Learning*, ed. Christian Jones. Cambridge: Cambridge University Press.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_8

8. Corpus Linguistics and Vocabulary Teaching

Leo Selivan[1, 2] ✉
(1) David Yellin College of Education, Jerusalem, Israel
(2) Oranim Academic College of Education, Kiryat Tiv'on, Israel

Keywords Vocabulary – Lexical approach – Wordlists – Frequency – Collocation – Collocation fork – Chunk – Multi-word unit

8.1 Introduction

If there is one insight we have gained from corpus linguistics, it is that vocabulary is inseparable from other areas of language—specifically, from grammar. However, since this chapter is devoted to what we would traditionally refer to as 'vocabulary', it will look at how corpus research has contributed to our understanding of the nature of vocabulary, and how corpus tools that are freely available today can be used for teacher preparation and classroom instruction.

The contribution of corpus linguistics to L2 vocabulary teaching has been manifold but could be broadly grouped into three main strands: lexicographical, theoretical and pedagogical. On the one hand, corpus linguistics has informed a new generation of corpus-based learner dictionaries. The COBUILD project alone resulted in 16 dictionaries and reference books, most notably the *Collins Cobuild Dictionary* (1987), and spurred a lexicographical revolution. On the other hand, insights from corpus research have given rise to new fascinating theories of language, such as Hunston and Francis's 'Pattern Grammar' (2000) and Hoey's 'Lexical Priming' (2005). Finally, and perhaps more pertinently, insights from corpus research have influenced the work of prominent

ELT methodologists, such as Dave Willis and Michael Lewis, who went on to propose influential methodological frameworks.

Using statistical evidence from the COBUILD project, Willis (1990) proposed a lexical syllabus based on learning the most frequent words in English and studying patterns associated with them. Lewis's proposal, outlined in his *The Lexical Approach* (1997) and subsequent titles (1997, 2000), set out to shift the emphasis to more holistic multi-word units or 'chunks' of language as an organising principle. In the same vein, this chapter will first focus on single words—traditional units of meaning—and then on multi-word units. I will start by looking at how two corpus research techniques—the generation of frequency lists and concordancing—can be used in the classroom, respectively, for deciding what vocabulary should be prioritised in teaching and for *data-driven learning* (Johns 1991). The second part of the chapter is devoted to the complex area of multi-word units, our understanding of which has been greatly enriched by corpus research, and particularly by John Sinclair's (1991) pioneering work. I will then return to corpus-derived frequency lists, this time of multi-word units, and examine corpus-based tools for identifying and extracting them. All the tools and resources mentioned in the chapter are open access, and all the practical tasks and suggestions have been trialled on students.

8.2 Words

8.2.1 Word Frequency

Rough estimates suggest that English has somewhere between 500,000 and 600,000 words (Crystal 1995). How do we choose which words to teach? How do we choose which words are more important when you start learning English? Where do we begin?

A common way of grading English vocabulary is to see how frequently words occur in discourse. One of the first attempts to put together a list of the most frequent words predates corpus linguistics as we know it today. Manually compiled by Michael West, this list was first published in 1936 and revised in 1953 and is known as the 'General Service List'(GSL). The idea behind the endeavour was to select words that would be of the greatest general service to learners of English, hence the name.

The advent of computers vindicated West's vision. Computer technology afforded a more rigorous and efficient analysis of data than was possible in West's days. The analysis of huge text corpora showed that around 80% of the running words in a text consist of 2000 most common words (Nation and Waring 1997). Seen another way, 2000 words cover 80% of almost any English text. This is known as 'text coverage'.

In the 2000s, there followed a plethora of other lists (see Folse and Youngblood 2017 for overview). These include the Academic Word List (AWL) (Coxhead 2000) and lists based on specific corpora, such as the BNC2000 (Nation 2006) based on the British National Corpus, the Word Frequency List of American English (Davies and Gardner 2010) based on the Corpus of Contemporary American English, and the BNC/COCA2000 (Nation 2012), which combines the two. Another recent addition are word lists for the CEFR levels compiled as part of the English Vocabulary Profile (EVP) project (Capel 2010).

Meanwhile GSL underwent a much-needed modernization. In 2013, on the 60th anniversary of West's GSL, a team of corpus linguists headed by Charles Browne released the New General Service List (NGSL)[1] and the New Academic Word List meant to supplement the former. A much larger corpus was used this time: a 273-million-word subsection of the 1.6 billion word Cambridge English Corpus (CEC), compared to West's corpus of just 2 million words (a very small corpus by modern standards). In addition to using a larger corpus, the list provides a greater coverage (92%) and is more up to date: gone are such antiquated words as *footman* and *milkmaid*, which have fallen out of frequent circulation since the days of the original West's list.

Brown's website also hosts a slew of publicly available specialized and disciplinary word lists, such as the Business Service List (BSL), the Fitness English List (FEL), the TOEIC Service List (TSL), and the New Dolch List (NDL) for young learners.

After this brief primer on what wordlists are available, I will now turn to their pedagogical uses.

8.2.2 Uses of Wordlists

The creation of frequency lists and the resulting notion that 2000 words provide 80% coverage has a number of pedagogical applications.

With the help of such lists, learners' attention can be drawn to, arguably, the most useful words in the English language.

Establishing frequency-based core vocabulary helps establish a lexical inventory for teaching purposes and benchmarks for assessment purposes and can therefore provide an empirical basis for defining syllabuses, developing materials and designing tests. Wordlists have also been instrumental in producing graded readers, such as Penguin Readers or Macmillan Readers, which allow learners to start reading in English with—or despite—their limited vocabulary.

Frequency lists find support among key scholars in the field of L2 vocabulary acquisition. The long-time proponent of such lists, Paul Nation, highlights the importance of drawing a distinction between high- and low-frequency vocabulary, pointing out that the former group of words is relatively small but "important no matter what use is made of the language" (2011: 3), whereas the latter group consists of tens of thousands of words that "are often restricted to certain subject areas" (ibid), which do not deserve the same amount of attention in the classroom. Schmitt and Schmitt (2014) refine Nation's categorisation of vocabulary proposing high-, mid- and low-frequency categories, but otherwise echo Nation's suggestion that high frequency words require deliberate learning.

It has to be noted, however, that there is no universal agreement among applied linguists on whether frequency should be the main vocabulary selection principle. For example, Milton (2009) asserts that an effective language course would require a combination of high frequency and low frequency vocabulary because teaching materials tend to be arranged thematically. It is inevitable that, in addition to high frequency words, learners will have to learn words that are thematically important although not necessarily frequent. A coursebook unit on travel, for instance, may include such mid-frequency items as *boarding* (K4)[2], *departure* (K4), *passport* (K5), and *luggage* (K6). Assuming frequency plays an important—albeit not the sole—role in vocabulary selection, let us look at some teaching applications of corpus frequency data.

8.2.3 Word Frequency in the Classroom

Although frequency is an important factor in vocabulary selection, frequency lists themselves do not have to be a starting point for vocabulary teaching. Indeed, many teachers might have reservations about teaching off a list of decontextualised words. Rather, lists can be consulted when deciding whether a text is appropriate for a given level. This can be done with the help of vocabulary profilers, which determine how many words from a given word list a text contains.

VocabProfiler on Lextutor (www.lextutor.ca/vp), a corpus research platform developed by Tom Cobb (2022), provides access to most of the above-mentioned lists and other corpora. After a text is pasted into VocabProfiler and analysed, it is returned with words colour-coded according to their frequency bands. In addition, lists of words are generated for each band. This can serve two pedagogical purposes. Firstly, it allows you to verify the suitability of a chosen text by determining its lexical coverage. Research shows that for adequate text comprehension the coverage of a text should be between 95% and 98%, i.e. the proportion of unknown words should be no higher than 5%, but optimally no more than 2% (Hu Hsueh-chao and Nation 2000; Laufer and Ravenhorst-Kalovski 2010). Secondly, profiling a text with the help of VocabProfiler can help when deciding which items may have to be pre-taught or glossed (before reading).

A similar tool, which can be found on the English for Academic Purposes (EAP) Foundation website, is known as AWL Highlighter (www.eapfoundation.com/vocab/academic/highlighter). As the name suggests, it highlights words from Coxhead's Academic Word List (AWL), which might be particularly useful for EAP instructors. There is also a gapfill feature, which can be used to create fill-in-the-blank exercises with AWL items blanked out. Such exercises allow for more word-focused activities in the context of the studied text after the initial focus on comprehension.

Of course, in order to fully exploit the potential of text profiling tools, it helps to know the learners' vocabulary size—how many words they know. In practice, many teachers might not know or do not have time to measure their learners' vocabulary size, but they do know their proficiency level. With the help of a profiling tool provided by VocabKitchen (www.vocabkitchen.com), a lexical profile of a text can be generated based on the CEFR levels.

8.2.4 Problems with Looking at Words in Isolation

A few words of caution are necessary when working with wordlists or determining word frequencies in texts. The most frequent words in English are words such as *any*, *do, in, of, the*, *to.* These structure and function words carry little informational content and contribute little to meaning. These are, nevertheless, important to master for developing grammatical competence. The *Collins COBUILD English Course* (Willis and Willis 1988), which was an outcome of the COBUILD project, had precisely that goal. Using highly frequent words as a starting point—and dispensing with the traditional grammar syllabus in the process—the authors exploited "the enormous power of the common words of English" (Willis 1990: 46) to introduce dozens of lexico-grammatical patterns based on those words.

A further problem worth highlighting arises when looking at words in isolation: many words in English are polysemous, i.e., they have multiple meanings. Lexical profiling of a text does not disambiguate between different meanings of a word; even if words in corpora are tagged for parts of speech, they are not tagged for meaning. Ironically, it is a small group of highly frequent words that is highly polysemous. In fact, the relationship is proportional: the higher the word's frequency, the more meanings it is likely to carry (Zipf 1945; Nagy and Scott 2000).

From these two points it follows that it is essential to look at words in their context of use. To that end, I now turn to another corpus handling technique—concordancing.

8.2.5 Concordancing

One of the advantages of corpus linguistics is having fast and easy access to numerous samples of language use in context. This allows for rich language exposure which learners, especially EFL learners, may lack. Such exposure plays a key role in developing awareness of language patterns.

One of the basic tools of corpus analysis, in addition to the generation of frequency lists, is concordancing, which retrieves all instances of a search term (a word or a string of words) from a corpus. The result consists of numerous random samples (concordance lines), which show the search term in its textual environment. The most

common format for displaying the result on the screen is Key Word in Context (KWIC), in which the target item (keyword) appears in the middle. Today, there are many free online concordancing programs available to assist teachers and learners wishing to look at naturally occurring language.

The process of examining concordance lines in language teaching is known as *data-driven learning* (DDL) (Johns 1991) and can be approached in two ways (see also Timmis and Templeton, this volume). Learners can themselves access corpora and retrieve concordances on computers or hand-held devices. Alternatively, the teacher can access a corpus and prepare concordance lines for learners to study in a handout or PowerPoint presentation. These two versions of DDL are known as *hard* and *soft* respectively (Leech 1997; cited in Gabrielatos 2005). The hard version—and, to a lesser extent, the soft version—are "akin to the activities of corpus linguists" (Cobb and Boulton 2015: 481) in that learners discover facts about language through attested instances of use in much the same way as corpus linguists observe and interpret patterns of use. The position of the keyword in the centre and the sorting of words to the left and to the right of the keyword make patterns easily discernible (Bernardini 2004).

This is in not to suggest that such examination of concordance lines should replace extensive reading, which is vital for the development of reading skills and repeated encounters with new vocabulary (Gabrielatos 2005; see also Curado Fuentes, this volume). Rather, examining the keyword in context, which constitutes a form of vertical reading (as opposed to horizontal reading of texts), provides learners with 'condensed exposure' (ibid. 10), promoting noticing of recurring language patterns. It can therefore be seen as a form of discovery learning, which is empowering and motivating for learners (Bernardini 2004).

Let us now look at how both the soft and hard version can work in practice to support teachers with two different classroom scenarios.

8.2.6 Concordancing in the Classroom: Practical Examples

Questions such as "Can I say X?" or "What is correct: X or Y"? often arise in class. Concordancers can be extremely useful when teachers are confronted with—and stumped by—such linguistic dilemmas. For

example, to find out when *arrive* is followed by *in* and when is it followed by *at*, learners can be instructed to perform a search using the concordancing tool on Lextutor (https://lextutor.ca/conc/eng). In the main search screen (see Fig. 8.1 below), the target word is inserted into the search field after **Keyword(s) equals**. Clicking on **Get concordance** brings up the concordance lines.

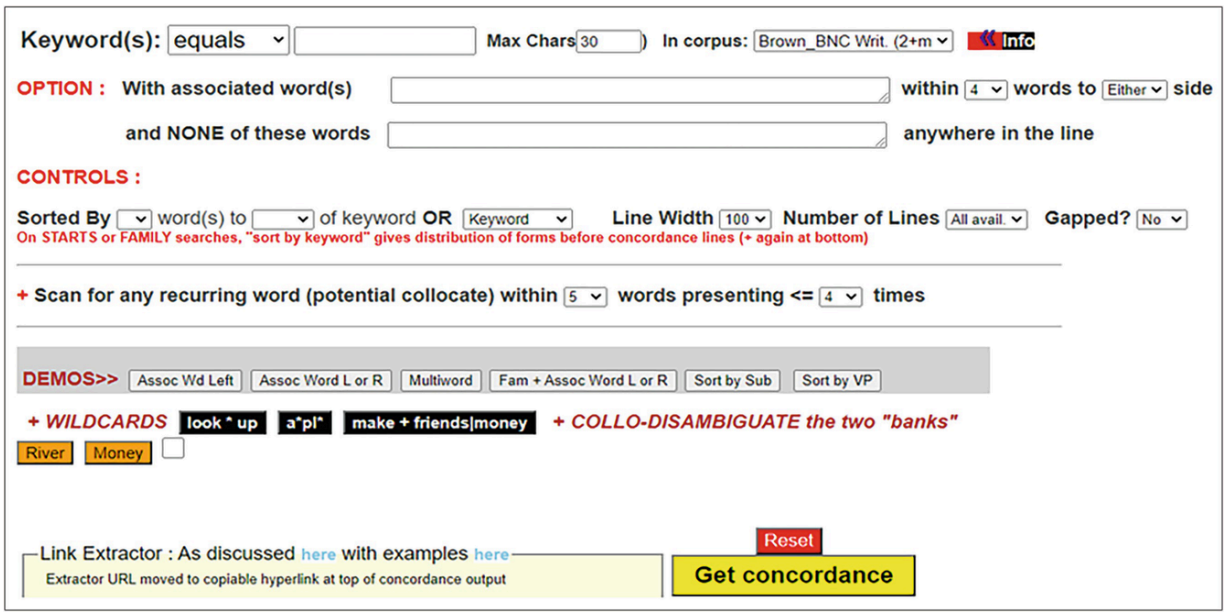

Fig. 8.1 Lextutor concordance screen

In addition to confirming our intuitions—or verifying our assumptions—about language use, concordancers can also be used to point students towards correct usage. In practical terms, instead of correcting the mistake, the teacher can provide the learners with a list of concordances with the target item, so that they can study its patterns of use and discover the error for themselves. This corrective function of concordancers—where learners compare their own writing with authentic language data—is particularly useful for fossilized errors (Nesselhauf 2004). For example, if learners consistently produce *I recommend you this film / this restaurant* when writing (film/restaurant) reviews, the teacher can prepare a worksheet with concordance lines for the word *recommend* to help the learners gain a better understanding of how the target word should be used (for greater elaboration on the uses of corpus linguistics for teaching writing, see Friginal et al., this volume). For this search the user should

adjust the search parameters by sorting the concordance lines by 1 word to the right of the keyword.

Because many learners find the raw linguistic data contained in concordance lines messy or misleading, unclear or unwelcome instances can be adjusted or removed altogether (Braun 2005). Such tidying up of concordance lines, also known as pedagogic mediation (Widdowson 2003), is recommended when using the soft version of DDL.

After exploring concordance lines for various target words—possibly over a course of a few lessons—the same concordance lines can be used as a basis of gapfill activities. The teacher can remove the keyword and ask students to reinsert it by using the surrounding text to the right and left of the target word (co-text) as clues. Highly frequent words particularly lend themselves to concordance-based gapfill exercises. Marsha Chan's (2021) website (http://streaming.missioncollege.org/mchan/media/voc/index.html) has a number of such exercises, where seven concordance lines are presented for each high frequency word.

Language practitioners (e.g., Peachey 2021) have also reported turning to concordancers for the following queries:

- Finding examples when teaching a new word or chunk, by using concordancing to find and copy out examples. This is particularly useful if students struggle to understand how a word is used in English, which is often the case when a word does not have a direct translation equivalent in their L1.
- Exploring different uses/meanings. Students look at concordance lines of a polysemous word and group the examples according to different senses of the word.
- Recording collocations. When learning a new word, students can enter it into a condordancer, look through concordance lines and find and record common collocations with the word.

In the following section I will look at multi-word units, and specifically collocations, in more detail.

8.3 Multi-word Units

8.3.1 From the Idiom Principle to the Lexical Approach

Knowing a word involves knowing many other words that co-occur with it. Although collocation—the co-occurrence of words—was first described by Harold Palmer in the 1930s and expounded upon by John Firth in the 1950s, we owe our understanding of the phenomenon to the work of John Sinclair, who led the ground-breaking COBUILD project. Based on corpus evidence, Sinclair (1991) argued that when we produce language, we draw on a vast stock of semi- and fully-prefabricated phrases.

This insight is encapsulated in what Sinclair (1991) termed *the idiom principle*, the idea that text is mostly constructed from ready-made phrases, and contrasted with the *open choice principle*, the traditional view according to which the only constraints we have on text production are grammatical (i.e., choosing the correct tense, word form, syntax). At around the same time, calls started to be made for a language pedagogy that recognises the central role played by recurrent word combinations: by Nattinger and DeCarrico (1992) and, most notably, Lewis (1993, 1997)

Published around the same time as corpus research started to reveal fascinating insights into lexical patterning, Lewis's *The Lexical Approach* (1993) attempted to bring lexis to the centre of L2 pedagogy. The core principle of the approach is the oft-cited claim that "language consists of lexicalised grammar not grammaticalised lexis" (Lewis 1993: vi). Lewis was, of course, not the first one to assert the central role of vocabulary. Willis, whose proposal for a lexical syllabus was discussed in the first section, moved vocabulary to the centre stage in the *Collins COBUILD English Course* (Willis and Willis 1988). A fundamental difference between Willis's (1990) lexical syllabus and Lewis's lexical approach is that Lewis claimed that it is specifically multi-word units (chunks), rather than single words, that should be accorded the most important role within a syllabus.

Another difference is that Lewis advocated a different approach rather than a syllabus—and the lack of a syllabus specification was something his critics were quick to point out (Lindstromberg 2003; Thornbury 1988). If we take an approach to mean the teachers' underlying assumptions and beliefs about language and language learning (Richards and Rodgers 2014), a new approach requires a shift

in the teacher's mindset. However, the main obstacle to changing the teacher's mindset—and, possibly, one of the main reasons for slow uptake of Lewis's ideas—might be the elusive and even enigmatic nature of chunks (Wood 2020). Not only does this element of language —which I have chosen to call here by the teacher-friendly term *chunk* (as do Harwood 2002; Krishnamurthy 2003; O'Keeffe et al. 2007, among others)—elude clear-cut definition, but also many different labels have been used for it: *formulaic sequences*, *prefabs*, *ready-made phrases*, *routinized expressions*, *holophrases*, to name but a few.

Further compounding the issue is the sheer number of multi-word units in English and a variety of different kinds of them—phrasal verbs, collocations, pragmatic formulas, idioms etc.—each one posing its own set of challenges for learners and practitioners who might not know where to start (Wolter 2020). While corpus linguistics cannot resolve methodological issues it can clearly help with selecting chunks and verifying their holistic validity.

As in the first part of this chapter, I shall briefly survey the existing published corpus-based lists before turning to how teachers can themselves avail of corpus tools for teaching multi-word units. I will not focus too much on multi-part verbs (phrasal verbs) and idioms, both of which were given attention (sometimes even an inordinate amount of attention) before the advent of corpus linguistics. Instead, I will focus on the complex area of multi-word units, and specifically collocations, which are not necessarily idiomatic but perfectly semantically compositional. The complexity stems from the fact that their frequency in language production is not immediately obvious to the 'naked eye', but their pervasiveness is confirmed by computational analysis.

8.3.2 Lists of Frequent Multi-word Items

One of the first attempts to generate a list of frequent collocations was undertaken by Shin and Nation (2008). Using the ten-million word spoken sub-corpus of the British National Corpus (BNC), they extracted the 100 most frequent combinations in English. Quite predictably, such combinations as *a bit*, *as well* and *a lot of* appear in the Top 10 with *you know* being the most frequent. Many of these, however, would not qualify as meaningful units for the purposes of language production

and would require additional elaboration, contextualization and co-textualization (e.g., *a bit of time* / *money* / *luck*) by the teacher.

Later attempts at the compilation of multi-word lists did not rely purely on a frequency-based approach. The Academic Formulas List (AFL), a list of the most frequent multi-word sequences in academic discourse, compiled by Simpson-Vlach and Ellis (2010) included a qualitative phase. In this phase, a group of English teachers were asked to rate the value of each sequence.

The resulting list contains 607 items and includes such chunks as *in terms of, in other words* and *with respect to*. Despite the addition of a qualitative refinement stage, the list still contains syntactic fragments such as *this is a* and *there is no*, which may seem to be of little pedagogical relevance.

To be pedagogically relevant, a list "should include only formulaic sequences that realize meanings or functions", observe Martinez and Schmitt (2012: 303), whose Phrasal Expression List is yet another attempt to compile a pedagogically useful list of multi-word units. Aptly abbreviated as the PHRASE list, it consists of 505 items and particularly focuses on multi-word units such as *at all*, *in fact*, *take place* etc., whose meaning is not apparent from the individual words they consist of, i.e., they are non-compositional chunks. The fact that compositional chunks (e.g., *go on a trip*, *conscious decision*) were excluded should not be taken as a sign that they do not merit our attention in the classroom—the role of chunks in the development of fluency and pragmatic competence (Wood 2010), among other things, cannot be underestimated. The authors of the list stress that they focused on the items that may pose difficulty when reading, following Martinez and Murphy's (2009) finding that the presence of non-compositional chunks in a text can seriously hamper reading comprehension.

In recent years, several other lists of multi-word expressions have been developed, mostly based on academic corpora. The Academic Collocation List (Ackermann and Chen 2013) features circa 2500 frequent (lexical) collocations from across academic disciplines, while the Lexicogrammar of the Academic Vocabulary List (Green 2019) takes the AWL as a starting point and provides the most common grammatical pattern for each entry, for example:

(1) *couple* v-prep-*with*

This coupled with strong winds create further pressure on low lying coastal areas.

These last two lists come closest to the kind of list I, as a practitioner, can make use of when implementing the lexical approach and contain the kind of chunks Lewis would probably have wanted to see included in a syllabus. Since both lists are organised alphabetically, it is extremely easy to find the word and its most common collocates or patterns, for example:

(2) *allocate*: *resources*
carry out: *research / (the) task*

8.3.3 Using Corpus Tools for Teaching Chunks

In *Lexical Grammar* (Selivan 2018), I have outlined a procedure for identifying chunks in texts citing a lack of corpus-based chunk extraction tools. This lack, however, has since been rectified. The Université catholique de Louvain has recently devised an algorithm for extracting chunks called IdiomSearch (https://idiomsearch.lsti.ucl.ac.be/). Still in the Beta version at the moment of writing this, IdiomSearch extracts all manner of chunks of varying degrees of fixedness—not only those we would traditionally refer to as idioms, as the name may suggest. The corpus used for extraction, the WaCky Corpus, consists of 200 million tokens, built by web crawling (Baroni et al. 2009). After copy-pasting the text into the input window, the text is returned with all chunks (termed 'set phrases') highlighted in distinct colours. The colours represent different types of multi-word units which are shaded according to how fixed they are: darker shades represent more fixed phrases. After identifying and focusing on relevant chunks in a text, students can be given, in a subsequent lesson, the same text with parts of the chunks blanked out and asked to recall the missing parts.

As mentioned above, IdiomSearch is still in its infancy and does not highlight all the chunks that merit pedagogical attention. So, the teacher's intuition still has a role to play; but intuitive guesses about the usefulness of a certain chunk can be easily verified using COCA. Just like with single words, COCA's KWIC function can be used to retrieve

concordances of a chunk. This can be helpful, for example, when students struggle to understand how a certain chunk is used in context.

Of course, in the soft version of DDL, a handout can be prepared and students asked to look up, for instance, which words tend to occur to the left or to the right of the target chunk, exactly as with single words. It is helpful to first introduce students to the concept of big data before inviting them to explore some chunks that have recently come up in class. The following is an example of an activity for a B2 level class.

Example 1
Part I: Concordancing steps

1. Navigate to www.english-corpora.org and select a corpus from the list. Recommendation: COCA, a constantly updated corpus of American English from both written or spoken sources, or iWeb, based on carefully selected web-based sources (nearly 100,000 websites)

2. Click on the + sign to display a full list of search options, then select **KWIC** (the last option in the list)

3. Enter the chunk in the search field and, right under the search field, set the search span to three words to the left

4. Click on the **Keyword in context** (KWIC) button

5. After the search has been performed the total number of times the chunk occurs in the chosen corpus is displayed. Click on the chunk to see concordance lines, i.e. examples showing how it is used in context.

6. Limit the number of results to 100. Select 100 from the box next to #KWIC

Part II: Further activity with the following chunks:

In moderation, marked increase, to top it all off, growing evidence, BOIL down to, positive outlook

NB. BOIL should be entered in all caps to include all possible forms of the verb (*boils*, *boiled*, etc.)

- What do the examples tell us about their use? Pay attention to which words occur to the left of the chunks: parts of speech (verbs, nouns, conjunctions etc.) or punctuation (e.g., full stop). Note down some examples to share with your classmates later. Which patterns are you going to adopt and use in your own speaking/writing?

The colour-coded KWIC display allows users to easily identify parts of speech. Before students embark on their KWIC exploration, it is advisable to draw their attention to the colour-coding: for example, verbs are coloured pink and prepositions are coloured yellow. The key for colours can be found by clicking on the question mark in the top-right hand corner (next to Re-sort) of the KWIC display.

The following are possible student answers:

alcohol / done ***in moderation***
bas been / seen ***a marked increase***—often goes with the present perfect
and / . [full stop] ***to top it all off***—the phrase tends to occur at the beginning of a sentence/clause
there is / there's / add to ***the growing evidence***
It all / What it / It really ***boils down to***
have a (more) / keep a / maintain a / develop a more ***positive outlook***

8.3.4 Oral Corpora

Although corpus research is normally associated with a study of written samples (even though they can be derived from both written and spoken sources), it does not have to be limited to the written dimension of language proficiency. The compilation of oral corpora should not be too far down the line and some initial elements of a research agenda in that direction have been laid out by Adolphs and Carter (2013). Already today there are online tools, such as YouGlish (https://youglish.com) and PlayPhrase (https://playphrase.me) which are databases of samples of online video recordings which can, in effect, be considered oral corpora. Since prosodic patterns are an important feature of lexical chunks, such tools can be used for focusing on lexical chunks aurally.

One sample activity described in *Lexical Grammar* (Selivan 2018) consists of playing 3–4 recordings featuring the same chunk and getting learners to identify them in a stream of speech and note them down.

Example 2. A list of chunks for playing on YouGlish—for B1/B2 level

1. Navigate to youglish.com. Enter the chunks listed below into the search box and click on *Say it!* The search engine will find several YouTube videos containing each chunk. Use a new tab in the browser for each chunk.

as soon as possible
as a matter of fact
better safe than sorry
don't you think?
has changed dramatically
come up with this idea
to make matters worse
Have you ever seen it?
put the cart before the horse
serves the/its/a purpose

2. Play a few seconds of each video clip. Make sure that students can hear the audio but not see the image with subtitles. Alternatively, hide the subtitles from view.

3. Students listen to the clips and note down chunks before comparing them with partners and practising saying them.

8.3.5 Collocations

A collocation—a frequent co-occurring combination of two content words (e.g. *blind spot*, *attend a conference, unemployment rate*)—is the most important kind of chunk, which deserves the most attention in vocabulary teaching, according to proponents of the lexical approach (Lewis 2000). As Firth (1957: 11) famously put it, "you shall know a word by the company it keeps". Corpus research, and particularly

Sinclair's work, has borne out this assertion, revealing a highly patterned nature of language. Following Firth, Sinclair established his work on the principle that words are not isolated units of meaning—they are closely linked with their collocational patterns.

Sinclair's approach to meaning and Hoey's ensuing theory of lexical priming (2005), posit that each word is stored in our mental lexicon together with its own collocational field (as well as its preferred colligational patterns, semantic prosody, etc.). This 'fine grained approach' elucidates learners' difficulties, especially when they sound unnatural, despite using the correct grammar and the seemingly correct vocabulary. But it also places a heavy burden on teachers (Granger 2011). Indeed, it would be impossible to teach all the word's collocations especially given that most of them are probabilistic rather than deterministic events (O'Keeffe et al. 2007). Some collocational learning may indeed be left to incidental learning, as many proponents of word-based vocabulary learning advocate (e.g., Schmitt 2008). This does not preclude, however, teaching a handful of most probable collocations, which can be done with the help of corpus tools, such as Sketch Engine for Language Learners (SkELL) (Kilgarriff et al. 2015).

Unlike other publicly available corpus tools on the web, SkELL, available at www.sketchengine.eu/skell, has been specifically designed with language learners in mind. To that end, the concordance lines, which may seem messy and fragmented in a pure, unadulterated corpus, and thus may put learners off, have been tidied up and made look more appealing and less scary for learners (Kilgarriff et al. 2015). SkELL has three features: examples, word sketch and similar words. Just as with COCA, the Example feature allows you to search for whole chunks, for example *to make matters worse* or *in this respect*. The Word Sketch feature is particularly useful for helping learners master collocations. The collocations of a target word are neatly categorised according to parts of speech, and if the target word itself can be more than one part of speech (e.g. *access*), the word sketch for each part of speech is presented on separate pages.

A classroom activity involving SkELL could be based on the idea of 'collocation forks' (Selivan 2018), a format for recording collocations with the node word in the handle and the three most common collocates in the prongs (see Fig. 8.2 below). This is particularly useful

for distinguishing near synonyms, such as *mild* and *soft*. Learners can be asked to look up collocates of each word using SkELL's Word Sketch feature and record the three most useful ones in the forks. Similar meanings can be grouped together, as shown below. For example, *soft* can be used with *skin/leather/hair*, all smooth surfaces pleasant to the touch.

	winter / climate / weather
mild	symptoms
	soap

	drink
soft	skin / leather / hair
	spot

Fig. 8.2 Collocation forks

8.4 Conclusion

I began this chapter by discussing how frequency is an important (but not the sole) factor in vocabulary selection, and how teachers can use frequency information and generate frequency-based lexical profiles of texts. However, frequency information provides raw data and does not reveal all the subtleties of word use, especially since highly frequent words are also highly complex. This is where concordancing comes in: concordancers can help both the teacher and students uncover patterns of use that would otherwise be missing in instruction based on L1 translation or definitions of single words.

Multi-word units, specifically collocations, were the second focus of this chapter. Multi-word units, or chunks, have been neglected in the teaching context until recently with little guidance for teachers on which multi-word items should be taught (and how). Recent compilation of multi-word lists is a welcome, if not overdue, development. However, considering that some multi-word units are more frequent than highly frequent words, pedagogically useful lists should ideally combine both types of items. Unfortunately, at the time of writing this, such corpus-based lists do not exist.

On the positive side, various large corpora are readily available online and can be accessed directly by teachers and students. Indeed, a vast array of corpus tools existing today can turn both parties into miners of big data. Of course, which tools and which DDL activities teachers opt for and in which order will ultimately depend on their

beliefs about vocabulary learning and teaching. They can start with single words chosen on the basis of their frequency and, after the initial form-meaning link has been established, move on to focusing on more contextual aspects of word knowledge with the help of concordancing. Conversely, those who prioritize multi-word items in their teaching, such as the adherents of the lexical approach, may want to start by mining texts for chunks of language and drawing the learners' attention to the company words keep. This can be done by using chunk extractors and collocation search tools. Whichever approach is taken, corpus data and the tools of corpus linguistics can prove invaluable for enhancing vocabulary learning and teaching.

References

Ackermann, Kirsten, and Yu-Hua Chen. 2013. Developing the Academic Collocation List (ACL)—A Corpus-driven and Expert-Judged Approach. *Journal of English for Academic Purposes* 12 (4): 235–247.

Adolphs, Svenja, and Ronald Carter. 2013. *Spoken Corpus Linguistics: From Monomodal to Multimodal*. New York: Routledge.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43 (3): 209–226.

Bernardini, Silvia. 2004. Corpora in the Classroom. An Overview and Some Reflections on Future Developments. In *How to Use Corpora in Language Teaching*, ed. John Sinclair, 235–247. Amsterdam: John Benjamins Publishing.

Braun, Sabine. 2005. From Pedagogically Relevant Corpora to Authentic Language Learning Contents. *ReCALL* 17 (1): 47–64.

Brezina, Vaclav, and Dana Gablasova. 2015. Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics* 36 (1): 1–22.

Capel, Annette. 2010. A1–B2 Vocabulary: Insights and Issues Arising from the English Profile Wordlists Project. *English Profile Journal* 1: 1–11.

Chan, Marsha. 2021. English Vocabulary Quizzes. Mission College. https://streaming.missioncollege.edu/mchan/media/voc/index.html. Accessed 6 March 2022.

Cobb, Tom. 2022. *Compleat Lextutor Tutor*. Lextutor 2021. www.lextutor.ca. Accessed January 2022.

Cobb, Tom, and Alex Boulton. 2015. Classroom Applications of Corpus Analysis. In *Cambridge Handbook of Corpus Linguistics*, ed. Douglas Biber and Randi Reppen, 478–497. Cambridge:

Cambridge University Press.

Coxhead, Averli. 2000. A New Academic Word List. *TESOL Quarterly* 34 (2): 213–238.

Crystal, David. 1995. *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press.

Dang, Thi Ngoc Yen, Stuart Webb, and Averil Coxhead. 2020. Evaluating Lists of High-frequency Words: Teachers' and Learners' Perspectives. *Language Teaching Research*. https://doi.org/10.1177/1362168820911189.

Davies, Mark, and Dee Gardner. 2010. Word Frequency List of American English. www.wordfrequency.com

Firth, John R. 1957. A Synopsis of Linguistic Theory, 1930–55. Oxford: Basil Blackwell.

Folse, Keith S., and Alison M. Youngblood. 2017. Survey of Corpus Based Vocabulary Lists for TESOL Classes. *MexTESOL Journal* 41 (1): 1–15.

Gabrielatos, Costas. 2005. Corpora and Language Teaching: Just a Fling or Wedding Bells? *TESL-EJ* 8 (4). www.tesl-ej.org/wordpress/issues/volume8/ej32/. Accessed 20 December 2021.

Granger, Sylviane. 2011. From Phraseology to Pedagogy: Challenges and Prospects. In *The Phraseological View of Language*, ed. Thomas Herbst, Susen Faulhaber, and Peter Uhrig, 123–146. Berlin/Boston: De Gruyter Mouton.

Green, Clarence. 2019. Enriching the Academic Wordlist and Secondary Vocabulary Lists with Lexicogrammar. Toward a Pattern Grammar of Academic Vocabulary. *System* 87: 102158. https://doi.org/10.1016/j.system.2019.102158.
[Crossref]

Harwood, Nigel. 2002. Taking a Lexical Approach to Teaching: Principles and Problems. *International Journal of Applied Linguistics* 12 (2): 139–155.

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

Hu, Hsueh-chao Marcella, and Paul Nation. 2000. Unknown Word Density and Reading Comprehension. *Reading in Foreign Language* 13 (1): 403–430. https://scholar.google.com.tw/citations?user=e3ATgcwAAAAJ&hl=zh-TW

Hunston, Susan, and Gill Francis. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins Publishing.

Johns, Tim. 1991. Should You Be Persuaded: Two Examples of Data-Driven Learning. In *Classroom Concordancing*, ed. Tim Johns and Philip King. *ELR Journal* 4: 1–16.

Kilgarriff, Adam, Fredrik Marcowith, Simon Smith, and James Thomas. 2015. Corpora and Language Learning with the Sketch Engine and SKELL. *Revue française de linguistique appliquée* 20 (1): 15–30.

Krishnamurthy, Ramesh. 2003. Language as Chunks, Not Words. In *JALT2002 Conference Proceedings: Waves of the Future*, ed. Malcom Swanson and Kent Hill, 288–294.

Laufer, Batia, and Geke C. Ravenhorst-Kalovski. 2010. Lexical Threshold Revisited: Lexical Text-Coverage Learners' Vocabulary Size and Reading Comprehension. *Reading in Foreign Language* 22 (1): 15–30.

Leech, Geoffrey. 1997. Teaching and Language Corpora: A Convergence. In *Teaching and Language Corpora*, ed. Anne Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles, 1–23. New York: Addison Wesley Longman.

Lewis, Michael. 1993. *The Lexical Approach: The State of ELT and a Way Forward*. Hove: Language Teaching Publications.

———. 1997. *Implementing the Lexical Approach*. Hove: Language Teaching Publications.

———, ed. 2000. *Teaching Collocation*. Boston, MA: Thomson-Heinle.

Lindstromberg, Seth. 2003. My Good-bye to the Lexical Approach. *Humanising Language Teaching* 5 (2).

Martinez, Ron, and Norbert Schmitt. 2012. A Phrasal Expressions List. *Applied Linguistics* 33 (3): 299–320.

Martinez, Ron, and Victoria A. Murphy. 2009. Effect of Frequency and Idiomaticity on Second Language Reading Comprehension. *TESOL Quarterly* 45 (2): 267–290.

Milton, James. 2009. *Measuring Second Language Vocabulary Acquisition*. Bristol: Multilingual Matters.

Nagy, William F., and Judith A. Scott. 2000. Vocabulary Processes. In *Handbook of Reading Research: Volume III*, ed. Michael L. Kamil, Peter Mosenthal, Peter D. Pearson, and Rebecca Barr, 268–284. Mahwah, NJ: Lawrence Erlbaum Associates.

Nation, Paul. 2006. How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review* 63 (1): 59–82.

———. 2011. Research into Practice: Vocabulary. *Language Teaching* 44 (4): 529–539.

Nation, Paul, and Robert Waring. 1997. Vocabulary Size, Text Coverage and Word Lists. In *Vocabulary: Description, Acquisition and Pedagogy*, ed. Norbert Schmitt and Michael McCarthy, 6–19. Cambridge: Cambridge University Press.

Nattinger, James R., and Jeanette S. DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Nesselhauf, Nadja. 2004. Learner Corpora and Their Potential for Language Teaching. In *How to Use Corpora in Language Teaching*, ed. John Sinclair and John McHardy, 125–156. Amsterdam: John Benjamins Publishing.

O'Keeffe, Anna, Michael McCarthy, and Ronald Carter. 2007. *From Corpus to Classroom Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Peachey, Nick. 2021. Concordancers in ELT. *Teaching English*. BBC British Council. www.teachingenglish.org.uk/article/concordancers-elt. Accessed 28 November 2021.

Richards, Jack C., and Theodore S. Rodgers. 2014. *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.

Schmitt, Norbert. 2008. Instructed Second Language Vocabulary Learning. *Language Teaching Research* 12 (3): 329–363.

Schmitt, Norbert, and Diane Schmitt. 2014. A Reassessment of Frequency and Vocabulary Size in L2 Vocabulary Teaching. *Language Teaching* 47 (4): 484–503.

Selivan, Leo. 2018. *Lexical Grammar: Activities for Teaching Chunks and Exploring Patterns*. Cambridge: Cambridge University Press.

Shin, Dongkwang, and Paul Nation. 2008. Beyond Single Words: The Most Frequent Collocations in Spoken English. *ELT Journal* 62 (4): 339–348.

Simpson-Vlach, Rita, and Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics* 31 (4): 487–512.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Thornbury, Scott. 1988. Lexical Approach: A Journey without Maps? *Modern English Teacher* 7: 7–13.

Widdowson, Henry G. 2003. Expert beyond Experience: Notes on the Appropriate Use of Theory in Practice. In *Mediating between Theory and Practice in the Context of Different Learning Cultures and Languages*, ed. David Newby, 23–30. Graz: Council of Europe Press.

Willis, Dave. 1990. *The Lexical Syllabus*. London: Collins.

Willis, Jane, Dave Willis, and Paula Walker. 1988. *Collins COBUILD English Course*. London: Collins.

Wolter, Brent. 2020. Key Issues in Teaching Multiword Items. In *Routledge Handbook of Vocabulary Studies*, ed. Stuart Webb, 493–510. London and New York: Routledge.

Wood, David. 2010. *Formulaic language and second language speech fluency: Background, evidence and classroom applications*. London, UK: Continuum.

Wood, David. 2020. Classifying and Identifying Formulaic Language. In *Routledge Handbook of Vocabulary Studies*, ed. Stuart Webb, 30–45. London and New York: Routledge.

Zipf, George Kingsley. 1945. The Meaning-Frequency Relationship of Words. *The Journal of General Psychology* 33 (2): 251–256.

Footnotes

1 There is another New General Service List compiled by Brezina and Gablasova (2015). To distinguish between the two, Browne's list is usually referred to as NGSL and the latter as new-GSL.

2 Frequency bands (K1, K2 etc.) shown in brackets in this section are based on Nation's BNC-COCA list. This list was found to be the most useful to L2 learners compared to new-GSL or NGSL (Dang et al. 2020).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_9

9. Culture in English Language Teaching: Let the Language Do the Talking

Kieran Harrington[1] ✉
(1) Faculty of Cultural Studies, TU Dortmund University, Dortmund, Germany

✉ **Kieran Harrington**
Email: kieran.harrington@tu-dortmund.de

Keywords Culture and language – Corpus linguistics – English language teaching – Data-driven learning

9.1 Introduction

In this chapter I consider the role of corpus linguistics in the context of the raising awareness of culture in English language teaching (ELT). I will begin with the concept of culture itself and its embodiment in language. The role of corpus linguistics will then be considered in relation to the use of simple concordancing to identify and raise awareness of the covert—as opposed to overt culture-specific keywords and phrases—manifestations of culture in language, with the pedagogical objective of expediting eureka moments or 'serendipitous' learning' (Bernardini 2004) in the ELT context.

In order to demystify the use of corpus linguistics for teachers who may not on the one hand be totally corpus literate (Zareva 2017) or who on the other hand be over-burdened with time and syllabus concerns, and even worried that their students might be frightened off

by “seemingly complex statistics and computations” (O’Keeffe and Farr 2003: 393), I advocate a threshold, introductory-level application, which here I demonstrate with a user-friendly software platform, Compleat Lexical Tutor, Version 8.3. (Cobb 2021).

The chapter will end with a discussion of the incorporation of concordancing activities into a data-driven model, either indirectly, with the teacher preparing material such as printed concordance lines, or directly, with learners themselves using concordancing software in the classroom.

9.2 Culture

Kroeber et al. (1952) found 160 definitions of culture in the literature. Since then many more have been added from fields of research in sociology, philosophy, literature, linguistics, and, of course, cultural studies.

Zhu Hua (2014) identifies four major conceptual tendencies in the definition of culture: the compositional approach, the interpretive approach, the action approach and the critical approach. In the compositional approach, *people* are presented as delimited by nationality, gender, race, ethnicity, social class, and religion, and *things* are viewed from binary perspectives as represented, for example, in Edward T. Hall’s (1976) iceberg metaphor, in Hofstede’s (2001) onion metaphor and in his model of cultural dimensions. Interpretive and action approaches include Geertz (1973), who proposes the uncovering of meaning in everyday interaction (with the learner as ethnographer in the education context), Street (1993), who speaks of ‘culture as verb’ in active meaning-making, Fay (1996), who considers culture as an evolving connected activity, and Holliday (1999), who advocates a perspective of constantly negotiated and changed rules of interaction in generational, occupational and educational ‘cultures’.

The critical approach, from the field of cultural studies, while also advocating the idea of process and change, principally focusses on power differences within and between groups and how the media, history, education and politics influence everyday human activity (Barker 2004). An important component of cultural studies, especially for the perspective espoused in this chapter, is the question of

representation and signification in culture, or as a part of culture. Williams (1981) and Stuart Hall (1997) refer to signification as the production of meaning in social practice. Williams (ibid.), in particular, goes as far as defining culture as the *realized signifying system*, which has led to the redefining of culture in cultural studies as "the production, circulation, and consumption of meanings that become embodied and embedded in social practice" (Storey 2017: 15).

9.2.1 Culture and Language Teaching

The Sapir-Whorf Hypothesis holds that language controls or influences thought. Most linguists, after almost a century of debate, accept that the influence of language over culture (linguistic relativity) rather than the control (linguistic determinism) is a more reasonable proposition (Wardhaugh 2003). In this contribution, I shift the focus to the embeddedness of culture in language in line with the approaches to culture by Geertz (1973), Street (1993), Fay (1996), Holliday (1999) and Williams (1981) as mentioned above. Rather than consider that human beings are at the mercy of language (Sapir 1929) in their thought processes, I will take the stance that human beings, and their quest for intersubjectivity in the immediacy of communicative interaction, are at the mercy of the meanings that have become embedded in particular contexts of situation and social practices, whether that be in the small tribal community in the fight for survival, or in the very sophisticated modern corporate practice community. People represent and negotiate such meaning using language—with all its interactional and pragmalinguistic resources (Leech 1983)—as a tool (Everett 2012) and perpetuate and rearticulate the meanings and the language over time.

This perspective undercuts approaches such as 'teaching culture as content' (see Hua 2014: 4–6) in language education. It also obviates the need for intercultural learning (as part of the curriculum of language teaching) that is conceived of in terms of political and social education with respect to human rights and diversity (see Sandu and Lyamouri-Bajja 2018: 7). As Edmondson and House (1998) point out, this type of learning might be better undertaken in other school subjects such as the social sciences and history. They also consider that intercultural learning in a language teaching classroom is superfluous, because

teaching a language equates to teaching culture, a view which fits with the perspective of the embodiment of culture in language that I take in this chapter.

The perspective also undercuts two popular models which are closely related to the Council of Europe's (Sandu and Lyamouri-Bajja 2018) view of culture, and which have been incorporated into English language teaching in central European curricula—teaching intercultural sensitivity (Bennett 1993), and Byram's (1997) *savoir*s model of intercultural communicative competence. Both these models, despite the context of the English language teaching classroom, focus minimally on language use (see for example, Bennett et al. 2003). They tend toward a compositional and binary approach to culture, and promote the awareness of clashes of ideologies and civilizations by means of critical incident scenarios, all which is clearly at odds with the context of present-day globalization and the drive to educate for global citizenship.

The Curriculum (2008) for English language teaching of the Ministry for Education and Development of North Rhine Westphalia[1] in Germany is an example of how the impact of these models leads to the prioritization of non-language teaching agenda, such as shaping the students "own reality of life … [and developing] open-mindedness … and tolerance" (ibid.: 74).[2] Interestingly, Nold and Rossa (2008), when developing a tool for the measurement of intercultural competence within this German curricular context, saw the need to create a second component (in addition to the critical incident component) which they call 'socio-pragmatic language awareness' (see also Nold 2009).

In what follows, I have a similar focus, although I contend that the embodiment of ways of life is omnipresent in all language, thereby eliminating the need for critical incident scenarios in the language teaching classroom. Everett (2012), for example, refers to a simple interaction between husband and wife in which the former says, "Are you ready yet?" and the latter answers, "Pour yourself a drink." Apart from the culturally significant sexism in the exchange, even an individual word such as *drink* can be analysed for its cultural association, not only with regard to the cognate and different interpretations in other languages, but also with regard to how that word is used and what it signifies across Englishes (see Buschfeld and

Weidle, this volume). Carter (1998), in the same vein, considers the deeply embedded cultural understandings of such individual lexical words. Just as with grammar, the awareness of these embedded cultural understandings are part of the unconscious competence of the native speaker (Harlow 1990) and as such constitute a concomitant acquisition goal of the student of a second language.

9.2.2 Sociopragmatic Consciousness Raising

The term *sociopragmatics* refers to how the social context impacts on the linguistic resources that are available to communities (Leech 1983). Sociopragmatic competence combines the concepts of sociolinguistic competence, which is knowing how to communicate with appropriateness (Hymes 1972), with pragmatic competence, the effective use of language (Thomas 1983). *Affective* use of language is also an important consideration as people strive to reach local and contextual intersubjectivity with the least amount of collateral damage possible (Lakoff 1973) and strive for the presentation of an (agreeable) self in line with the social customs (Goffman 1959). Leech (1983: 10) distinguishes between sociopragmatics, the knowledge of the social conditions that impact on language use, and pragmalinguistics, the knowledge of the strategies for conveying particular intentions and the linguistic items used to express these intentions. My focus in this contribution is on raising awareness of how these linguistic items are used in talk-in-interaction, in "specific 'local' conditions on language use" (ibid.), because this in turn, brings us closer to the given context and culture.

In language teaching, students can be given explicit grammar instruction and they can be shown various strategies for vocabulary learning. Difficulties arise with elements that require first-hand experience with use. In this sense, the appropriate application of the pragmalinguistic resources that are available in particular contexts and cultures is especially difficult and it usually takes a period of time living abroad and immersion in the culture and language of the host country (Rose 2007; Fraser 2010) before the specific and the local interactional use—which leads in turn to a greater 'feel' (Carter 1998) for the language and the culture—is internalised.

Many frameworks and models for teaching sociopragmatic competence suggest a combination of different approaches, such as evaluation of the social situation, consciousness awareness and production tasks, which include role play activities though modelled dialogues (Cohen and Ishihara 2012; Padilla Cruz 2013; Taguchi 2011; Olshtain and Cohen 1991).

Most of the articles in a dedicated volume, *Culture in Second Language Teaching and Learning*, (Hinkel 2007), which is frequently found on reading lists for ELT university programmes and for courses such as DELTA and CELTA, also promote methodologies of awareness-raising, but it is the call for authentic L2 data that is the common denominator in this volume. Rose (ibid.) also refers to the difficulty of access to such authentic data in the context of the teaching of English as Foreign Language (EFL) as opposed to the context of the teaching of English as a Second Language (ESL). While authenticity and what it means is disputed and debated in much research on language teaching (see for example, Willis 1990; Widdowson 1998), it is clear that textbooks tend to present idealized and neatly-packaged speech acts (such as *I am sorry that +* to redress or express regret for an offence), and that teachers through introspection or the 'apprenticeship of observation' (Lortie 1975) will devise similar formulae. The use of corpora would seem to be the solution to such difficulties, as they "are always authentic in the sense that they contain naturally occurring language data" (Gilquin and Granger 2010: 359) and students and teachers have easy access to them. Corpus linguistics, then, which usually involves computer-based empirical analyses (both quantitative and qualitative) of this naturally-occurring language data, allows the language to 'do the talking' in the sense that it facilitates the unprejudiced engagement of the user with the discourse. Course-book writers and material developers and teachers have less control over the language. Thus, rather than working with what Carter (1998: 50) calls mediated 'culturally disinfected dialogues', students engage with 'real' English in the here-and-now of immediate interaction (see Little et al. 2017), more immediately relevant to them and their understanding of cross-cultural communication.

9.3 Tracing Culture in ELT Using Corpus Linguistics

For work on the interface between language and culture, the facilitation of engagement with naturally occurring language in action that simple concordancing furnishes, is sufficient, I suggest, at least at entry level use. Teachers, adhering to a more indirect data-driven model, can search for the language and culture interface and present it to their students in text format. Within a more direct data-driven approach, on the other hand, concordancing platforms can be used as part of classroom and homework activities (see Sect. 9.4 below).

With regard to culture, little research has been done on *tracing*, to use Schneider's (2018) word, and working with manifestations of culture in corpora for classroom elucidation in the ELT context. There are some exceptions, such as Carter (1998), who explores 'speaking cultures', and O'Keeffe and Farr (2003), who explore sociocultural grammatical choices. O'Keeffe et al. (2007) also consider cultural nuances, but in overt key cultural expressions such as *Oh my God, Jesus Christ* and *God help us*.

Here, I will focus on language use in the context of the immediacy of communication, because this is where one can see clearly the influence of the local interactional context on "systematic differences in group forms and styles" (Stubbe 1998: 257). And that is what is essential in getting the 'feel', as Carter (1998) calls it, for the culture from which the language originates. Moreover, I have chosen a single word, *sorry* (from Old English *sarig* (OED, s.v. *sorry*), meaning distressed, grieved, or full of sorrow) to elucidate the difference between the overt use of that word as presented in EFL text books as the primary means of apology in the context of offence (see Limberg 2016), and as widely researched (Ogiermann (2009: 61) in the context of the speech act of apology, with its 'covert' use in the quest for local intersubjectivity in the immediacy of communication.

9.3.1 Using Compleat Lexical Tutor (Version 8.3) to Investigate the Use of Sorry

The choice of corpora and concordancing software is important for the success of corpus linguistics used in the context of language teaching. The concordancing software, especially when a teacher is in the process of introducing corpus linguistics, should be user-friendly, while the corpora should have a wide range of spoken and written texts; although at the introductory stages there may not be a need for very large corpora such as the Corpus of Contemporary American English (COCA), a one-billion word corpus (Davies 2021).

I have chosen Compleat Lexical Tutor because the platform is user-friendly, and it is a colourful platform similar to gaming platforms that younger students may already be using on their hand-held devices. Furthermore, the platform gives access to corpora in four different languages. For English the main sources in Compleat Lexical Tutor are the British National Corpus (BNC) and COCA. Here I will access one of its smaller corpus samples—one million words of spoken English from the BNC which were recorded in casual and formal settings.

9.3.2 Concordancing in Compleat Lexical Tutor (Version 8.3): Basic Steps

This following section is principally directed at readers who are unfamiliar with concordance programs. I will begin with a step by step outline of the initial logging-in process, access to the concordancing feature and the inputting of *sorry*. The exploration of *sorry* is intended as an example of an activity that can be replicated in the classroom.

The link www.lextutor.ca takes the user to the main webpage.

Steps

1. Click on **English** (highlighted in blue on the platform). The concordancing screen shown in Fig. 9.1 opens.

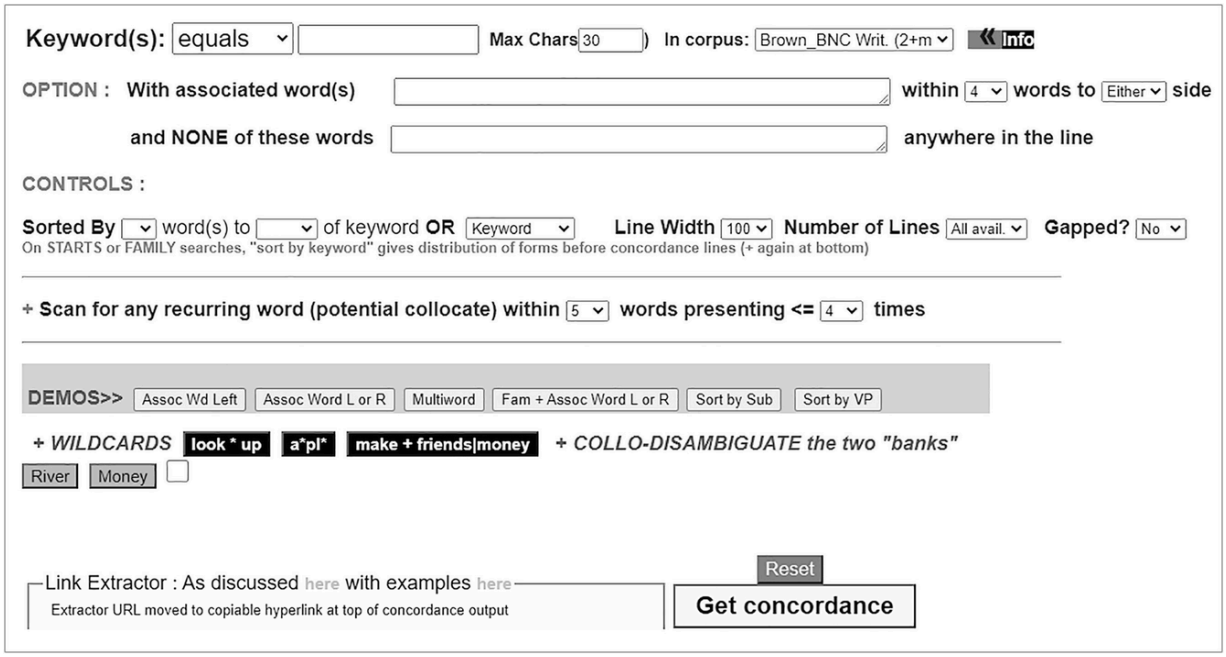

Fig. 9.1 Concordance screen (www.lextutor.ca, last accessed January 2022)

2. Type in *sorry* in the empty box reserved for *Keyword(s)* on the first line.

3. Choose a corpus in the far right box (In **corpus**). This is set by default at **Brown BNC**. Click on the arrow to bring up other corpora. Scroll down and click on **BNC Spoken Sampler (1m)**.

4. Now click on the **GET CONCORDANCE** yellow (on the live platform) box at the bottom of the screen.

The screen that opens shows the 419 hits of *sorry* in concordance lines (which computes at 434 hits per million words), the first ten of which are visible in the screenshot in Fig. 9.2. The node-word (in corpus linguistics terminology the word or string of words you are searching for) *sorry* is in the centre, with multiple words on either side.

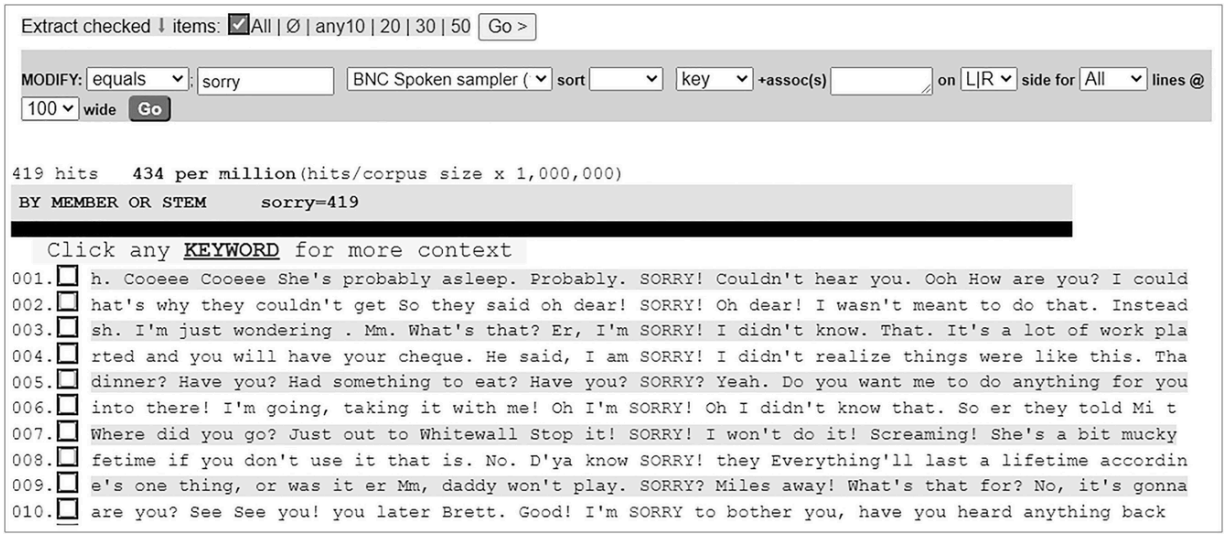

Fig. 9.2 Concordance lines for *sorry*

Whilst on the one hand the perspective of total number of concordance lines here is useful for both teachers and students in its presentation of the vastness and variation of language, sorting the results, on the other hand, facilitates a quick but very interesting view of patterns. Concordancing programmes have wide sorting capacity, but for the sake of simplicity, again, here I show how to search with just one word to the left, ordered alphabetically. In the main concordancing screen (see below CONTROLS in Fig. 9.1), I scroll down to 1v in the 'Sorted By' box and then scroll to Left in the box after 'Word(s) to'. This sorts one word to the left of the node. The operation opens a screen (Fig. 9.3) with clear patterns discernible in the concordance lines, above which is presented the frequency of words that collocate to the left of *sorry*.

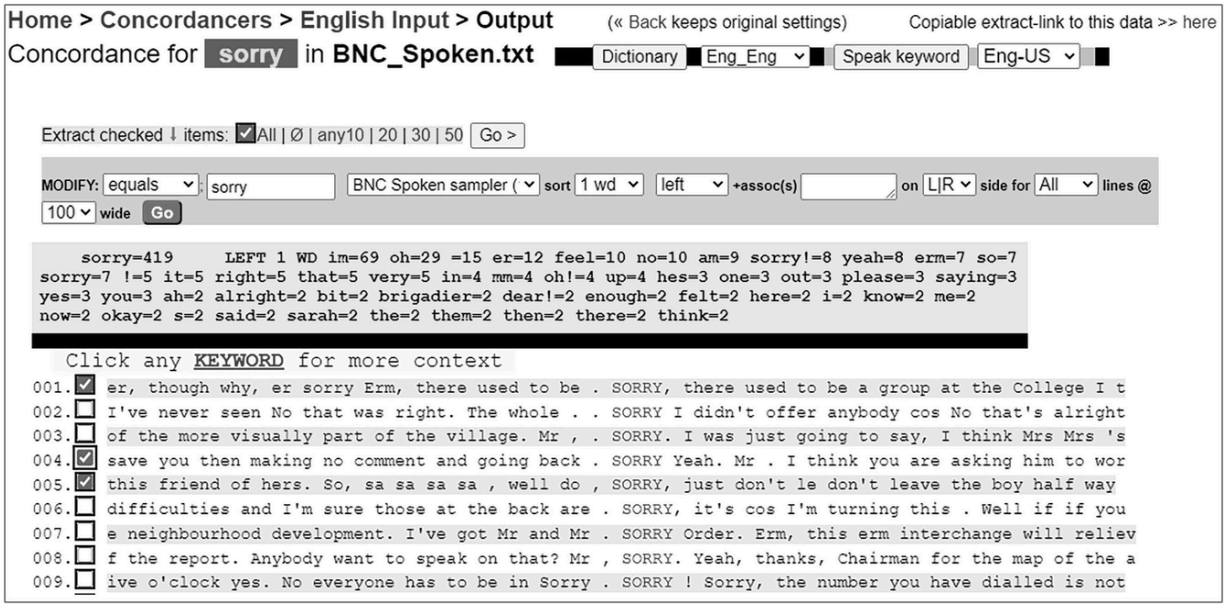

Fig. 9.3 Concordance lines of *sorry* sorted alphabetically one word to the left

The frequency of *I'm sorry* (69 tokens) will not be too surprising as that coincides with its presentation as the standard formula of apology in textbooks (Limberg 2016) and research (Barron 2009; Bergman and Kasper 1993; Aijmer 1996). But the user should find the frequency of collocations (taken as words that occur together frequently) such as *Oh sorry* (29), *Er sorry* (12), *Erm sorry* (7), and *Sorry* after a period (15), worthy of closer examination. This is done by scrolling down the lines and examining these one by one. Normally, after sorting, the concordance lines and the patterns they highlight will give enough information on the use of a particular item; however, the larger (expanded) context of the item can also be called up by most concordancing programmes, and this, apart from providing further clarity, provides the all-important authentic text for classroom use. Figure 9.4 shows the larger context that is displayed after clicking on the node *sorry*, for concordance line 84.

Now Mollie that's, I haven't given you a large helping but it's a ginger pudding. I've been dying to make a good ginger pudding for years!

I like ginger pudding.

So wait and see if it's any good, I don't know. There now Mollie, there's some cream. Erm, SORRY, some custard.

Fig. 9.4 Expanded text for concordance line 84

9.3.3 Patterns and the Connection with Culture

As mentioned previously, some of the patterns (*I am sorry, I am sorry about, I feel sorry about*, etc.), viewable after sorting one word to the left, coincide with the typical description of the apologetic use of *sorry* in the context of offence in English. However, as one might anticipate from the presentation of frequencies (see Fig. 9.3), a more detailed examination (scrolling down through the sorted 419 hits) shows that such use and function is not the most frequent. This can proportion a eureka moment in itself for students who have been told and see in their textbooks that *sorry* is a word used for apologies. What is neglected pedagogically is the use of *sorry* as a discourse marker, that is as part of the procedural use of language which contributes to what Schegloff (1982) calls 'interactional achievement' in the quest for intersubjectivity. This I contend is more important to cross-cultural communication than neatly packaged speech acts and their associated covert keywords and phrases (such *as I am sorry that/for*), and keywords such as *behaviour* which we only use when we *discuss* culture.

There is a significant body of research on discourse markers in the literature (see Harrington 2018 for review), but suffice to say here that, principally, these items help to organize and monitor the discourse (taken here its most basic sense as language above and beyond the sentence), marking openings and closings and speaker incipiency (*right, now, okay, well, yeah*), marking shared knowledge in interpersonal usage (*you know, I see*), marking the speaker's stance (*clearly*, a*ctually*), marking repair and reformulation in a difficult

stretch of talk, (*I mean, oh, hold on*) and in thinking processes (*I mean*), and they also carry out mitigating functions in the preservation of face (*sort of, kind of*).

While there are 116 (27.7%) clear-cut uses of the pattern Subject+*be/feel*+*sorry that* (or *for*) + clause or noun complement in the corpus sample, most occurrences of *sorry* pertain to discourse marking functions in the immediacy of everyday conversation, either on its own before a pause, or collocating with other discourse markers, such as *sorry* (after a pause and at the beginning of an utterance), *er sorry, erm sorry, oh sorry, sorry sorry, yeah sorry, yes sorry, no sorry, mm sorry* and *right sorry*

With the objective of investigating this frequent use as a discourse marker in everyday interaction, I manually searched through all the concordance lines and the expanded contexts of the two-word patterns, and the use of *sorry* before a pause. In these cases, in this one-million-word sampler, the word *sorry* primarily functions as a discourse marker in the management of talk-in-interaction (repair, topic management, interactional order), as a discourse marker of mitigation and face-saving, as a marker before offers invitations and requests, as a marker of new information, and as a pause filler. In Table 9.1, I give some examples of such usage from the corpus.

Table 9.1 Discoursal functions of *sorry* after a pause and in two-word collocations

Line	Concordance	Function
061	*The second of September. December. Sorry, I mean December*	Self-repair
260	*Oh sorry. Wait a minute. Blue.*	Self-repair
001	*Sorry, there used to be a group at the college.*	Speaker incipiency
083	*Er, sorry, not fifty years, twenty five years*	Other repair
263	*Oh sorry. Yeah I mean in terms of*	Clarification
088	*Erm sorry, just to come back, er I work for an amateur theatre company*	Topic renewal
074	*Er sorry, Liz McColgan's just knocked ten seconds off the record.*	Topic change/interruption
079	*Er, sorry, just to follow up on B Sky B again, are you gonna*	Mitigation
003	*Sorry, I was just going to say*	Mitigation

Line	Concordance	Function
078	*Er sorry. Helen? Can I just ask one thing?*	Face-saving/mitigation
085	*It goes something like erm, sorry it's 9.9.*	Hesitation
072	*Ah but everybody er sorry, they did things for me, oh aye bastard*	Filler
049	*Sorry, will you have some toast?*	Offer
084	*Erm sorry, some custard, some sauce*	Offer
266	*Oh sorry. Do you want to put that on the full screen?*	Request initiation/mitigation

It is conceivable that many of these uses could be traced back to the normally associated meaning of *sorry* with apology, but in these examples *sorry* is not functioning as an expression of regret for any salient offence; it is functioning as a tool in the negotiation of meaning. While *sorry* is typically presented in academic articles, but more importantly in textbooks, as the word that is used for apologies, in the context of redressing offence, in 'real' language a substantial part of its usage corresponds to discourse marking functions and the negotiation of meaning in the immediacy of interaction. Engaging with the collection of examples of *sorry* and their analysis exposes students to a wide variety of sociopragmatic applications of the discourse marker to an extent that they could not have experienced from textbooks and in traditional classroom discourse alone. From the engagement with the pragmatic marker *sorry*, students can gain insights into a wide variety of different uses of this and other pragmatic markers investigated in a similar fashion. Alternatively, it may, of course, also be the case that the investigated marker in the target language is only used in a subset of the functions in which a corresponding marker in the students' own language(s). This in turn facilitates a diverse and naturalistic view of the use of language and a clearer view of the impact of the local, cultural interactional context.

9.4 Pedagogical Application and Data-Driven Learning

Data-driven Learning (DDL) (see Timmis and Templeton, this volume) in the context of language pedagogy "consists in using the tools and

techniques of corpus linguistics for pedagogical purposes" (Gilquin and Granger 2010: 359). While espoused as essentially a "new form of grammatical consciousness raising" (Hadley 2002: 99; see also Rutherford 1987: 160), concordancing, as the principal tool of DDL, and as part of a discovery-based approach, can also serve the purposes of languacultural consciousness-raising. It can facilitate the development of sensitivity toward the embeddedness of culture in language and an understanding of how social and communicative processes operate and intertwine. DDL learners have been described as 'travellers' (Bernardini 2001: 22) and 'researchers' (Johns 1997: 101), but as far as language and culture is concerned, they can become linguistic ethnographers (Geertz 1973), learning about communities through their ways with words.

9.4.1 A 'soft' Data-Driven Approach

Teachers can use corpus linguistics as part of a data-driven model, indirectly, that is preparing and printing the concordance lines and authentic texts from their own corpus exploration, and guiding the students in the examination of this material. This is called the "soft" approach to data driven learning (Gabrielatos 2005) and is of particular use in situations where a lead-in to more hands-on corpus linguistics work in the classroom is needed, or in situations where the language level of the students is not high enough to cope with all the concordance lines. Such an activity can be incorporated into different classroom formats, as part, for example, of both presentation and practice in the traditional Presentation Practice Production (PPP) lesson plan, or as part of tasks or problem-solving in a Task-Based Learning approaches (see Mishan 2011). The teacher may have to clean-up discourse that at times looks messy and untidy, and adapt and simplify for different levels, with as much care as possible taken in order not to compromise authenticity (Johns 1994). For example, the teacher can omit concordance lines that are too difficult by simply extracting for use those that are more manageable. Compleat Lextutor offers the facility of selective extraction of concordance lines. The box on the left hand side of the concordance line (see Fig. 9.4) is ticked and the concordance lines that are of special interest are extracted when

the grey button GO> (on the top line of the display) is pressed. Table 9.2 shows the concordance extraction for *sorry*.

Table 9.2 Extraction of *sorry*

Concordance extract for equals SORRY	
001	Sorry, there used to be a group at the college
003	Sorry. I was just going to say
005	Sorry, just don't le don't leave the boy half way

Students can examine these concordances and larger contexts in collaboration, looking at the variation of meaning and function and debating interpretations. But whether done in isolation or in collaboration, the process is inductive, consistent with the pedagogical position that learning is more meaningful and lasting if the learner has critically reflected *en route* to acquisition (Thornbury 2004). The teacher can scaffold the students' examination of corpora (Hadley 2002; Hadley and Charles 2017; Boulton 2010) and advise them, in the context of languagacultural consciousness raising, for example, to consider firstly the taken-for-granted meaning of a word, and then to focus on its function in interaction, the function that leads to final intersubjectivity between speaker and listener. School students may not be familiar with terms such as *repair*, *self-repair*, *other-repair*, *face-saving*, *mitigatio*n, *filler*, etc., but this could also constitute an advantage, because they can tease out in lay-terms, in either their own language or the L2, what *sorry* (in the case of the example I have given) is doing—what its function is. In this way, the students come closer to ways of life that people have articulated and rearticulated in words over time and arrive at that eureka moment: the realization of why native speakers of the L2 say *what* to *whom where*, and *when* and *how* they do it effectively and affectively. This realization can help students of any language to better understand the sociopragmatics of their target language and thus the target culture.

9.4.2 A 'hard' Data-Driven Approach

We speak of a 'hard' version of data-driven learning when learners have direct access to the corpora (Gabrielatos 2005). In the introductory phase the teacher can still provide guidance, but once the students

become more experienced in the use of the software programme and with negotiating concordance lines, the teacher then, especially in the context of language and culture, can allow the students to explore the texts for themselves with a view to sharing their findings and views to the class in a 'plenary' session. That notwithstanding, the software platform I have used here, Compleat Lexical Tutor, and the discovery processes that I have exemplified above, are so straightforward as to create little difficulties for any student, especially given the fact that most people now, from an early age, use digital devices. Once the students engage with the concordance lines, they engage with the connection between language and culture in a neutral interface, that is an interface that is both decentralised from the teacher and decentralized from the student. This also fits perfectly with the promotion and facilitation of learner autonomy in the sense of Holec's (1981) original call for democracy in learning and the development in the student of a sense of awareness and liberation, and Little's (1997) understanding of autonomy as a capacity for detachment and critical reflection on language and culture.

9.5 Conclusion

In this chapter I have considered the role of corpus linguistics in raising awareness of the embeddedness of culture language in the context of ELT. I focussed on examples taken from spoken communicative interaction, which I propose is the kernel of the interface of culture and language, where meanings as associated with ways of life that are constantly articulated and rearticulated using language as a tool in the quest for interpersonal intersubjectivity. As such, all language can be seen to be impacted by culture and we could theoretically investigate and see the legacy of culture in any linguistic item. In order to elucidate the role of corpus linguistics in this context, however, I focussed on sociopragmatics as that is where L2 learners will perhaps need greater experience when studying a second language. The focus is on how words are used in specific local and interactional contexts where patterns and principals are just as important as forms and rules. I focussed on the word *sorry* to show how elusive such words are until

we experience them in *real* interaction, as opposed to how we experience them in language learning textbooks.

I chose a user-friendly concordancing programme and described an activity with a step by step approach, in order to demystify corpus linguistics, and suggest that such an activity is replicable in both teacher preparation of texts and as a part of a direct data-driven activity with corpora and the tools of corpus linguistics in the classroom. The advantages for the student are clear: there is neutral and direct engagement with real English (in this case) and therefore direct and neutral engagement with what I contend is the centre of the cultural communicative universe: the here and now of everyday interaction.

I acknowledge difficulties of accommodating such activities in the busy classroom where the syllabus takes priority, and I also acknowledge the greater role of the teacher with lower levels in achieving a balance between scripted texts of unreal English, which may be easier to comprehend and thus pedagogically useful when the focus is on forms, and the unscripted texts of real English, which may be more difficult to comprehend and less useful pedagogically in certain learning contexts.

As mentioned previously, and as demonstrated above, the use of corpus linguistics in the classroom has the potential for raising awareness of the embeddedness of culture in all language, not just at the sociopragmatic level. And of course, items such as *sorry* can be looked at across Englishes (accessing other corpora) to provide insight into the different development of a particular language in its cultural context. While I would invite teachers to harness this potential, future research can also, of course, gauge and test this use of corpus linguistics as a means of coming closer to culture and coming closer to those eureka moments when students of a language identify *where* the native speakers *are coming from*.

References

Aijmer, Karin. 1996. *Conversational Routines in English: Convention and Creativity*. London: Longman.

Barker, Chris. 2004. *The SAGE Dictionary of Cultural Studies*. London: SAGE.

Barron, Anne. 2009. Apologies Across the USA. In *Language in Life, and a Life in Language: Jacob Mey – A Festschrift*, ed. Bruce Fraser and Ken Turner, 9–17. Howard House: Bingley.

Bennett, Milton J. 1993. Towards Ethnorelativism: A Developmental Model of Intercultural Sensitivity (Revised). In *Education for the Intercultural Experience*, ed. R. Michael Paige. Yarmouth: Intercultural Press.

Bennett, Janet, Milton Bennett, and Wendy Allen. 2003. Developing Intercultural Competence in the Language Classroom. In *Culture at the Core: Perspectives on Culture in Second Language Learning*, ed. Dale A. Lange and R. Michael Paige. Greenwich, CN: Information Age Publishing.

Bergman, Marc, and Gabriele Kasper. 1993. Perception and Performance in Native and Non-Native Apology. In *Interlanguage Pragmatics*, ed. Gabriele Kasper and Soshana Blum-Kulka. Oxford: Oxford University Press.

Bernardini, Silvia. 2001. Spoilt for Choice: A Learner Explores General Language Corpora. In *Learning with Corpora*, ed. Guy Aston, 220–249. Athelstan: Houston.

———. 2004. Corpora in the Classroom. An Overview and Some Reflections on Future Developments. In *How to Use Corpora in Language Teaching*, ed. John Sinclair, 15–36. Amsterdam & Philadelphia: Benjamins.

Boulton, Alex. 2010. Learning Outcomes from Corpus Consultation. In *Exploring New Paths in Language Pedagogy: Lexis and Corpus-Based Language Teaching*, ed. María Moreno Jaén, Fernando Serrano Valverde, and María Calzada Pérez. London: Equinox.

Byram, Michael. 1997. *Teaching and Assessing Intercultural Communicative Competence*. Clevedon, UK: Multilingual Matters.

Carter, Ronald. 1998. Orders of Reality: CANCODE, Communication and Culture. *ELT Journal* 52 (1).

Cobb, Tom. *Compleat Lextutor Tutor*. Accessed January 2022. Lextutor 2021. https://www.lextutor.ca/

Cohen, Andrew D., and Noriko Ishihara. 2012. Pragmatics. In *Applied Linguistics Applied: Connecting Practice to Theory Through Materials Development*, ed. Brian Tomlinson and Hitomi Masuhara. London: Continuum.

Davies, M. 2021. *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/

Edmondson, Willis, and Juliane House. 1998. Interkulturelles Lernen: ein überflüssiger Begriff. *Zeitschrift für Fremdsprachenforschung* 9: 161–188.

Everett, Daniel. 2012. *Language: The Cultural Tool*. London: Profile Books.

Fay, Brian. 1996. *Contemporary Philosophy of Social Science: A Multicultural Approach*. Oxford: Oxford Blackwell.

Fraser, Bruce. 2010. Pragmatic Competence: The Case of Hedging. In *New Approaches to Hedging*, ed. Gunther Kaltenböck, Wiltrud Mihatsch, and Stefan Schneider. Bingley: Emerald.

Gabrielatos, Costas. 2005. Corpora and Language Teaching: Just a Fling Or Wedding Bells? *Teaching English as a Second Language Electronic Journal* 8 (4): 1–35.

Geertz, Clifford. 1973. *The Interpretation of Cultures*. New York: Basic Books.

Gilquin, Gaëtanelle, and Sylviane Granger. 2010. How Can DDL Be Used in Language Teaching. In *The Routledge Handbook of Corpus Linguistics*, ed. Anne O'Keeffe and Michael McCarthy. London: Routledge.

Goffman, Erving. 1959. *The Presentation of Self in Everyday Life*. New York: Doubleday.

Hadley, Gregory. 2002. An Introduction to Data-Driven Learning. *RELC Journal* 33 (2): 99–124.

Hadley, Gregory, and Maggie Charles. 2017. Enhancing Extensive Reading with Data-Driven Learning. *Language Learning & Technology* 21 (3): 131–152.

Hall, Edward T. 1976. *Beyond Culture*. New York: Double Day.

Hall, Stuart, ed. 1997. *Representation: Cultural Representations and Signifying Practices*. London: Sage Publications.

Harlow, Linda. 1990. Do They Mean What They Say? Sociopragmatic Competence and Second Language Learners. *The Modern Language Journal* 74 (3): 328–351.

Harrington, Kieran. 2018. *The Role of Corpus Linguistics in the Ethnography of a Closed Community: Survival Communication*. London: Routledge.

Hinkel, Eli. 2007. *Culture in Second Language Teaching and Learning*. Cambridge University Press.

Hofstede, Geert. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. 2nd ed. Thousand Oaks, CA: SAGE Publications.

Holec, Henri. 1981. *Autonomy and Foreign Language Learning*. Oxford: Pergamon.

Holliday, Adrian. 1999. *Intercultural Communication and Ideology*. London: Sage.

Hua, Zhu. 2014. *Exploring Intercultural Communication*. London: Routledge.

Hymes, Dell. 1972. On Communicative Competence. In *Sociolinguistics*, ed. John Pride and Janet Holmes, 269–293. Harmondsworth: Penguin.

Johns, Tim. 1994. From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning. In *Perspectives on Pedagogical Grammar*, ed. T. Odlin. Cambridge: Cambridge University Press.

———. 1997. Contexts: The Background, Development and Trialling of a Concordance-Based CALL Program. In *Teaching and Language Corpora*, ed. Anne Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles, 100–115. London and New York: Longman.

Kroeber, Alfred, Clyde Kluckjohn, Wayne Untereiner, and Alfred Meyer. 1952. *Culture: A Critical Review of Concepts and Definitions*. New York: Vintage Books.

Lakoff, Robin. 1973. The Logic of Politeness: Or, Minding Your p's and q's. In *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society.

Leech, Geoffrey. 1983. *Principles of Pragmatics*. London: Longman.

Limberg, Holger. 2016. Teaching How to Apologize: EFL Textbooks and Pragmatic Input. *Language Teaching Research* 20 (6): 700–718.

Little, David. 1997. *Learner Autonomy 1: Definitions, Issues and Problems*. Dublin: Authentik Books.

Little, David, Leni Damm, and Lienhard Legenhausen. 2017. *Language Learner Autonomy: Theory, Practice and Research*. Bristol: Multilingual Matters.

Lortie, Dan. 1975. *Schoolteacher: A Sociological Study*. Chicago: University of Chicago Press.

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. 2008. *Lehrplan Englisch. Richtlinien und Lehrpläne für die Grundschule in Nordrhein-Westfalen*, 69–84. Ritterbach Verlag GmbH.

Mishan, Freda. 2011. Whose Learning Is It Anyway? Problem-Based Learning in Language Teacher Development. *Innovation in Language Learning and Teaching* 5 (3): 253–272.

Nold, Günter. 2009. Assessing Components of the Intercultural Competence Reflections on DESI and Consequences. In *Interkulturelle Kompetenz und fremdsprachliches Lernen: Modelle, Empirie, Evaluation: Models, Empiricism, Assessment*, ed. Hu Adelheid and Michael Byram, 173–178. Narr Francke Attempto Verlag.

Nold, Günter, and Henning Rossa. 2008. Sprachbewusstheit Englisch. In *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie*, ed. DESIKonsortium (Hrsg.), 157–169. Weinheim: Beltz.

O'Keeffe, Anne, and Fiona Farr. 2003. Using Language Corpora in Initial Teacher Education: Pedagogic Issues and Practical Applications. *TESOL Quarterly* 37: 3.

O'Keeffe, Anne, Michael J. McCarthy, and Ronald Carter. 2007. *From Corpus to Classroom*. Cambridge: Cambridge University Press.

Ogiermann, Eva. 2009. *On Apologising in Negative and Positive Politeness Cultures*. Amsterdam: John Benjamins.

Olshtain, Elite and Andrew D. Cohen. 1991. Teaching Speech Act Behavior to Nonnative Speakers. In *Teaching English as a Second Or Foreign Language*, ed. Mariam Celce-Murcia, 154–165. Boston, MA: Heinle & Heinle.

Oxford English Dictionary Online. https://www.oed.com/view/Entry/184948. Accessed January 2022.

Padilla Cruz, Manuel. 2013. An Integrative Proposal to Teach the Pragmatics of Phatic Communion in ESL Classes. *Intercultural Pragmatics 10*: 131–160.

Rose, Kenneth R. 2007. Teachers and Students Learning About Requests in Hong Kong. In *Culture in Second Language Teaching and Learning*, ed. Eli Hinkel, 167–180. Cambridge: Cambridge University Press.

Rutherford, William E. 1987. *Second Language Grammar: Learning and Teaching*. London: Longman.

Sandu, Oana Nestian, and Nadine Lyamouri-Bajja. 2018. *T-Kit 4. Intercultural Learning*. Council of Europe.

Sapir, Edward. 1929. The Status of Linguistics as a Science. *Language* 5: 207–214.

Schegloff, Emanuel. 1982. Discourse as an Interactional Achievement: Some Uses of 'uh huh' and Other Things That Come Between Sentences. In *Analyzing Discourse: Text and Talk*, ed. Deborah Tannen, 71–93. Washington, DC: Georgetown University.

Schneider, Edgar. 2018. The Interface Between Cultures and Corpora: Tracing Reflections and Manifestations. *ICAME Journal* 42: 97–132.

Storey, John. 2017. The Politics of Culture. *Middle East – Topics & Arguments*. meta-journal.net. https://archiv.ub.uni-marburg.de/ep/0003/issue/view/179. University of Marburg.

Street, Brian. 1993. Culture is a Verb: Anthropological Aspects of Language and Cultural Process. In *Language and Culture*, ed. David Graddol, Linda Thomson, and Michael Byram, 23–43. Clevedon: Multilingual Matters.

Stubbe, Maria. 1998. Are You Listening? Cultural Influences on the Use of Supportive Verbal Feedback in Conversation. *Journal of Pragmatics.* 29 (3): 257–289.

Taguchi, Naoko. 2011. Teaching Pragmatics: Trends and Issues. *Annual Review of Applied Linguistics 31*: 289–310.

Thomas, Jenny. 1983. Cross-Cultural Pragmatic Failure. *Applied Linguistics* 4: 91–112.

Thornbury, Scott. 2004. *How to Teach Grammar*. Harlow, Essex: Longman.

Wardhaugh, Ronald. 2003. *An Introduction to Sociolinguistics*. London: Blackwell.

Widdowson, Henry G. 1998. Context, Community, and Authentic Language. *TESOL Quarterly* 32: 705–716.

Williams, Raymond. 1981. *Culture*. London: Fontana.

Willis, David. 1990. *The Lexical Syllabus: A New Approach to Language Teaching*. London: Collins Cobuild.

Zareva, Alla. 2017. Incorporating Corpus Literacy Skills into TESOL Teacher Training. *ELT Journal* 71 (1): 69–79.

Footnotes

1 Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (2008).

2 My own translation (K. H.).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_10

10. World Englishes and the Second Language Classroom: Why Introducing Varieties of English Is Important and How Corpora Can Help

Sarah Buschfeld[1] ✉ and Emily Rose Weidle[1] ✉
(1) TU Dortmund University, Dortmund, Germany

✉ **Sarah Buschfeld (Corresponding author)**
Email: sarah.buschfeld@tu-dortmund.de

✉ **Emily Rose Weidle**
Email: emily.weidle@tu-dortmund.de

Keywords English Language Teaching (ELT) – Teaching linguistic variation – Corpus Linguistics – Student and teacher attitudes – Linguistic prejudices and stereotypes – Native speaker ideal – Varieties of English

10.1 Introduction

English has been spreading around the globe ever since the heyday of the British Empire. As a result, different first- and indigenized second-language varieties of English have emerged, e.g. Australian, Singaporean, Indian, and Nigerian Englishes. In addition, due to the forces of our modern and globalizing world, English is being increasingly taught as the first foreign language in schools around the globe, but with a strong teaching focus on traditional native speaker norms, prominently British (BrE) and American English (AmE). Other

varieties of English have only recently found their way into curricula and textbooks, and their presentation and prominence leave much to be desired. However,

> [t]he rapid and remarkable shifts in the linguistic tide worldwide over the last quarter of a century have challenged English-language learning and teaching in unprecedented ways. The spread and natural evolution of English itself, combined with the transience in the population of English-language users, have forced a re-examination of the goals of English-language learning and teaching, as well as a reconceptualization of the English language itself […]. (Nero 2012: 153)

We are therefore confronted with two interlinked realities that urgently need to be catered for in English language teaching (ELT).[1] As World Englishes research and related disciplines have shown, the English language cannot be conceptualized as a monolithic whole. With ever-increasing globalization and a worldwide speaker population in which native speakers of English are clearly outnumbered by non-native speakers, preparing students for "effective communication across varieties" (Kachru 2005) should be the ultimate aim of ELT. As Matsuda and Friedrich (2012: 17) state, the heterogeneity of the English language "challenges some of the fundamental assumptions of English language teaching (ELT) and requires that we revisit our pedagogical practices".

In this chapter, we bring together different linguistic approaches and disciplines, i.e. the study of English as a world language, the teaching of English, and corpus linguistics. Our aim is to outline and discuss the pedagogical implications of World Englishes research and related paradigms and the opportunities provided by corpus linguistic approaches to impart the heterogeneity of English and its various repercussions in the ELT classroom. We postulate the need to expose students to different varieties of English (see also Matsuda 2002) and claim that introducing different varieties of English and the concomitant social and cultural values via corpus linguistic oriented pedagogies can be fruitful for students in two ways. First of all, they obtain a better understanding of the heterogeneity of the English language and are exposed to different varieties and not just the two

standard varieties, BrE and AmE. This will give them far better skills for later life than any standard-oriented language teaching since the chances that they meet speakers of BrE or AmE in business encounters are much smaller than meeting speakers from an, e.g., Indian, Russian, or Chinese background. Secondly, they will be introduced to technical skills for data mining, which are needed for the analysis and exploitation of large-scale text- and data bases. Since data collection and scraping have become ever so easy and fast, the relevant information in mega data sets cannot be accessed on the basis of manual work alone and these skills have gained in importance.

To this end, we inquire into the following practical question: How can corpus linguistics be implemented as a useful tool for increasing students' awareness of the linguistic and cultural heterogeneity of English as a world language?

In Sect. 10.2, we provide a brief introduction to the historical foundations and the evolution of English worldwide as well as to some theoretical findings of World Englishes research and related paradigms and their repercussions for the ELT classroom. We further introduce corpus linguistics as one of the major approaches employed for analyzing varieties of English (Sect. 10.2). In Sect. 10.3, we pinpoint the theory-practice divide between the scientific insights introduced in Sect. 10.2 and their practical implementation in ELT classrooms, using ELT in Germany as an example. We discuss student and teacher attitudes towards varieties other than BrE and AmE, as these are considered to be of crucial importance for a successful restructuring of ELT towards practical consideration of the heterogeneity of the English language (see also Galloway 2017: 15).

In Sect. 10.4, we make practical suggestions for implementing the research findings discussed in Sect. 10.2. We show how corpora and corpus linguistic methods can introduce students to the linguistic and cultural diversity of the English language. The data come from a corpus of first language (L1) child Singaporean English (SingE). We show how the linguistic situation and developments in Singapore question our traditional views of the native speaker. We further discuss selected examples from the corpus and illustrate how they present linguistic characteristics typical of SingE and how these, in turn, reflect the regio-

specific cultures. As an integral part of this discussion, we consider how students would profit from such approaches.

10.2 An introduction to World Englishes and Corpus Linguistics

10.2.1 World English: What Needs to Be Understood

Ever since its genesis in the mid-fifth century, the English language has undergone a vast variety of linguistic changes at all levels of linguistic description, i.e. phonology, morphosyntax, lexis, and pragmatics. Language contact has been identified as one of the driving forces behind such changes, in its various stages of development and in particular in the spread of English as a world language through and in the aftermath of British colonization. In three diasporas, English has spread to mostly all corners of the world: First, it was transported to Wales, Ireland, and Scotland; in the second diaspora, between the seventeenth and nineteenth centuries, it spread overseas to North America, Australia, and New Zealand; the third diaspora brought English to other diverse linguistic and cultural contexts, most importantly, different parts of Asia and Africa, via colonial expansion between the seventeenth and twentieth centuries (Kachru et al. 2006: 13–14). Therefore, the language has come into contact with a large number of languages of different language families, typologies, and cultural backgrounds, e.g. Celtic languages, North American and Antipodean languages, and a variety of African and Asian languages. Different types of contact intensity, ways of language acquisition, and "universal laws of ontogenetic second-language acquisition and phylogenetic language shift", e.g. simplification and overgeneralization (Schneider 2007: 89; see also Williams 1987: 169–70), have led to the emergence of first and second language varieties all around the globe and beyond the historical parent variety of BrE. As a consequence of these expansionist processes and contacts, English can no longer be conceptualized as a single or monolithic language oriented towards the erstwhile metropolis but needs to be understood as a rich set of so-called 'World Englishes', each with their own history, structural properties, and contexts of use. These insights have led to the birth of

the World Englishes research paradigm and related fields of study, such as Global Englishes, International Englishes, and English as a Lingua Franca (ELF). Despite some differences in precise orientation (e.g. Galloway 2017 for an overview), they all account for the same phenomenon, i.e. the spread of English worldwide, its different manifestations and usage contexts and conditions, and, to different extents, the implications for ELT.[2]

In recent decades, English has been the most prominent language in academia, international business, politics, and diplomacy. It is the language of the media and popular culture, and the most often taught second language worldwide (see also Galloway 2017: 2).

The commercialization of the Internet and its rapid spread into private households worldwide since the 1990s, the introduction of the smartphone, and an increasing number of widely used communication and social networking platforms have made worldwide communication easier than ever before. These resources have long reached global audiences and have promoted access to English almost everywhere and this has created even more complex realities of the distribution of the English language and contact with other languages—and new and unprecedented challenges for ELT (cf. Galloway 2017: 1 for a similar argument).

In general, these new challenges have been most prominently discussed as part of the ELF approach, which does not focus on the use of English in circumscribed geographical settings but in communicative encounters between speakers of different first languages and more or less fluid and dynamic communities of practice (Wenger 1998). As Wang and Jenkins (2016: 39) point out:

> Undoubtedly, the scholarly insights into the development of ELF have implications for English education in the NNESs'[non-native English speaking] contexts. Given that the unprecedented global spread of English has caused changes to the role of English for NNESs in intercultural encounters. (Jenkins 2015; Seidlhofer 2011), it is time to rethink the subject matter of English education today (Jenkins 2012; Widdowson 2003)

However, realities are still characterized by a "mismatch between the language presented in the ELT classroom and how it functions as a Lingua Franca" (Galloway 2017: 9) or how English is used beyond the British or American contexts. What clearly distinguishes ELF encounters from the nation-based description of varieties of English is that in ELF communication, a direct connection between language and a particular culture does not necessarily exist. This, however, is an important component of World Englishes research and should likewise be considered important for ELT. How this aspect can be investigated on the basis of speech corpora is briefly illustrated in Sect. 10.2.2 and further discussed in Sect. 10.4.

In general, World Englishes research has reacted to the complex developments and realities of English worldwide ever since its early days. Researchers in the paradigm have refined their assumptions, approaches, and models to better understand the complex realities of the development, distribution, functions, and uses of the English language today. They have described the development and characteristics of geographically-defined varieties of English (e.g. Kortmann and Schneider 2004; Kortmann et al. 2020); business encounters between non-native speakers as lingua franca uses and the communication dynamics and linguistic characteristics of such exchanges (e.g. Seidlhofer 2005, 2011; Jenkins 2000, 2007); and, more recently, as grassroots Englishes, i.e. usage and communication forms of rather low proficiencies, which are acquired for utilitarian purposes and beyond any formal education (e.g. Schneider 2016). However, since "the prevailing monolingual myth in ELT perpetuates favourable attitudes towards both the idealised construct of 'native' English and the 'native' English speaking teacher" (Galloway 2017: 1), ELT classrooms have widely failed to adequately accommodate the heterogeneity of the language, despite some first attempts to account for it in textbooks and curricula (cf. Sect. 10.3).

10.2.2 Studying World Englishes Through Corpora

One of the most prominent and well-suited approaches to studying linguistic variation is the corpus linguistic approach (for a more detailed introduction of the approach, see the introduction to this volume). As part of the World Englishes paradigm, a number of corpora

of different varieties of English have been assembled over the last five decades. The first systematically assembled English language corpus belongs to the Brown Family of corpora.

Whereas the Brown corpora exhibit a major focus on the traditional native varieties of English, in particular BrE and AmE varieties, the ICE (International Corpus of English 2017) project initiated by Sidney Greenbaum in 1990 has a strong focus on second language varieties and therefore reflects the turn towards giving non-native Englishes a voice in the World Englishes complex.

The most recent large-scale corpus addition to Corpus Linguistics and the World Englishes paradigm is the Global Web-Based Corpus of English (GloWbE), a 'big data' corpus of approximately 1.9 billion words of informal blogs (ca. 60%) and more formal texts (ca. 40%) scraped from the internet (Davies 2013). It, too, represents a large variety of Englishes from around the globe (for a detailed summary and broader approach to the topic of World Englishes and corpus linguistics, see, e.g., Lange and Leuckert 2020).

Further corpora exist, for example learner corpora such as those provided by the Institute for Language and Communication of the UCLouvain (https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html). Apart from these and similar 'institutionalized' corpora which are available to a wider public, researchers also collect their own corpora, which are often of much smaller scale but come with their very own advantages, most importantly detailed, first-hand information on speaker profiles and their individual linguistic backgrounds. The CHEsS corpus (Children's English in Singapore; Buschfeld 2020), for example, contains spoken data from 37 Singaporean children aged 1;4 (1 year; 4 months) to 12;1, of mainly Indian and Chinese descent. The data were elicited by means of different data collection methods: language attitudes and use questionnaires (filled in by the parents and containing detailed information on the demographic and sociolinguistic background of each child), free interaction with and between the children, a story retelling task, the Rice/Wexler Test (aimed at eliciting particular grammatical structures such as past tense markings; Rice and Wexler 2001), and a self-designed picture naming task. The corpus consists of

48,360 words as produced by the children and their orthographic transcriptions (see Buschfeld 2020: Ch.4).

Using corpus linguistics can help identify and compare linguistic characteristics of varieties of English (e.g. the eWAVE project; Kortmann et al. 2020) or exemplify the link between language and culture (e.g. Schneider 2018). Rethinking and reconceptualizing this relationship has also been identified as a major implication for ELT (Warschauer 2000: 514). Coming from a linguistic perspective, Schneider (2018), for example, shows how culture is reflected through different varieties of English via three layers of reflections, i.e. cultural objects, dimensions of cross-cultural analysis, and syntactic constructions, and how the corpus linguistic approach can be utilized to assess the occurrence and weight of such cultural markers. We come back to this in Sect. 10.4 and outline how both aims of World Englishes corpus linguistics could be transferred to the practical level and thus the ELT classroom.

Using corpora in the classroom is not a new approach in principle. A large number of publications focus on how corpora could be exploited in language teaching, dating back as far as the 1990s. We cannot portray the full scholarly treatment of the application of corpora in the (English) language classroom here. For more precise and detailed suggestions of how to employ corpora in the classroom, the interested reader is referred to, for example, Bennett (2010), Cobb and Boulton (2015), or McEnery and Xiao (2010), and to chapters in the present volume that deal specifically with this question.

10.3 World Englishes in the Classroom: Scholarly Reception vs. Current Realities

10.3.1 The Theory-Practice Divide

World Englishes research and related, more teaching-oriented disciplines such as Global/International Englishes and ELF research have clearly succeeded in raising awareness of the plurality of the English language and the linguistic, communicative, and cultural implications for ELT (e.g. Galloway and Rose 2015, 2017; Jenkins 2015; Matsuda 2017; Syrbe 2018). Even though these approaches constitute a

step in the right direction, the situation is still characterized by a strong theory-practice divide since practical suggestions for implementing these considerations and actual applications are still largely missing. An important factor here are the students' and teachers' attitudes towards varieties of English and their still clearly prevailing preference for native English norms in many ELT settings (Galloway 2017: 15). In this respect, teacher education plays a crucial role for changing this longstanding mindset. First attempts have been made through teacher training programs incorporating classes on Global/International English (e.g. Matsuda 2017 for an overview), but these are still far from being incorporated in the standard curricula of teacher training and university education. The same is true for teaching materials. Even though varieties of English have found their way into current textbooks, ELT still focuses on traditional native speaker norms (i.e. British and American English) and the treatment of varieties other than these remains rather superficial and theoretical (cf. Galloway 2017: 17–18). Therefore, it is high time that the awareness of the heterogeneity of the English language is more widely recognized by concerned parties including teachers and policy makers.

10.3.2 Student and Teachers' Attitudes and Perceptions

As pointed out in Sect. 10.3.1, student's perceptions of and attitudes to varieties of English need to be addressed and ultimately corrected for successfully reorienting ELT towards more up-to-date conceptions and teaching of the English language. In this respect, Meer et al. (2021) analyze the 'folk linguistic perceptions' of Englishes among German learners of English. Their investigation of 166 upper secondary level school students (*Sekundarstufe II*, *Gymnasium*) from four different schools within North-Rhine Westphalia, Germany, has revealed that students show general awareness of varieties of English other than BrE and AmE, but that they consider BrE as the norm and AmE as most likeable and cool, while attaching mostly negative, stereotyped values to varieties such as IndE (Meer et al. 2021). When it comes to the perceptions of English language teachers, a study by Sadeghpour (2020) shows that a similar bias towards standardized ELT exists, together with prejudices against varieties other than BrE and AmE. The data come from 56 English language teachers of different ethnicities

based in Australia and were collected by means of semi-structured interviews and questionnaires on the participants' sociolinguistic and demographic backgrounds. Most importantly for the current chapter, her findings show that the teachers believe that an exposure to World Englishes in ELT increases students' awareness of varieties of English.

10.3.3 Current Realities in the German Classroom

As pointed out in Sect. 10.3.1, teaching varieties of English and counteracting old, stereotyped attitudes towards varieties other than BrE or AmE have not sufficiently reached the practical level yet, neither internationally nor in the German ELT classroom (Callies et al. 2021: 1). In accordance with student and teacher attitudes (cf. Sect. 10.3.2), orientation towards BrE and AmE as norms and standards is still at the core of German curricula (cf. *Ministerium für Schule und Weiterbildung* (*MSW*)[3] 2008, 2013, 2018, 2019). In German secondary school, the focus is on teaching Inner Circle varieties (i.e. BrE, AmE, AusE, NZE) and postcolonial literature and culture. Postcolonial (second language) varieties of English are largely neglected (Meer et al. 2021: 6). In this respect, Bieswanger (2008, 2012) and Syrbe (2018) identify a lack of teaching of non-British and non-American varieties in lower-level secondary school (*Sekundarstufe I*). They stress that ELT mainly focusses on communicative competence during these years. This is, in principle, a good approach, but true global communicative competence cannot be developed on the basis of teaching BrE and AmE only. In comparison, Meer et al. (2021), investigating the teaching of varieties of English at upper-level secondary school (*Sekundarstufe II*), were able to identify learning objectives in the German curricula that are associated with Global English Language Teaching (GELT) (Galloway and Rose 2015), such as acknowledging English as a Lingua Franca (Meer et al. 2021: 6). Even though the focus does not explicitly lie on World Englishes but on acquiring intercultural communicative competence (MSW 2019: 8), students encounter spoken data from outer circle varieties in the classroom (e.g. AfrE or IndE). The term English as a "Global Language" is mentioned in the curriculum (MSW 2019: 8) and India, New Zealand, Africa, and Australia are listed as English-speaking countries that should be considered in the classroom beyond Great Britain and America; the focus, however, is again on intercultural

communicative competence as an integral part of Lingua Franca uses of English rather than the varieties as such (MSW 2019: 8). In general, learning objectives, i.e. what to acquire in terms of lexis, grammar, and pronunciation, are geared towards the two standard varieties; Outer Circle varieties are not taken into account (MSW 2019: 17). This trend also shows in ELT materials. Most of the lower-level secondary school English textbooks are set in Great Britain as far as their main characters and places are concerned (Klett *Green Line*, Cornelsen *English G 21*, Westermann *Camden Town*). However, in upper secondary level, teaching materials (e.g. Schöningh Verlag *Pathway*) are based on the curriculum for the *Zentralabitur* (central German high-school diploma) and therefore include topics such as 'The British Empire', 'The American Dream', 'Voices from Africa', 'Globalization', 'Utopia', and also 'Global Englishes' (MSW 2018: 22). The textbook *Pathway* (Edelbrock 2015) includes a section on 'English, Englishes…Globish-English Around the World', which discusses Paul Roberts' article 'Set us free from standard English'.

In general, the new German curriculum (as of 2018) seems to at least partly include Global Englishes, but this rather new addition urgently needs to be promoted.

10.4 Teaching World Englishes Through Corpora: Some Practical Considerations

In the following sections, we sketch out two interrelated options to introduce variation in English in ELT. We aim at the senior classes (B1 and B2 CEFR levels) of, for example, German *Gymnasia* (grammar schools) since detailed work with authentic speech corpora and understanding linguistic variation requires solid language competence.

On the basis of authentic speech samples from Singapore, we first document how the traditional native speaker ideal needs to be reconsidered in order to widen the students' perspective on the English language. We then show how, on the basis of the CHEsS corpus (Buschfeld 2020), both oral and transcribed speech data can be utilized to illustrate linguistic variation in World Englishes. In a final step, we illustrate how intercultural knowledge and understanding, which are

an integral part of central European curricula and can be of crucial importance in negotiating pragmatic meaning in daily encounters (e.g. business encounters, university teaching, or simply intercultural understanding in private exchanges), can be increased and can profit considerably from working with authentic corpus data from varieties of English. This approach is meant to serve as an example for how to tackle any variety or heterogeneous speech situation.

10.4.1 Reconsidering the Traditional Native Speaker Ideal

The generally acknowledged criteria and definitions for native speakerism include a number of different aspects, revolving around the idea that "[t]he first language a human being learns to speak is his native language; he is a native speaker of this language" (Bloomfield 1933: 43). Cook (1999: 185) lists the following criteria for native-speaker status: (1) subconscious knowledge of rules; (2) intuitive grasp of meanings; (3) ability to communicate within social settings; (4) range of language skills; (5) creativity of language use (cf. Stern 1983: 344); (6) identification with a language community (Johnson and Johnson 1975: 227); (7) ability to produce fluent discourse; (8) knowledge of differences between his or her own speech and that of the "standard" form of the language; (9) ability "to interpret and translate into the L1 of which she or he is a native speaker" (Davies 1996: 154). However, these criteria no longer apply to traditional native speakers only. A number of former second language Englishes are nowadays acquired as first languages by ever-increasing cohorts of children, and Singapore English (SingE) pioneers this development (Buschfeld 2020).

As the result of nearly 150 years of British rule (1819–1965), English was introduced to Singapore as a second language. However, due to a number of sociolinguistic developments and language-political decisions, English started gaining ground as a first language some time in the 1980s. This trend has increased, in particular in the last three decades. The 2020 Census reports that in the group of 5 to 14-year-olds, 77.4% of the Chinese, 63% of the Malay, and 69.8% of the Indian segments of the population nowadays speak English as the most frequently used home language (Department of Statistics Singapore 2020: 29). In comparison, in central London, 78% of all households reported to have English as their main language (Office for National

Statistics 2011); in the US, approximately 78.7% of the overall population speak English as their home language (U.S. Census Bureau 2013–2017). Even though the latter numbers include English-only speakers across generations and the cases are not fully comparable, the comparison shows that the number of Singaporeans who have English as their most important language at home is only marginally lower than the US or central London numbers. In addition, the central criteria for native speakerism are fulfilled by the Singaporean children. Still, speakers of L1 SingE are not publicly accepted as native speakers of English.

Whatever reasons motivate such a neglect—they basically revolve around nescience and attitudinal factors—the Singaporean example clearly shows that we need to reconceptualize our native speaker ideals.

The only way to overcome such stereotyping and negative attitudes is to address and clarify such misconceptions and prejudices, and schools would constitute the ideal mouthpiece for this aim. This can be done in several ways. Corpus-based teaching, for example, can exemplify how such new L1 Englishes are characterized by their very own linguistic, often local, characteristics, while at the same time displaying the exact same level of speaker fluency when compared to the traditional native varieties of English (cf. Sect. 10.4.2).

10.4.2 Teaching Linguistic Variation Via Spoken and Written Material from Corpora

To facilitate the students‘ understanding of linguistic diversification, i.e. the development of ever more varieties and communities of practice of English, through the emergence of local linguistic structures and usage forms, authentic speech corpora can be utilized. We will illustrate this on the basis of a sample analysis of CHEsS corpus data (Buschfeld 2020).

If audio- or video files are available, the teacher can, first of all, play examples of (audio)-recordings to the students to provide a first impression of what the variety of English sounds like. Phonological variation is often the most striking aspect of speech varieties and the first to be noticed by listeners. Anyone who has ever listened to authentic corpus data knows that such data is very different from what

is prototypically presented as oral sound samples accompanying most traditional textbooks. Teachers may ask the students to describe why and in what way they think the respective variety of English is different from what they have so far encountered in school. L1 SingE, for example, is characterized by a number of local pronunciation features (Buschfeld 2020: 141–142).

Grammatical, lexical, and pragmatic features of a variety can also be identified on the basis of audio data. However, transcribed corpus data might be the better choice to introduce such characteristics since they give students more time to really focus on the structures. The following extracts from the CHEsS corpus (examples 1 through 5) illustrate some features of L1 SingE (Buschfeld 2020: 143–154).

1. Zero subject pronouns

   Researcher: Ah, there is the CD, right! Wow, you found it. Very good.

   Child (5;0, female, Chinese): **[Ø IT]** Was in here.

2. Zero plural inflection

   Child (5;8, female, Chinese): First, the three little pig**[Ø PL.]** hug her mommy and kiss her mommy […].

3. Local past tense marking strategies (here: *finish* as postverbal past tense marker)

   Child (6;7, female, Chinese): He **eat finish** everything.

4. Local discourse particles

   Child (6;9, female, Chinese): Help to pack, **leh**! (indicates tentative suggestion or request)

5. Reduplication (for emphasis)

   Child (6;9, female, Chinese): Oh, yah, **can can**.

If teachers cannot draw on their own linguistic knowledge to identify and describe such features, they can always resort to research publications that introduce and describe the respective characteristics. However, this will be time-consuming on the part of the teacher. As a less work-intensive alternative, teachers could simply focus on the overt linguistic manifestations or differences.

Teachers could either present students with examples and their respective explanations or, ideally, let the students discover linguistic variation themselves by working on the corpus transcripts, for example, with the help of concordance software (see below for further details). This can be done via an indirect/soft or a direct/hard data-driven learning approach, both well-known and widely discussed in the ELT literature (e.g. Gilquin and Granger 2010; Gabrielatos 2005; Hadley 2002; Timmis 2015; and Templeton and Timmis, this volume). The teacher could then explain the linguistic background of the feature (e.g. sociolinguistic or language typological reasons for their emergence) and maybe even introduce the linguistic terms (though this would be a secondary aim).

The field of corpus linguistics offers a number of software programs for concordancing and further analysis, such as AntConc (Anthony 2022) or WordSmith Tools (Scott 2020). These programs and methods cannot be introduced in detail, but see, for example, Lange and Leuckert (2020: Ch.3) for some further details on these programs and Bennett (2010), Cobb and Boulton (2015), or McEnery and Xiao (2010) for suggestions for the implementation of such and similar methods in (ELT). Their implementation certainly requires some additional skills on the part of the teacher. However, students may strongly profit from an introduction to corpus software, since being able to handle a variety of different programs and to process and automatically filter out information from large texts constitute important skills in our ever-digitizing age.

10.4.3 Increasing Intercultural Awareness and Understanding Via Corpora

As pointed out in Sect. 10.2.2, corpora have also been used to show how language reflects culture. Since teaching and increasing intercultural awareness and competence has long been part of the worldwide

curricula (e.g. teaching English as a Lingua Franca (MSW 2013: 11)), this approach may also be fruitful for the ELT classroom (see Riegler, this volume). On the basis of an analysis and comparison of different ICE corpora (cf. Sect. 10.2.2), Schneider (2018), for example, shows how culture- and regio-specific objects, habits, plants, and animals reflect different cultures in corpora. As a further important facet that differentiates cultures from one another he points to the dimension of whether a culture pursues and promotes collectivism or individualism. Generally, western cultures traditionally value individualism whereas eastern and Asian cultures orient towards collectivism and the community (Fang 2012: 28). Schneider (2018: 109–111) shows how this sociocultural dimension finds expression in language choice and use.

His corpus analysis clearly illustrates that collectivism is much stronger in the Asian and African varieties than in BrE. Transferred to our real lives, such knowledge might be of crucial importance for the success of, for example, intercultural business encounters.

To transport this knowledge to the classroom, teachers could first ask the students whether they can think of similar examples from their own experiences with language variation (if they have any) and collect ideas in form of a mind-map, poster, or padlet. In a next step, teachers can either orally and/or in writing present such examples to the students or let them work out similar examples on the basis of corpus material. Again, the latter approach would be pedagogically more efficient and valuable but also the more difficult approach for both the teacher and the students (cf. our discussion in Sect. 10.4.2).

If such insights were taken up in ELT, they would clearly help increase the understanding of the connection between language and culture and how the latter is reflected by the former (see also Harrington, this volume). With a class of advanced and (corpus)linguistically experienced students, one could even explore such issues in corpora and relate the findings to insights from sociocultural studies. Being able to take such interdisciplinary viewpoints and applying interdisciplinary methods is of increasing importance in our modern world, but taking such an approach would require that students not only have good competences in standard varieties of English, but also have some solid knowledge of varieties of

English and corpus linguistic methods. To count and compare frequencies, they need to have been introduced to the techniques and programs of corpus linguistics, as briefly introduced in Sect. 10.4.2.

10.5 Conclusion

In the present chapter, we have shown how a stronger integration of varieties of English in the ELT classroom would be advantageous to students in a number of respects. They would be much better prepared for the linguistic realities and communicative encounters outside the classroom, which are normally dominated by non-native speakers of English, who all bring to the conversations their specific varieties of English. We have further shown how debunking the native speaker myth would help to overcome linguistic prejudices and negative attitudes towards speakers of varieties other than BrE or AmE. Finally, we have illustrated how corpora can be utilized to convey linguistic variation and introduce students to different varieties of English and the connections between language and culture.

We are still far away from a full incorporation of such topics and approaches in our teaching curricula worldwide, but we hope to have shown how rethinking our focus in ELT, changing the existing curricula, and, finally, implementing these changes would be worth the effort. Quite a number of universities around the world have successfully implemented World Englishes approaches and corpus linguistic methods in their teaching, and our own teaching experience has clearly shown how students not only profit from such classes and modules but also enjoy this type of hands-on experience. Ultimately, if we manage to equip our future teachers with the knowledge and methodological know-how to fully implement varieties of English via corpus linguistic methods and if they manage to dissipate their own prejudices, that would be half the battle.

Ideally, one day, we not only accomplish to realistically depict the heterogeneity of the English language in the ELT classroom but also to overcome "the prevailing monolingual myth in ELT" (Galloway 2017: 1) and to raise students' awareness that multilingualism has turned into the norm around the world.

References

Anthony, Laurence. 2022. *AntConc (Version 4.0.5) [Computer Software]*. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software.

Bennett, Gena R. 2010. *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers*. Michigan ELT.

Bieswanger, Markus. 2008. Varieties of English in Current English Language Teaching. *Stellenbosch Papers in Linguistic* 38: 27–47.

———. 2012. Varieties of English in the Curriculum. In *Codification, Canons and Curricula: Description and Prescription in Language and Literature*, ed. Anne Schröder, Ulrich Busse, and Ralf Schneider, 359–371. Bielefeld: Aisthesis.

Bloomfield, Leonard. 1933. *Language*. London: Allen & Unwin.

Buschfeld, Sarah. 2020. *Children's English in Singapore: Acquisition, Properties, and Use*. London: Routledge.

Callies, Marcus, Stefanie Hehner, Philipp Meer, and Michael Westphal, eds. 2021. *Glocalising Teaching English as an International Language: New Perspectives for Teaching and Teacher Education in Germany*. London: Routledge.

Cobb, Tom, and Alex Boulton. 2015. Classroom Applications of Corpus Analysis. In *Cambridge Handbook of English Corpus Linguistics*, ed. Douglas Biber and Randi Reppen, 478–497. Cambridge: Cambridge University Press.

Cook, Vivian J. 1999. Going Beyond the Native Speaker in Language Teaching. *TESOL Quarterly* 33 (2): 185–209.

Corpus of Global Web-Based English. https://www.english-corpora.org/glowbe/. Accessed 9 January 2022.

Davies, Alan. 1996. Proficiency Or the Native Speaker: What Are We Trying to Achieve in ELT? In *Principle and Practice in Applied Linguistics*, ed. Guy Cook and Barbara Seidlhofer, 145–157. Oxford: Oxford University Press.

Davies, Mark. 2013. Corpus of Global Web-Based English: 1.9 Billion Words from Speakers in 20 Countries (GloWbE). Available Online at https://corpus.byu.edu/glowbe/.

Department of Statistics Singapore. 2020. Census of Population 2020. Statistical Release 1: Demographic Characteristics, Education, Language and Religion. https://www.singstat.gov.sg/-/media/files/publications/cop2020/sr1/cop2020sr1.pdf. Accessed 22 January 2022.

Edelbrock, Iris. 2015. *The New Pathway Advanced Lese- und Arbeitsbuch Englisch für die Gymnasiale Oberstufe*. Paderborn: Schöningh Verlag.

Fang, Tony. 2012. Yin Yang: A New Perspective on Culture. *Management and Organization Review* 8: 25–50.

Gabrielatos, Costas. 2005. Corpora and Language Teaching: Just a Fling Or Wedding Bells? *TESL-EJ* 8 (4): 1–35.

Galloway, Nicola. 2017. *Global Englishes and Change in English Language Teaching. Attitudes and Impact*. London and New York: Routledge.

Galloway, Nicola, and Heath Rose. 2015. *Introducing Global Englishes*. London: Routledge.

———. 2017. Incorporating Global Englishes into the ELT Classroom. *ELT Journal* 71 (1): 3–14.

Gilquin, Gaëtanelle, and Sylviane Granger. 2010. How Can DDL Be Used in Language Teaching. In *The Routledge Handbook of Corpus Linguistics*, ed. Anne O'Keeffe and Michael McCarthy. London: Routledge.

Hadley, Gregory. 2002. An Introduction to Data-Driven Learning. *RELC Journal* 33 (2): 99–124.

International Corpus of English. 2017. https://www.ice-corpora.uzh.ch/en/joinice/Teams.html. Accessed 9 January 2022.

Jenkins, Jennifer. 2000. *The Phonology of English as an International Language*. Oxford: Oxford University Press.

———. 2007. *English as a Lingua Franca: Attitude and Identity*. Oxford: Oxford University Press.

———. 2012. English as a Lingua Franca from the Classroom to the Classroom. *ELT Journal* 66 (4): 486–494.

———. 2015. *Global Englishes: A Resource Book for Students*. London: Routledge.

Johnson, David W., and Roger T. Johnson. 1975. *Learning Together and Alone: Cooperation, Competition & Individualisation*. Inglewood Cliffs, NJ: Prentice-Hall.

Kachru, Yamuna. 2005. Teaching and Learning of World Englishes. In *Handbook of Research in Second Language Teaching and Learning*, ed. Eli Hinkel, 155–173. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Kachru, Braj B., Kachru, Yamuna, and Nelson, Cecil L. 2006. Introduction: The World of World Englishes. In *The Handbook of World Englishes*, ed. Braj B. Kachru, Yamuna Kachru, and Cecil L. Nelson, 1–14. Malden, MA: Blackwell.

Kortmann, Bernd, and Edgar W. Schneider, eds. 2004. *A Handbook of Varieties of English*. Vol. 2. Berlin: Mouton De Gruyter.

Kortmann, Bernd, Kerstin Lunkenheimer, and Katharina Ehret, eds. 2020. *The Electronic World Atlas of Varieties of English*. Zenodo. https://ewave-atlas.org. Accessed 27 January 2022.

Lange, Claudia, and Sven Leuckert. 2020. *Corpus Linguistics for World Englishes: A Guide for Research*. London: Routledge.

Matsuda, Aya. 2002. "International understanding" Through Teaching World Englishes. *World Englishes* 21 (3): 436–440.

———, ed. 2017. *Preparing Teachers to Teach English as an International Language*. Bristol: Multilingual Matters.

Matsuda, Aya, and Patricia Friedrich. 2012. Selecting an Instructional Variety for an EIL Curriculum. In *Principles and Practices of Teaching English as an International Language*, ed. Aya Matsuda, 17–27. Bristol: Multilingual Matters.

McEnery, Tom, and Richard Xiao. 2010. What Corpora Can Offer in Language Teaching and Learning. In *Handbook of Research in Second Language Teaching and Learning*, ed. Eli Hinkel, 364–380. London and New York: Routledge.

Meer, Philipp, Johanna Hartmann, and Dominik Rumlich. 2021. Folklinguistic Perceptions of Global Englishes Among German Learners of English. *European Journal of Applied Linguistics* 9 (2): 391–416.

Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen (MSW). 2018. *Zentralabitur 2022 – Englisch*. https://www.standardsicherung.schulministerium.nrw.de/cms/zentralabitur-gost/faecher/getfile.php?file=4989. Accessed 27 January 2022.

Ministerium für Schule und Weiterbildung des Landers Nordrhein-Westfalen (MSW). 2008. In *Richtlinien und Lehrpläne für die Grundschule in Nordrhein-Westfalen*. Frechen: Ritterbach. https://www.schulentwicklung.nrw.de/lehrplaene/upload/klp_gs/LP_GS_2008.pdf. Accessed 27 January 2022.

———. 2013. *Kernlehrplan für die Sekundarstufe II Gymnasium/Gesamtschule in Nordrhein-Westfalen – Englisch*. Frechen: Ritterbach. https://www.schulentwicklung.nrw.de/lehrplaene/lehrplan/19/KLP_GOSt_Englisch.pdf. Accessed 27 January 2022.

———. 2019. *Kernlehrplanfür die Sekundarstufe I Gymnasium in Nordrhein-Westfalen – Englisch*. Frechen: Ritterbach. https://www.schulentwicklung.nrw.de/lehrplaene/lehrplan/199/g9_e_klp_%203417_2019_06_23.pdf. Accessed 19 January 2022.

Nero, Shondel. 2012. Languages Without Borders: TESOL in a Transient World. *TESL Canada Journal* 29 (2): 143.

Office for National Statistics. 2011. *2011 Census: Population and Household Estimates for the United Kingdom*. https://www.nomisweb.co.uk/census/2011/ks206ew. Accessed 22 January 2022.

Rice, Mabel L., and Kenneth Wexler. 2001. *Rice/Wexler Test of Early Grammatical Impairment. Examiner's Manual*. San Antonio, TX: The Psychological Corporation.

Sadeghpour, Marzieh. 2020. *Englishes in English Language Teaching*. London: Routledge.

Schneider, Edgar W. *Postcolonial English: Varieties Around the World*. Cambridge: Cambridge University Press.

———. 2016. Grassroots Englishes in Tourism Interactions. *English Today* 32 (3): 2–10.

———. 2018. The Interface Between Cultures and Corpora: Tracing Reflections and Manifestations. *ICAME Journal* 42: 97–132.

Scott, Mike. 2020. *WordSmith Tools Version 8*. Stroud: Lexical Analysis Software.

Seidlhofer, Barbara. 2005. English as a Lingua Franca. *ELT Journal* 59 (4): 339–341.

———. 2011. *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Sketch Engine. https://www.sketchengine.eu. Accessed 9 January 2022.

Stern, Hans Heinrich. 1983. *Fundamental Concepts of Language Teaching*. Oxford: Oxford University Press.

Syrbe, Mona. 2018. Evaluating the Suitability of Teaching EIL for the German Classroom. *International Journal of Applied Linguistics* 28 (3): 438–450.

Timmis, Ivor. 2015. *Corpus Linguistics for ELT. Research and Practice*. London and New York: Routledge.

U.S. Census Bureau. 2013–2017. American Community Survey 5-Year Estimates. https://archive.ph/20200214060808/https:/factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_5YR_S1601&prodType=table#. Accessed 22 January 2021.

Wang, Ying, and Jennifer Jenkins. 2016. "Nativeness" and Intelligibility: Impacts of Intercultural Experience Through English as a Lingua Franca on Chinese Speakers' Language Attitudes. *Chinese Journal of Applied Linguistics* 39 (1): 38–58.

Warschauer, Mark. 2000. The Changing Global Economy and the Future of English Teaching. *TESOL Quarterly* 34 (3): 511–535.

Wenger, Etienne. 1998. *Communities of Practice: Learning, Meaning, and Identity*. Cambridge: Cambridge University Press.

Widdowson, Henry G. 2003. *Defining Issues in English Language Teaching*. Oxford: Oxford University Press.

Williams, Jessica. 1987. Non-native Varieties of English: A Special Case of Language Acquisition. *English World-Wide*, 8. Amsterdam: John Benjamins, 161–199.

Footnotes

1 As we are not convinced of the strict separation between second and foreign language English for a variety of reasons that cannot be discussed as part of the chapter, we will use ELT (classroom) as a cover term without making any implications about the sociopolitical status of a language.

2 If we do not explicitly aim to point out the specifics of the individual approaches, we use the label "World Englishes" as a cover term for all linguistic approaches concerned with the spread, heterogeneity, and use of English worldwide.

3 The *Ministerium für Schule und Weiterbildung* (*MSW*) (Ministry of School and Education) is the school or education ministry of the German state of North Rhine-Westphalia and one of twelve ministries of the North Rhine-Westphalian state administration.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_11

11. Annotating VOICE for Pedagogic Purposes: The Case for a Mark-up Scheme of Pragmatic Functions in ELF Interactions

Stefanie Riegler[1] ✉
(1) University of Vienna, Vienna, Austria

✉ **Stefanie Riegler**
Email: stefanie.riegler@univie.ac.at

Keywords English as a lingua franca (ELF) – Vienna-Oxford International Corpus of English (VOICE) – Corpus annotation – Pragmatic annotation – Error tagging – English language teaching (ELT) – Learner corpus research (LCR) – Communication strategies – Pragmatic practices – Normativity – Language pedagogy

11.1 Introduction

Research at the interface between linguistic corpora and language pedagogy—though widely debated—has been very prolific. The chapters included in this volume are indicative of this relation. Likewise, the pedagogic relevance of studying corpora of English as a lingua franca (ELF) has become increasingly recognized. After all, the function of English as the currently predominant global lingua franca entails that many, if not most, students nowadays will be learning English primarily to be able to communicate in international, lingua franca contexts. For this reason, descriptions of ELF use necessarily

concern English language classrooms. With that said, the question that arises is how ELF corpora and English language teaching (ELT) relate to each other. In what respect is this link different from more traditional understandings of the relationship between corpus linguistics and pedagogy, and what is it that ELF research proposes for pedagogic implementation?

One aspect that is frequently neglected in the relationship between linguistic description and pedagogic prescription is the role of corpus annotation, i.e. the provision of supplementary interpretative and language-related information to linguistic data in a corpus (Leech 1997: 2). Even if the idea of using corpus tagging, a type of corpus annotation, to mediate between corpora and ELT is quite common in learner corpus research (LCR) and second language acquisition (SLA), its uptake outside of these fields is relatively minimal (for an exception see e.g. Simpson-Vlach and Leicher 2006). As for ELF corpora, these have been principally compiled for linguistic description rather than pedagogic purposes. This, of course, is also reflected in their design, which primarily targets linguists rather than language education stakeholders such as teachers or material designers. Therefore, this paper argues the need for an annotation scheme for ELF corpora that points out aspects of pedagogic relevance in ELF interactions to make them accessible for use in language classrooms.

This contribution zooms in on corpus tagging for pedagogic purposes, addressing the above-mentioned questions on the interplay between ELF corpora and ELT in Sect. 11.2. It critically investigates existing LCR and SLA tagging schemes in Sect. 11.3 to make the case for a non-normative approach in the annotation of ELF corpora. Section 11.4 outlines which aspects are desirable to annotate in corpora of English as a lingua franca communication from an ELF-based language teaching and learning perspective. It establishes the conceptual basis for a pedagogically oriented annotation system for the *Vienna-Oxford International Corpus of English* (VOICE) (VOICE 2021), a one-million-word corpus of naturally-occurring spoken ELF discourse. The aim of developing this annotation scheme for VOICE is to provide an alternative to norm-referenced mark-up systems. These tag deviations from norms of first language (L1) usage as errors irrespective of their functional value. What is needed instead is an annotation scheme that

orients to communicatively significant aspects in ELF interactions, which a pedagogically focused system for annotating pragmatic functions would provide.

11.2 ELF Corpora and ELT

When some 20 years ago the conceptual basis for English as a lingua franca corpora was established, it was already suggested that descriptions of ELF interactions might yield implications for English language teaching (cf. Seidlhofer 2001: 133). Since then the *Vienna-Oxford International Corpus of English*, the corpus of *English as a Lingua Franca in Academic Settings* (ELFA) (ELFA 2008), and the *Asian Corpus of English* (ACE) (ACE 2014) have been built. They have provided fruitful bases for corpus-based research into ELF communication, which has been defined as "*any use of English among speakers of different first languages for whom English is the communicative medium of choice, and often the only option*" by Seidlhofer (2011: 7, original emphasis). And as envisaged, out of this descriptive interest that originally motivated the compilation of ELF corpora, a number of insights of pedagogic relevance have emerged in the meantime.

Before moving on to these, the question that needs to be clarified first is what justifies pedagogic proposals from ELF research. What is it that makes these different from more conventional implications for language classrooms derived from descriptions of English as a native language (ENL), some of which are also represented in this volume? The traditional reasoning is that corpus linguistic findings of 'real' and 'authentic' L1 usage of English provide the benchmark for language students and should therefore directly determine items for English language instruction and learning (see e.g. Römer 2010). This way, descriptions of ENL have become norm-providing and established as standard for language students.

This transfer of insights from corpus linguistics to language education, however, disregards that the context in most foreign language classrooms is rather different from the original context of L1 communication. Likewise, the shift omits that the language as produced by ENL users may not necessarily be feasible for language learning, thereby confounding linguistic 'samples' with pedagogic 'examples'

(Widdowson 2009: 211–214). By contrast, interactants in ELF communication appropriate linguistic resources for communicative purposes and in doing so, "are not borrowing somebody else's language but using the language of their own learning. Their communication is their learnt language put to use" (Widdowson 2009: 214).[1] The way language is exploited to communicative effect in these contexts then shows what elements of the taught language become realized, what practices are drawn upon and what becomes transferred from language learning, all of which are findings that are highly significant for ELT (Widdowson 2009: 214). This being so, the function of linguistic evidence in pedagogic insights derived from ELF research is different. It does not provide a model for replication but indicates how linguistic resources are exploited in lingua franca settings. So corpus data do not yield items for teaching but inform the actual process of learning in guiding language students to recognize communicative functions.

Browsing recent corpus-based studies of ELF interactions reveals that there are indeed many insights of pedagogic relevance to take away from descriptions of ELF communication. One finding relates to the functional significance of multilingual elements in ELF interactions, which challenges monolingual orientations in ELT (Cogo 2018: 359). Another concern in ELF research has been speakers' use of non-conventional idioms and metaphors. Evidence here suggests a reorientation in English language classrooms from obliging students to imitate L1 idiomatic usage to pointing out the functional load that creative idioms and metaphors carry in communication (Pitzl 2018: 243–247). Providing examples of non-codified word-formations from VOICE, Pitzl (2017: 42) proposes that rather than merely accepting the occurrence of innovative lexical items, language teachers should be encouraged to view them as indicators that students have successfully learnt how to apply morphological mechanisms (Pitzl 2017. Equally questioning the deficit perspective in many language classrooms, other ELF studies imply that ELT is disproportionally occupied with having students accurately re-produce lexis and grammar items that prove difficult for them, since they also correspond to those aspects of lexicogrammar that ELF speakers flexibly and variably draw on, and that are unlikely to harm communication (Seidlhofer 2011: 207–208; Cogo and Dewey 2012: 172–173).

Staying with the notion of communicative effectiveness, it has been shown that ELF speakers are much invested in managing interaction and achieving shared understanding. For these purposes, interactants in lingua franca settings draw on a range of pragmatic strategies, which clearly puts these forward as an explicit focus in ELT (Seidlhofer 2011: 205). This obviously highlights the need to equip language students with strategic competencies (e.g. Vettorel 2019: 204) and enhances the role of accommodation skills in the pedagogic context (Cogo and Dewey 2012: 176). More detailed research assuming a pedagogic perspective in the study of the pragmatics of ELF communication identifies more frequent and effective communication strategies for low-proficiency learners (e.g. direct appeal for help or repeating triggers) and recommends that specific attention be paid to these in the classroom (Sato et al. 2019: 30–31). Along these lines, Kaur (2016: 243) suggests that English language instruction should engage students in such pragmatic processes, and outlines the need for materials that target the effectiveness of learners' language use.

This overview of pedagogically relevant findings that analyses of ELF interactions have yielded is certainly cursory and selective. Yet, even this brief survey indicates what emerges as the central proposition that ELF research puts forward for language education: a move from an orientation to L1 *usage* and the approximation of ENL forms to a more pragmatic and communicative conceptualization of language *use* (for a consideration of use/usage see Widdowson 2012: 14–15). In addition, several theoretical frameworks have been proposed to facilitate the integration of findings from research on ELF communication in language teacher education and ELT classrooms such as the concept of an "ELF-informed pedagogy" (Seidlhofer 2011: Ch. 8), the "post-normative approach" (Dewey 2012) or the notion of "ELF awareness" (Sifakis 2019), to mention just a few. Unfortunately, though, the uptake of these certainly significant and fundamental language educational proposals has been somewhat limited. One issue that ELF research reveals (cf. e.g. Seidlhofer 2011: 41; Cogo and Dewey 2012: 173; Widdowson 2012: 22–23) is that current practices in ELT, even when they claim to be communicative, in some respects fail to reflect a genuinely communicative perspective in that the aim still conforms to L1 speakers' language usage. Such a view treats form-function

mappings, i.e. the use of a linguistic form to fulfill a communicative function, as fixed rather than adaptive and, thus, implies a form-focused orientation to language in the classroom (Widdowson 2012: 22). For communicative pedagogies to take account of the underlying theoretical perspectives, they would need to be equally concerned with semantic and pragmatic meanings (Widdowson 2009: 204). The prevailing imbalance towards a focus on linguistic forms, i.e. semantic meaning, in ELT suggests that findings on the way ELF speakers use language for pragmatic purposes have not yet been properly taken on board in English language education. What then, we need to ask, sustains this disproportionate concern with linguistic forms?

11.3 Normative Approaches in Corpus Annotation

As indicated above, the relation between ELF corpora and ELT remains somewhat ambivalent with many pedagogic proposals being put forward that are struggling to gain traction. By contrast, the set of corpora that is essentially dedicated to language pedagogy and puts forward proposals to support language learning, i.e. learner corpora, has had more profound impact upon language education. Granger (2015: 489–494) for instance, illustrates how findings in the context of learner corpus research have found their way into language classrooms through learners' dictionaries or pedagogic grammars. The essential point though is that what is carried over from learner corpora to language pedagogy generally focuses on standard language forms, i.e. the traditional learning target. That is, items for teaching and learning derived from learner corpora traditionally rely on reference to the conventions of L1 speakers' usage.

This may be due to approaches adopted to mark aspects of pedagogic significance in language corpora. As can be seen from established procedures in the tagging of learner corpora (for examples see Dagneaux et al. 1998; Nicholls 2003; Lozano and Díaz-Negrillo 2019), these are prevalently norm-dependent, with the common practice of error tagging that assesses learner performance against a benchmark of accepted L1 usage.

To illustrate this aspect, Extract 1 provides an example from the *NICT Japanese Learner English Corpus*[2] (NICT JLE) (NICT 2012a), a spoken learner corpus consisting of transcripts of audio-recorded L2 English oral proficiency interview tests.[3] The passage exemplifies a focus in the error tagging system on linguistic non-conformities rather than the pragmatic use of linguistic resources.

Extract 1

A = examiner, B = student

9 <A>So, <H pn="B's name">XXX02</H>, are your exams over? <.> </.> Finished?</A>

10 <B><F>Um</F>. <F>Er</F> <SC>yeah</SC> <F>er</F> no. <SC>I don't have</SC> I didn't have a test <SC>this ter</SC> in this term. Because I am a graduate school student. So <F>um</F> <at odr="1" crr="a"></at> freshman of <F>er</F> **<o_lxc odr="2" crr="graduate school">graduated school</o_lxc>**. I <v_tns odr="1" crr="finished">finish</v_tns> <F>er</F> all <n_num odr="2" crr="subjects">subject</n_num>. <F>Eh</F>. Now, <SC>I have</SC> <F>er</F> I only have to <F>er</F> <SC>res</SC> <F>er</F> **<v_lxc odr="1" crr="do">study</v_lxc> my research**. Yes.</B>

11 <A><F>Uum</F>. Can you tell me why you decided to go to a graduate school?</A> (File00546)

Speaker A, the examiner, models active involvement in achieving understanding for the interlocutor by self-rephrasing *are your exams over?* with *Finished?* in utterance 9. That is, the interviewer performs a listener-oriented communicative strategy to increase explicitness. The examinee, speaker B, mirrors this behavior in the following utterance, when he reformulates the term *graduate school student* with a non-conventional expression, *freshman of graduated school*. The term *graduated school* is marked <o_lxc> to display a collocation error and the mark-up suggests a correction (indicated in the tag by crr="…") to *graduate school* (despite the fact that the student has already signaled that *graduate school* forms part of his lexicon). From descriptive findings on ELF interactions we know that increasing explicitness to enhance clarity may yield non-codified expressions displaying pragmatically motivated, variable language use (Seidlhofer 2011: 96).

Utterance 10 shows how the student exploits linguistic resources for pragmatic effect, which is an entirely communicative undertaking. In uttering *freshman of graduated school*, speaker B also performs the pragmatic strategy of paraphrase that research has found to be an effective means in the negotiation of understanding in ELF conversations (for examples in ELF discourse cf. e.g. Kaur 2016: 244–246). Another interesting pragmatic phenomenon occurs at the end of utterance 10, when the student utters the phrase *study my research*. In the NICT learner corpus, this is tagged as a lexical error and corrected to the more established expression *do my research*. Yet, pragmatically, speaker B's expression foregrounds the semantic relation between *study* and *research*, which is a useful communicative strategy to increase explicitness in ELF communication (Seidlhofer 2011: 99). It makes the meaning of the expression more transparent, and is—above all—intelligible to the interlocutor as it does not trigger any indication of non-understanding. In the learner corpus, it is nevertheless marked as outright erroneous.

The second extract uses examples from Dose-Heidelmayer and Götz's (2016) study into learners' use of the progressive. Their analysis draws on a data sample from the well-known and widely used *Louvain International Database of Spoken English Interlanguage* (LINDSEI) (Gilquin et al. 2010)[4] that has been marked up with an adapted version of the Louvain error tagging manual (Dagneaux et al. 2005). Similar to Extract 1, it shows that the error tagging has been applied regardless of the functional motivation that presumably lies behind the non-canonical uses of the progressive aspect in the exchange. Additionally, the mark-up runs contrary to findings on the use of the progressive in ELF interactions.

Extract 2

A = examiner, B = student

147 <B> . well .. I mean it's obvious that there is .. quite a difference between . the picture and .. how **(GVT)** she**'s looking $she looks$** . but .. (mm) <sighs> well . one can see that she's . pleased with the picture that she likes it and . that she'd like to look like it. I think so . which is to say that it's . a marvellous picture . <overlap/> and that . that **(GVT)** she**'s** </B>

148 <A> <overlap/> (mhm) </A>
149 <B> **looking $she looks$** beautiful (LSP) on in it . which . she <starts laughing> does <stops laughing> . but not . not it . it . not in real </B>
150 <B> no </B>
151 <A> no </A>
152 <B> . I don't think so </B>
153 <A> yeah okay . do you think your image of yourself your mirror image that is when you look in the mirror . is different . from what other people see </A>
154 <B> … ye= . yeah I think so </B>
155 <A> . why </A>
156 <B> … (erm) .. (er) I don't know if you're looking . in the mirror . (erm) . it's just .. one or two of (mm) minutes . that you . and I I would if **(GVT)** I**'m looking $I look$** in the mirror I don't smile at myself I don't talk to myself and then . **(GVT)** you**'re looking $you look$** different . when <overlap/> (GVM) you're . just have $you just have$ your </B>
157 <A> <overlap/> (mhm) </A>
158 <B> <overlap/> (LSF) mimics $facial expression$ or . gestures </B>
159 <A> <overlap/> (mhm) </A> (GE006)

This passage involves speaker B, the student, in a picture description task. Throughout the extract, the student's use of the progressive form of *look* has been tagged as *GVT* to indicate errors in the categories grammar (G), subcategory verb (V), subcategory tense (T). In all the tagged cases, the mark-up suggests a correction to the simple aspect since *look* denotes perception and therefore belongs to the category of stative verbs. These are conventionally not used in the progressive form, which means that the occurrences in Extract 2 constitute errors according to prescriptive grammar rules. Yet, findings from ELF interactions indicate that traditionally 'non-standard' uses, e.g. of stative verbs, in the progressive form might be prompted by functional motivations that qualify these occurrences. Ranta (2013: 102), for instance, argues that speakers in ELF contexts draw on the progressive to render the verb more salient "wanting to make sure that the most essential part of their message—the verb—is successfully conveyed". That is, the non-canonical occurrences of *look* in the

progressive could equally derive from the student's effort to ensure communicative effectiveness and represent a pragmatic strategy. Notably in Extract 2, all 'erroneous' uses of *look* in the progressive occur in the picture description part of the exam. Research into ELF interactions outlines that the progressive aspect may be drawn upon to render descriptions more vivid and immediate (Dorn 2011: 14). So an alternative interpretation of the above example may be that the student is engaged in creating a particularly lively description of the picture by using the progressive form of *look*, again indicating how quite naturally the pragmatics of the exchange influence the form it takes.

It is shown that the two tagging schemes presented above provide examples of norm-referenced mark-up systems. What is tagged are deviations from the standard target language to counteract difficulties that learners at a particular proficiency level with a specific L1 background typically experience. This mark-up that identifies 'pitfalls' in learner language should then be used in the classroom to indicate what aspects of the language system need to receive particular attention to avoid errors and to support learners in the approximation of L1 usage side-by-side with items derived from L1 corpora, making sure that the language does not diverge from how it is conventionally used (Granger 2015: 486–488). Such approaches in corpus annotation are likely to favor and continue normative, form-focused orientations in language teaching and learning rather than encourage a focus on the interactive processes language users are engaged in during communication, i.e. language use. While classroom practices focusing on language usage emphasize the linguistic forms and expressions users produce while interacting, approaches orienting to language use would give priority to the process of communication. That is, they would prioritize function over form and orient to the natural use of functions and forms as known to speakers from L1 communication, but which they are discouraged from exploiting in pedagogical contexts.

Additionally, the two illustrative data samples suggest how corpus annotation may contribute to normative approaches that have been conventionally adopted in classroom-oriented corpus-based research. It is difficult to imagine how subsequent research which essentially builds on error-tagged data could take on a non-normative perspective. This means that normativity in corpus annotation is likely to cause an

orientation to the corpus data from a deficit, form-focused and accuracy-centered perspective. While I am aware that not all corpus-based research starts from annotated data and that corpora will always enable an analysis of the raw and untagged data, the essential point that emerges is that corpus annotation for pedagogic purposes as traditionally conceived, may considerably encourage, and provide the means for form-focused approaches to linguistic data.

In contrast to norm-referenced schemes, proposals for mark-up systems that seek to take a less normative approach to learner corpus data are not that common in LCR and SLA (for an exception see e.g. Rastelli 2009). The underlying assumption in the majority of existing pedagogically aimed mark-up schemes appears to be that implications for language learning are to be drawn from tagged non-conformities or errors found in language data. It is then these, rather than successful linguistic practices, that serve as points of reference and samples for language learning. This conventional fixation on the 'inadequacies' in learner usage in pedagogically focused corpus research has considerably suppressed an orientation to the effectiveness of students' language use, those aspects that allow them to manage interaction and that they get right (Aston 2011: 11). "This would be the use of learner language: not to identify what is to be corrected, but what is to be encouraged—a genuine learner-centred approach", as Widdowson (2012: 24) argues. What I would like to put forward in the remainder of this paper is that classroom-oriented corpus annotation—contrary to its traditional, normative interpretation—can be reconceived to highlight what is to be encouraged in language use. Considering the pedagogic relevance of ELF corpora established above, the following section therefore gives insight into the development of an annotation system for pedagogic purposes, one that indicates those aspects of language use that ELF research deems significant.

11.4 Towards Pedagogically Oriented Annotation in an ELF Corpus

As illustrated in Sect. 11.3, pedagogically focused corpus tagging has fallen prey to a one-sided, deficit conceptualization in many cases. This

though does not necessarily have to be the case and there are alternative ways of conceiving corpus annotation that can be exploited for pedagogic purposes. The question then, we need to ask, is what aspects could be exploited and annotated in ELF corpora? Since clearly the focus should not be on language usage, i.e. linguistic forms, and since ELF research calls for a more pragmatically informed conceptualization of language use, the obvious desirable option is communicative function, which Aston (2011: 14) also suggests for communicative classrooms more generally. This would result in an annotation scheme that would point stakeholders in language education, like teachers or materials designers, not to what is to be corrected, in the way pedagogically oriented tagging schemes for English have conventionally done, but what is to be encouraged, thereby informing the process of language learning. Such annotation would indicate what is happening in the process of communication, as opposed to mark-up that focuses on the linguistic product speakers create.

Such an annotation system is now being developed for VOICE, a one-million-word corpus of naturally-occurring, spoken ELF communication that was originally released in 2009 as the first general and freely available ELF database (Breiteneder et al. 2009: 21). The elaborate corpus design reflecting the multilingual and variable nature of ELF interactions makes VOICE most suitable for communicatively oriented corpus annotation. VOICE already supplies pragmatically useful information, which explains why the corpus has already inspired a great deal of research into the pragmatics of ELF interactions. For instance, it makes available comprehensive metadata for each speech event on aspects such as speakers' power relations or their level of acquaintedness. Also, the mark-up system comprises detailed tags for pragmatic phenomena such as laughter, overlaps, repetition, pauses or pronunciation variations and coinages (see VOICE Project 2007). The part-of-speech (POS) tagged version of VOICE includes annotation of pragmatic categories like discourse markers or formulaic items (see VOICE Project 2014). Besides the design of VOICE, the new technologies that have been introduced in the recently launched 3.0 version of the corpus would be highly compatible with more detailed functional annotation. While several enhanced

functionalities have become available in this update, the main benefit that comes with VOICE 3.0 Online is that it is now possible to search for mark-up. Since a particular focus in the development of VOICE 3.0 Online was on ensuring its user-friendliness, such searches are easily accessible also for a non-expert audience since they can be conducted without knowledge of corpus query language (CQL). Provided that the annotation scheme for pragmatic functions, which is currently developed and intended to complement the existing annotation in VOICE with an additional functional layer, is eventually applied to the data and released, it would thus be possible to search for communicative functions in the corpus without difficulty.

In view of all this, it seemed highly desirable to embark on the establishment of a pedagogically motivated annotation scheme for pragmatic functions in VOICE. Attempts to refer to conventional approaches when annotating VOICE data indicate the inapplicability of established categories (Pitzl 2018: 90–92), and highlight the need to do justice to the kind of data captured in the corpus (Osimk-Teasdale and Dorn 2016: 382). This suggests an empirically informed approach that pays specific attention to the nature of the data to adequately reflect the interactive processes occurring in ELF communication. In addition to this empirical orientation, initial observations suggest several other points of reference that prove useful in the development of the annotation system. First of all, there are annotation schemes that, in principle, adopt a communicative orientation to language data. These are, however, primarily associated with application in corpus linguistic research rather than language pedagogy. The index available for the *Michigan Corpus of Spoken Academic English* (MICASE) (The University of Michigan English Language Institute 2002) represents a special case in this respect. It includes a selection of pragmatic features for a subcorpus of MICASE for language pedagogic purposes (cf. Simpson-Vlach and Leicher 2006: 68–71). Simpson-Vlach and Leicher (2006: 67) explain that the aim is to render communicative phenomena evolving in MICASE more transparent and to ease searches for samples of functional language use for teachers. These samples are, however, based on data of L1 usage and therefore less applicable to ELF interactions.

Other existing annotation systems equally relate to ENL/L1 corpora and remain mainly associated with use in corpus linguistic research. These schemes have been generally subsumed under the umbrella of speech act annotation. In this context, Searle's (1976) categories are still readily picked up as main mark-up units (see e.g. Kallen and Kirk 2012) or inform taxonomies developed to annotate language data (see e.g. Garcia 2004). Based as it is on invented examples and grammatically well-formed, isolated sentences, Searle's taxonomy, however, contributes little to the description of language functions in data of spoken language-in-use. Adolphs (2008: 45–47), for instance, convincingly exemplifies how difficult it is to empirically apply one of Searle's categories to attested spoken language data to conclude that conventional classification frameworks, like the one provided by Searle, fail to account for much of the discursive work speakers are involved in during communication.

Currently ongoing research applying existing classification systems to VOICE data (Riegler 2021a, 2021b) seems to support these findings. It displays how using established speech act annotation schemes to annotate ELF interactions only scratches the surface of what is actually happening interactively and points towards the need of a specialized categorization of pragmatic functions for ELF data. While it is early days yet, further points of reference are already being considered as useful supplements to available annotation systems. First, findings on the pragmatic practices adopted in ELF interactions inform the currently devised VOICE taxonomy. Existing research already locates pragmatic practices in actual ELF data, points out strategies of functional value that ELF speakers adopt as they communicate (for a useful overview cf. e.g. Sato et al. 2019: 13–14), and in doing so, already gives an indication of pedagogically significant aspects in ELF use. Studies into ELF pragmatics generally recognize the close relation between the more pedagogically oriented concept of communication strategy and the notion of pragmatic practice, which is why classification systems of communication strategies (see e.g. Dörnyei and Scott 1997) will be reflected in the VOICE annotation system for pragmatic functions. Second, conversation analysis (CA) is deemed compatible with pragmatic corpus annotation that is rooted in speech act theory (cf. e.g. Archer and Culpeper 2018: 501–502). This explains

why CA is drawn upon in current pragmatic, speech act-oriented annotations (see e.g. Garcia 2004) and will also inform the development of the annotation scheme for pragmatic functions in VOICE.

The design of the additional layer in the VOICE mark-up is an important step in rendering corpus data, more specifically ELF data, accessible for use in language pedagogy. Not only is it the pedagogic significance immanent in ELF communication, as addressed above, that is attended to in the development of this annotation scheme. The aim will be to conceptualize corpus annotation in such a way that it provides the means for pedagogic authentication. A pedagogic example, as defined by Widdowson (2009: 207), "is always an example of something, the token of a type […]. So a sample of textual data from a corpus has in some way to be interpreted as evidence of something typical for it to serve as an example". A pragmatically oriented annotation system for VOICE should highlight instances of pragmatic practices in ELF data, giving an indication of how ELF speakers' utterances can be functionally interpreted and, from a pragmatic perspective, what they can be taken as exemplifying. The samples of communicative processes, which the VOICE annotation would provide, could then be used in ELT, for example, to actually teach interactive strategies and to get language students actively involved in the same processes of negotiating understanding as observed in ELF interactions —to take up suggestions offered by Seidlhofer (2011: 189–205). That is, the samples may be applied in English classrooms to encourage "learning how to 'language', how to exploit the potential in the language for making meaning" (Seidlhofer 2011: 189). The development of the annotation scheme thus works towards strengthening the ties between English language pedagogy and ELF corpora. At the same time, the design of the annotation system for pragmatic functions in VOICE seeks to complement the aforementioned proposals for introducing concepts and findings from ELF research into ELT and facilitate the implementation of pedagogies for English as a lingua franca.

11.5 Conclusion

This contribution has sought to outline the need for a pedagogically oriented annotation scheme for data of English as a lingua franca communication. While giving an overview of recent research exploring the interface between ELF corpora and language pedagogy, and recognizing the pedagogic relevance of findings from descriptions of English as a lingua franca communication, it is argued that more mediation work is urgently required to enhance the relation between ELF corpora and language classrooms. The paper suggests corpus annotation as a way to render aspects of pedagogic significance in ELF interactions more readily available for use in English language education. From the analysis of established mark-up schemes, it emerges that pedagogically oriented corpus annotation so far has adopted a normative approach and focuses heavily on language forms. Corpus tagging thus needs to be reconceived to serve the purpose of pointing out communicative practices to be encouraged in language students rather than continuing the tradition of highlighting deviations from an L1 norm as errors in need of correction.

The paper then makes the case for a functional orientation to pragmatic practices in the annotation of ELF interactions, as would be desirable from an ELF perspective on language pedagogy. It gives insight into, and outlines the theoretical and conceptual points of reference informing the ongoing development of this annotation system for VOICE data. The design of the annotation scheme works towards strengthening the ties between ELF corpora and ELT by rendering aspects of pedagogic significance in ELF communication more readily available for use in language classrooms, thereby supporting the pedagogic implementation of an ELF perspective. Once available, the further layer of pragmatic functions in VOICE will render communicative practices in ELF interactions more accessible, for instance, for materials development, task design and use in communicative language pedagogies. The development of this pedagogically oriented annotation system for pragmatic practices in VOICE represents a necessary step in responding to the challenge of integrating findings from ELF research into pedagogic practice.

Appendix

NICT JLE tags in Extract 1 (from NICT 2012b)

<A>…</A> Interviewer's utterance

<B>…</B> Learner's utterance

<H pn="X">…</H> Learner's personal information

<F>…</F> Filler/Filled pause

<SC>…</SC> Self-correction

<at>…</at> Article

<o_lxc>…</o_lxc> Collocation

<v_tns>…</v_tns> Verb tense

<n_num>…</n_num> Number of noun

<v_lxc>…</v_lxc> Lexis (verb)

<… crr="X"> Corrected form

<… odr="X"> Number of tag applied within utterance

LINDSEI tags in Extract 2 (from Dagneaux et al. 2005)

<A></A> Examiner's utterance

<B></B> Learner's utterance

$…$ Corrected form/target hypothesis

(L) Lexis

(S) Single

(P) Preposition

(G) Grammar

(V) Verb

(M) Morphology

(F) False friend

References

ACE. 2014. *The Asian Corpus of English*. Director: Andy Kirkpatrick. Researchers: Wang Lixun, John Patkin and Sophiann Subhan. http://corpus.ied.edu.hk/ace/. Accessed 1 February 2018.

Adolphs, Svenja. 2008. *Corpus and Context: Investigating Pragmatic Functions in Spoken Discourse*. Amsterdam: Benjamins.
[Crossref]

Archer, Dawn, and Jonathan Culpeper. 2018. Corpus Annotation. In *Methods in Pragmatics*, ed. Andreas Jucker, Klaus Schneider, and Wolfram Bublitz. Berlin/Boston: Mouton de Gruyter.

Aston, Guy. 2011. Applied Corpus Linguistics and the Learning Experience. In *Perspectives on Corpus Linguistics*, ed. Vander Viana, Sonia Zyngier, and Geoff Barnbrook. Amsterdam: Benjamins.

Breiteneder, Angelika, Theresa Klimpfinger, Stefan Majewski, and Marie-Luise Pitzl. 2009. The Vienna-Oxford International Corpus of English (VOICE): A Linguistic Resource for Exploring English as a Lingua Franca. *ÖGAI Journal* 28 (1).

Cogo, Alessia. 2018. ELF and Multilingualism. In *The Routledge Handbook of English as a Lingua Franca*, ed. Jennifer Jenkins, Will Baker, and Martin Dewey. London: Routledge.

Cogo, Alessia, and Martin Dewey. 2012. *Analysing English as a Lingua Franca: A Corpus-driven Investigation*. London: Continuum.

Dagneaux, Estelle, Sharon Denness, and Sylviane Granger. 1998. Computer-aided Error Analysis. *System* 26.

Dagneaux, Estelle, Sylviane Granger, Fanny Meunier, Jennifer Thewissen, Sharon Denness, and JoAnne Neff. 2005. *Error Tagging Manual Version 1.2*. Louvain-la-Neuve: Université catholique de Louvain.

Dewey, Martin. 2012. Towards a Post-normative Approach: Learning the Pedagogy of ELF. *Journal of English as a Lingua Franca* 1 (1).

Dorn, Nora. 2011. The '-ing thing': Exploring the Progressive in ELF. *Vienna English Working Papers* 20 (2).

Dörnyei, Zoltán, and Mary Scott. 1997. Communication Strategies in a Second Language: Definitions and Taxonomies. *Language Learning* 47 (1).

Dose-Heidelmayer, Stefanie, and Sandra Götz. 2016. The Progressive in Spoken Learner Language: A Corpus-based Analysis of Use and Misuse. *International Review of Applied Linguistics in Language Teaching* 54 (3).

ELFA. 2008. *The Corpus of English as a Lingua Franca in Academic Settings*. www.helsinki.fi/elfa/elfacorpus. Accessed 1 February 2018.

Garcia, Paula. 2004. *Meaning in Academic Contexts: A Corpus-based Study of Pragmatic Utterances*. PhD thesis, Northern Arizona University, Flagstaff.

Gilquin, Gaëtanelle, Sylvie de Cock, and Sylviane Granger. 2010. *Louvain International Database of Spoken English Interlanguage*, Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, Sylviane. 2015. The Contribution of Learner Corpora to Reference and Instructional Materials Design. In *The Cambridge Handbook of Learner Corpus Research*, ed. Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier. Cambridge: Cambridge University Press.
[Crossref]

Kallen, Jeffrey, and John Kirk. 2012. *SPICE-Ireland: A User's Guide. Documentation to Accompany the SPICE-Ireland Corpus: Systems of Pragmatic Annotation in ICE-Ireland*. Belfast: Cló Ollscoil na Banríona.

Kaur, Jagdish. 2016. Using Pragmatic Strategies for Effective ELF Communication: Relevance to Classroom Practice. In *Exploring ELF in Japanese Academic and Business Contexts: Conceptualisation, Research and Pedagogic Implications*, ed. Kumiko Murata. London/New York: Routledge.

Leech, Geoffrey. 1997. Introducing Corpus Annotation. In *Corpus Annotation: Linguistic Information from Computer Text Corpora*, ed. Roger Garside, Geoffrey Leech, and Tony McEnery. London/New York: Routledge.

Lozano, Cristóbal, and Ana Díaz-Negrillo. 2019. Using Learner Corpus Methods in L2 Acquisition Research. *Spanish Journal of Applied Linguistics* 32 (1).

Nicholls, Diane. 2003. The Cambridge Learner Corpus: Error Coding and Analysis for Lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 Conference*, ed. Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery. Lancaster: Lancaster University.

NICT. 2012a. *The NICT Japanese Learner English Corpus* (version 4.1). https://alaginrc.nict.go.jp/nict_jle/index_E.html. Accessed 9 February 2021.

———. 2012b. *The NICT JLE Corpus Tag List*. https://alaginrc.nict.go.jp/nict_jle/src/taglist.pdf. Accessed 9 February 2021.

Osimk-Teasdale, Ruth, and Nora Dorn. 2016. Accounting for ELF: Categorising the Unconventional in POS-tagging the VOICE Corpus. *International Journal of Corpus Linguistics* 21 (3).

Pitzl, Marie-Luise. 2017. Communicative 'Success', Creativity and the Need for De-mystifying L1 Use: Some Thoughts on ELF and ELT. *Lingue e Linguaggi* 24.

———. 2018. *Creativity in English as a Lingua Franca: Idiom and Metaphor*. Boston/Berlin: Mouton de Gruyter.
[Crossref]

Ranta, Elina. 2013. *Universals in a Universal Language? Exploring Verb-syntactic Features in English as a Lingua Franca*. PhD thesis, University of Tampere, Tampere.

Rastelli, Stefano. 2009. Learner Corpora Without Error Tagging. *Linguistik Online* 38 (2).

Riegler, Stefanie. 2021a. 'I think this pragmatic things can be solved somehow': Annotating Pragmatic Functions in a Spoken ELF Corpus. Paper presented at *Outils et Nouvelles Explorations de la Linguistique Appliquée (ONELA)*, University of Toulouse, Hybrid/Toulouse, 19–21 October.

———. 2021b. 'It's a surface definition, it's not about the quality aspect': Reviewing Speech Act Classifications for Annotating Pragmatic Functions in ELF Communication. Paper presented at *17th International Pragmatics Conference (IPRA)*, Zurich University of Applied Sciences, Online/Winterthur, 27 June–2 July.

Römer, Ute. 2010. Using General and Specialized Corpora in English Language Teaching: Past, Present and Future. In *Corpus-based Approaches to English Language Teaching*, ed. Mari Carmen Campoy-Cubillo, Begoña Bellés-Fortuño, and Maria-Lluïsa Gea-Valor. London/New York: Continuum.

Sato, Takanori, Yuri Yujobo, Tricia Okada, and Ethel Ogane. 2019. Communication Strategies Employed by Low-Proficiency Users: Possibilities for ELF-Informed Pedagogy. *Journal of English as a Lingua Franca* 8 (1).

Searle, John. 1976. A Classification of Illocutionary Acts. *Language in Society* 5 (1).

Seidlhofer, Barbara. 2001. Closing a Conceptual Gap: The Case for a Description of English as a Lingua Franca. *International Journal of Applied Linguistics* 11 (2).

———. 2011. *Understanding English as a Lingua Franca*. Oxford: Oxford University Press.

Sifakis, Nicos. 2019. ELF Awareness in English Language Teaching: Principles and Processes. *Applied Linguistics* 40 (2).

Simpson-Vlach, Rita, and Sheryl Leicher. 2006. *The MICASE Handbook: A Resource for Users of the Michigan Corpus of Academic Spoken English*. Ann Arbor: The University of Michigan Press. [Crossref]

The University of Michigan English Language Institute. 2002. *Michigan Corpus of Spoken Academic English (MICASE)*. https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase. Accessed 18 September 2020.

Vettorel, Paola. 2019. Communication Strategies and Co-construction of Meaning in ELF: Drawing on 'Multilingual Resource Pools'. *Journal of English as a Lingua Franca* 8 (2).

VOICE. 2021. *The Vienna-Oxford International Corpus of English* (version 3.0 online). https://voice3.acdh.oeaw.ac.at. Accessed 7 October 2021.

VOICE Project. 2007. *Mark-up Conventions*. VOICE Transcription Conventions [2.1]. https://voice.acdh.oeaw.ac.at/wp-content/uploads/2021/04/VOICE-mark-up-conventions.pdf. Accessed 7 October 2021.

———. 2014. *Part-of-Speech Tagging and Lemmatization Manual*. https://voice.acdh.oeaw.ac.at/wp-content/uploads/2021/04/POS-tagging-and-lemmatization-manual.pdf. Accessed 7 October 2021.

Widdowson, Henry. 2009. The Linguistic Perspective. In *Handbook of Foreign Language Communication and Learning*, ed. Karlfried Knapp and Barbara Seidlhofer. Berlin: Mouton de Gruyter.

———. 2012. ELF and the Inconvenience of Established Concepts. *Journal of English as a Lingua Franca* 1 (1).

Footnotes

1 While Widdowson (2009: 214) originally referred to foreign language communication, the arguments seem to equally apply to ELF communication 2009).

2 Corpus created by the National Institute of Information and Communications Technology (NICT).

3 See appendix for the tags relevant in Extract 1 taken from *The NICT JLE corpus tag list* (NICT 2012b).

4 The extract relies on the error tagging applied to LINDSEI data by Dose-Heidelmayer and Götz (2016). For the purposes of this paper, additional (untagged) context was provided from LINDSEI.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_12

12. Detecting and Analysing Learner Difficulties Using a Learner Corpus Without Error Tagging

Gerold Schneider[1] ✉
(1) Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

✉ **Gerold Schneider**
Email: gschneid@ifi.uzh.ch

Keywords Data-driven approaches – Data-driven learning – Collocations – Learner corpus – Indirect corpus use – Surprisal

12.1 Introduction

This study aims to show how typical learner errors and areas of difficulty can be detected by using learner corpora in a data-driven fashion, that is by using an approach which brings overused patterns to the surface—very many of them contain errors. I demonstrate how we detect specific learner errors in a corpus of learner English and make suggestions as to how this information can be used, so that material producers expose the students to the correct forms, steering them clear from direct translations in difficult constructions. This suggestion is first and foremost directed at ELT material developers. However, we also advocate for the use of our insights by the inquisitive teacher in the sense that he or she could make a direct connection between research and practice by developing classroom materials or indeed by adapting concordance lines from corpora which showcase the naturally

occurring L1 uses of the items we uncover here, or by extracting authentic learner data from learner corpora and gradually moving toward direct data learning where the students themselves look at both types of corpora, mainly corpora of native English containing the difficult expressions, and possibly even corpora of learner English (for further discussion on the integration of data-driven learning in the classroom, see Timmis and Templeton, this volume). I focus on the most frequent types of errors in learner language according to Ng et al. (2014), namely determiner errors and prepositional errors. The latter also includes many collocation errors, which are not only frequent, but also hard to learn, address and correct. According to Gilquin and Granger (2011: 59–60), verb-preposition constructions are particularly difficult to acquire for language learners. In addition, phrasal verbs, which we also include in this class, represent "one of the most notoriously challenging aspects of English language instruction" (Gardner and Davies 2007: 339). Similarly, English determiners have been identified as one of the most difficult parts of English grammar for language learners, and one of the latest features to be fully acquired (Master 1990: 461).

In this study, I detect typical errors by comparing a learner corpus, the International Corpus of Learner English (ICLE, Granger et al. 2009) to a native speaker reference, the British National Corpus (BNC, Aston and Burnard 1998). Many approaches use error-annotated corpora, but as large error-annotated corpora are rare (Han et al. 2010) or not accessible to all researchers (Nicholls 2003), their scope is limited. I thus suggest the use of a large, unannotated learner corpus as a complementary approach. Detecting errors in unannotated corpora is more challenging, but the ICLE corpus is large enough to facilitate the detection of hundreds of typical errors.

To do so, I present an approach of indirect corpus use, e.g. the use of corpus materials that provides teachers and students with learning materials, in particular lists of typical idiomatic errors and difficult verb + preposition constructions. Indirect corpus use employs methods which allow researchers and teachers to compile teaching material and to avoid typical pitfalls that arise in the transfer from the speakers' L1 or when using semantic analogies. This is particularly useful for teachers of those students who find direct corpus use too challenging.

Chujo et al. (2016) confirm that the use of data-driven learning (Boulton 2011) is helpful, but they point out, concerning the use of corpus tools in classrooms, that many studies show them to be mainly beneficial for intermediate and advanced learners (Chujo et al. 2016: 262), be it because students do not know where to start, find query languages puzzling or do not fully understand the complex discourse which the concordance displays.

In these cases, indirect corpus use can provide additional help, as Chujo et al. (2016) also point out. Luo (2016) summarizes the mutual benefits and the complementary nature of direct and indirect data-driven learning (DDL) as follows:

> [B]oth direct and indirect DDL have their own distinctive advantages and disadvantages. To compare the effectiveness of the two ways, Yoon and Jo (2014) conducted a small-scale study investigating their different effects on L2 learners' error correction in a writing class. This study revealed that the self-correction rate was higher in indirect DDL than in the direct DDL for most learners, especially for lower-level learners; however, direct DDL activities may have more positive effects on learner autonomy especially for higher-level learners. Thus it could be found that indirect DDL is more appropriate for lower-level or novice learners. (Luo 2016: 2)

This study's approach contributes to the compilation of more complete inventories of problem areas and teaching materials for students, tailored to their native language background. Our guiding question will be: how can learner corpora be used to detect learner errors and areas of difficulty, to the benefit of language learners and teachers?

In the following we will first sketch relevant background, then we will introduce our methodology and data before we present our results and draw conclusions.

12.2 Background

Language learners tend to overuse familiar constructions and words at the cost of rarer and semantically more fine-grained realizations, because they do not know or cannot recall the appropriate word or construction. This tendency is known as the 'Teddy Bear' effect (Granger 2001; Ellis 2012). It often leads to errors, and also often marks the difference between understandable and native-like English (Pawley and Syder 1983). In a similar vein, Cowie (1994: 3168) argues that "native-like proficiency of a language depends crucially on knowledge of a stock of prefabricated units".

12.2.1 The Role of Corpus Linguistics

Corpus linguistics offers the possibility to language learners of accessing vast amounts of real-world utterances and remedying their lack of exposure to the target language. In practice, though, language learners are often overcharged when given the possibility to query a corpus and may not know how to approach the task (e.g. Chujo et al. 2016). To overcome this problem, an approach of indirect corpus use is helpful. From the vast collections of typical learner and non-learner language presented in corpora, we can cull typical learner errors by using statistical methods to detect these with data-driven approaches.

12.2.2 Learner Difficulties

While learners typically have a limited vocabulary, they spend a large amount of their learning time acquiring new words and are aware of this. When it comes to multiword expressions and idiomatic language use, learners are often less aware and make less effort to improve their skills. The acquisition of multiword expressions is reported to consistently lag behind the learning of single words (Forsberg 2010; Laufer and Waldman 2011; Li and Schmitt 2010; Peters 2014). Routinization is particularly difficult for learners, as e.g. McEnery and Xiao (2010) point out. Learning collocations (Sect. 12.3.1) and expected continuations (Sect. 12.3.2) are thus particularly important, and often underrepresented parts of language learning. As L2 learners are used to focusing on individual words (Wray 2002), El-Dakhs et al. (2022) report that learners are unlikely to notice new expressions and that it is very challenging for language teachers to teach multi-word expressions in a principled manner (Alali and Schmitt 2012). In order to acquire

routinization and idioms, a vast amount of exposure is needed, which, according to Pawley and Syder (1983), is difficult for learners to obtain. Even if multi-word expressions are shown to learners in texts, they do not necessarily notice them, as El-Dakhs et al. (2022: 6) report from a story retelling task.

The current study has revealed that EFL learners, particularly with lower or medium levels of proficiency, are likely to incorporate useful multiword expressions from an input text in their subsequent writing only marginally. This finding indicates that attempting to increase EFL learners' mining of multiword expressions through pairing reading with listening and/or typographically enhancing target expressions may not suffice for effective text mining. To achieve an adequate level of noticing (Schmidt 1990, 2010) and for vocabulary uptake to take place, focused instruction and sustained practice of multiword expressions seem necessary.

By using data-driven methods to detect the difficulties arising from the lack of exposure, we can precisely satisfy this need, specifically detecting difficult multi-word expressions in a systematic and transparent way, based on the evidence that these errors are frequent, and teaching their correct forms.

12.3 Methods and Data

It is no coincidence that the term *data-driven* appears both in the teaching method on which we rely here, data-driven learning (DDL) as a teaching method, and in the term data-driven pattern detection used as an error detection method. Both are essentially synonymous in a cognitive linguistic interpretation. The specific methods that we employ to uncover the typical error patterns are introduced in this section, namely collocations (Sect. 12.3.1) and expected sequences (Sect. 12.3.2)—cognitive statistical data that can be extracted from learner corpora and L1 reference corpora (Sect. 12.3.3).

12.3.1 Collocations

Collocations involve conventionalized use of linguistic expressions. A collocation is defined, according to Choueka (1988: 609), as "a sequence of two or more consecutive words, that has characteristics of

a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components". Criteria for determining collocations include non-compositionality, non-substitutability, limited modifiability, non-literal translations and statistical co-occurrence. While only the criterion of statistical co-occurrence can be measured trivially in corpora, this criterion has proven to be a generally appropriate measure, because it can relate measured collocation strength to the psycholinguistic entrenchment which is behind it. Gries and Wulff (2005, 2009) find strong correlations between collocation strengths and experimentally obtained sentence completions from advanced learners of English as a Foreign Language (EFL). This indicates that in measuring collocation strength we can form a model of language users' expectations. This suggestion is supported by Ellis and Ferreira-Junior (2009), who find that frequency of learner uptake is predicted by frequency of occurrence, and indeed by collocation measures.

A wide array of frequency-based collocation statistics has been suggested (see Evert 2009 for an introduction). We will use simple measures such as O/E and T-score here. O/E stands for Observed divided by Expected. The observed frequency of two words occurring together is divided by the naive expectation that words are randomly distributed, which would be the case if no collocational force existed. O is the co-occurrence of two words form a corpus, and E is the count that one would expect if words (say *x* and *y*) were independent events, i.e. if word order were random.[1] The additional measure of T-score directly uses the value of the T-test, a popular significance test. O/E delivers the same ranking of collocations as Mutual Information (MI), which is included in several concordance tools. After filtering out false positives, i.e. non-examples, and possibly irrelevant material, lists of strong collocations detected by such means can be used as teaching material. Such an example of verb-preposition constructions, reported by the collocation measure O/E, is given in Table 12.1. It is reproduced from Lehmann and Schneider (2011).

Table 12.1 Verb-Object-PP collocations, sorted by decreasing O/E

Verb	Object	Prep	Desc noun	t-score	O/E
Send	shiver	down	spine	5.74456	2.21E+08
Tap	esc	for	escape	6.40312	2.11E+08
Separate	shield	from	plate	6.78233	2.33E+07
Refer	gentleman	to	reply	8.24621	7.81E+06
Obtain	property	by	deception	5.2915	7.60E+06
Ask	secretary	for	affairs	6.40312	5.02E+06
Kill	Bird	with	stone	5.38516	3.38E+06
Add	Insult	to	injury	6.08276	2.22E+06
Throw	Caution	to	wind	5.09902	2.03E+06
Refer	Friend	to	reply	7.54983	1.36E+06
Report	Loss	on	turnover	7.14142	1.35E+06
Ask	secretary	for	science	5.56776	837702
Thank	friend	for	reply	5.56776	809822
Run	finger	through	hair	6.48073	734125
Ask	secretary	for	health	5.19615	715744
Keep	head	above	water	5.56776	714483

The entries are sorted by decreasing O/E. Table 12.1 only shows the very top of the list. The first several hundred entries contain many idioms. Examples from such a collection of collocations measured by O/E, such as *send (a) shiver down (the / one's) spine*, *separate (the) shield from (the) plate*, *kill (two) birds with (one) stone* provide an excellent repository of idioms that can be taught. False positives appear in this list, too, for example *ask secretary for science* which stems from a parser error, as the automatic parser (Schneider 2008) which Lehmann and Schneider (2011) used, wrongly considers *for science* to be attached to *ask*. There are also many examples which in fact are not idiomatic but fully compositional, such as *tap esc for escape*.

However, many collocations and idiomatic expressions are identical or similar in the learners' L1 and in English, or do not lead to errors because they are frequently taught. For example, the expression *on the one hand … on the other hand* is mostly used correctly by language learners in the ICLE corpus, and even tends to be overused as it is well known among many learners. The comparison of collocations in learner corpora with native speaker corpora offers the opportunity to detect

which errors typically occur, and which idiomatic expressions are used incorrectly. This knowledge is then used by material developers (e.g. to create lists of difficult idioms) and can be further developed by teachers to produce targeted teaching materials for their students, for example cloze tests.

The automatic detection of errors is now an established field of research in computational linguistics (for example, De Felice and Pulman (2008) and Ng et al. (2014)). Automatic detection works best if large amounts of manually annotated errors are available for training machine learning approaches, so-called 'supervised learning'. But annotated resources are not always available, and the main interest of the study is to discover errors that occur frequently and repeatedly, so that they can be targeted in teaching. The goal is not to detect each individual error automatically, but to detect most of the frequent and typical errors, from which teaching material can be created.

A number of typical errors can be extracted from learner corpora because they are so frequent that they reach collocational status in these corpora. But the vast majority of collocations in learner corpora are correct uses—so the method of simply applying collocation detection is not sufficient. In order to find learner errors, we need to find word combinations which 1) are conventionalized, i.e. frequent enough to reach collocation status, 2) are collocations in L2, but 3) are not collocations, or much less so, in L1. If we apply traditional collocation measures we fail to see point 3). A successful measure for 3) is the collocation ratio (Schneider and Zipp 2013), which simply divides the two collocation values.[2]

We find that the collocation ratio allowed us to efficiently detect hundreds of verb-preposition errors, as we describe in Sect. 12.3.2.

12.3.2 Surprisal

A further measure that goes back to Shannon's (1951) Information Theory is surprisal, which expresses how unexpected the use of a word is in a certain context. Surprisal is generally defined (Levy and Jaeger 2007) as the probability of a word to appear in the given context.[3] The probabilities with which a word commonly occurs in British English have been learnt from the British National Corpus (BNC, Aston and Burnard 1998). The BNC has been chosen as the source for extracting

the bigram surprisal values, because it is a large corpus that represents a larger selection of genres than, for instance, web-derived corpora, and this leads to a more balanced language model (see e.g. Lapata and Keller 2005).

Surprisal allows one to measure chunking (Altenberg 1998), i.e. the tendency that several words may form multi-word sequences which are retrieved as single units from our vocabulary. It also allows us to measure the competition between the idiom and syntax principle (Sinclair 1991), i.e. the competition between chunking and the creative use of syntax. It further models the cline from lexis to grammar (Langacker 1990; Boers 2014), showing for instance that lexicon entries and syntax are in complex interaction. A linguistic style that is dominated by the idiom principle has low surprisal, many chunks, and is easy to process, but it contains little information. A style dominated by the syntax principle is dense but often hard to process. Surprisal also correlates strongly with processing times. While high surprisal may be a sign of highly compressed language, for example in scientific writing, in learner language it very often is a sign of error. High surprisal indicates that an unusual sequence of word-forms has been used. For this reason, surprisal also forms the base of most error detection systems, usually coupled with other approaches (Ng et al. 2014). In particular, the automatic correction of errors is often much more difficult.

Finally, a word on the difference between collocation measures and surprisal is appropriate at this stage. While collocation measures and surprisal sometimes deliver similar results, there are also differences. A main difference is that surprisal is directional: while we can make statements about the probability of one word following another word, this does not entail that this relation also holds the other way round. For example, not all collocations have high surprisal: even though the noun *decision* is often preceded by the verb *take*, very many different objects may follow the verb *take*. Thus, the probability for the word *take* to be followed by the word *decision* is quite low, and as a result, surprisal is low as well even though any collocation measure will predict a high collocational status. Although there are clear differences between surprisal and collocation measures, it is far from obvious

which methods yield good results for which error type—here open-minded experimentation and evaluation are needed.

The differences in frequency that are detected between collocations in learner corpora and non-learner corpora such as the BNC are then compared systematically and finally the correct forms are used for teaching. In Sect. 12.4.1, we will show how surprisal can be used to detect determiner errors.

12.3.3 Data

The learner corpus used for this study is the International Corpus of Learner English (Granger et al. 2009), henceforth ICLE. The corpus contains 3.7 million words of essays by university students who range from higher intermediate to advanced learners of English. It covers a broad range of L1 backgrounds: Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish, and Tswana, each contributing at least 200.000 words.

As the L1 reference corpus, we use the written part of the British National Corpus (Aston and Burnard 1998), henceforth BNC. It contains 90 million words of written texts from a wide range of registers and we use it as a standard of comparison for formal varieties of native British English. Additionally, for one comparison involving article frequency, we use the International Corpus of English (Nelson et al. 2002) corpora, henceforth ICE. Each ICE component contains 1 million words of spoken and written text and contains the same genres.

12.4 Results

We now present the typical errors that our methods uncover, and assess their performance, in terms of how many of the suggested candidates for learner errors are indeed errors.

12.4.1 Determiner Errors

For the detection of the first error type, typical determiner errors in ICLE, we compare if found word sequences in ICLE are more surprising or less surprising with or without determiners in the BNC. This is done

by using surprisal (Levy and Jaeger 2007), the measure of surprise which we have introduced in Sect. 12.2.2.

In order to detect determiner omission errors, we check for each sequence of two words (a so-called bigram) in the ICLE corpus, whether, according to BNC, the surprisal for the bigram (*W1 W2*) is high (we define as "high" any measure above 17) and higher than the one for the trigram with a determiner (*W1 the/a W2*). With this method, 90% of the instances found in the learner corpus are true positives, i.e. real errors. For example, our method detects *have place* instead of *have a place*, *make decision* instead of *make a decision*, and *if person* instead of *if a person*. Table 12.2 gives an excerpt of the data and shows the top 70 candidates that seem to show determiner omission. We have sorted the list by frequency and secondarily by the O/E value. A high O/E value gives prominence to candidates that have collocational status and are thus especially worthy of being taught as fixed expressions or idioms and thus be added to teaching materials. Overall, over 200 candidates for collocational status have been determined. However, the quality further down in the list is lower, there are more false positives.

Table 12.2 Top 70 missing determiners in ICLE

F (ICLE)	W1	W2	O/E (W1,W2) with determiner	Surprisal w/o det	Surprisal with det	F (BNC with determiner)
16	is	same	2.0551	17.1530	16.8010	1077
10	have	same	3.8788	17.8462	14.7963	947
8	make	decision	22.4474	26.0000	21.1376	245
8	takes	long	14.0807	17.4407	17.0997	77
8	have	place	0.4712	26.0000	23.6109	87
7	just	few	19.0451	26.0000	15.2620	816
5	reading	newspaper	99.3530	26.0000	20.7651	35
5	are	same	2.3542	26.0000	15.6676	577
5	if	woman	1.2728	26.0000	20.6851	43
4	make	mistake	37.8728	26.0000	23.1630	93
4	find	solution	32.5326	26.0000	22.5816	77
4	watch	news	28.0419	26.0000	21.2711	21
4	making	decision	21.9975	26.0000	21.0268	83

F (ICLE)	W1	W2	O/E (W1,W2) with determiner	Surprisal w/o det	Surprisal with det	F (BNC with determiner)
4	make	effort	12.5807	26.0000	24.0338	63
4	after	few	11.0974	26.0000	16.6898	391
4	than	hundred	10.3224	26.0000	20.0382	242
4	have	job	1.8707	26.0000	22.8696	163
4	almost	only	1.8501	26.0000	15.8041	66
4	is	result	0.6150	26.0000	23.0505	116
3	smoking	cigarette	785.4514	26.0000	18.9606	36
3	solve	problem	72.5684	17.1530	16.3903	34
3	under	influence	57.1413	26.0000	22.1856	293
3	quite	few	50.9216	17.8462	15.1464	719
3	within	few	33.4934	26.0000	15.2742	533
3	on	contrary	16.0007	26.0000	25.9991	262
3	has	advantage	15.1054	26.0000	21.2041	243
3	pass	time	7.6929	17.1530	15.3248	70
3	have	advantage	5.9646	26.0000	23.0034	171
3	where	woman	2.5960	26.0000	20.5518	45
3	before	law	2.2133	26.0000	23.0219	34
3	with	lot	2.0901	26.0000	20.2268	324
3	is	difference	1.6627	26.0000	25.5289	159
3	have	game	0.8883	26.0000	25.7739	50
3	is	importance	0.5263	26.0000	25.5181	43
3	not	few	0.3216	26.0000	20.1496	52
3	have	house	0.2461	26.0000	25.5486	33
2	spend	lot	93.5468	26.0000	14.6779	163
2	understand	meaning	38.1033	26.0000	21.7650	39
2	play	role	36.4927	26.0000	20.7569	118
2	discussed	matter	33.2150	26.0000	18.3335	44
2	nearly	hundred	26.9564	26.0000	18.1484	45
2	tell	lie	24.4708	26.0000	21.7546	28
2	getting	job	21.6198	26.0000	19.2170	78

F (ICLE)	W1	W2	O/E (W1,W2) with determiner	Surprisal w/o det	Surprisal with det	F (BNC with determiner)
2	given	chance	19.8261	26.0000	20.3786	86
2	spend	Great	16.6030	26.0000	17.4136	42
2	with	Aim	10.3295	26.0000	25.0923	367
2	cause	Problem	8.2156	26.0000	19.1847	26
2	understand	Problems	8.0447	26.0000	21.0789	28
2	find	Place	7.0885	26.0000	19.9212	116
2	about	importance	6.7617	26.0000	23.4552	107
2	use	Correct	6.1066	26.0000	23.9235	21
2	is	Ultimate	5.9071	26.0000	25.4211	117
2	get	Fair	5.8102	26.0000	23.7186	35
2	within	Period	5.0809	26.0000	21.7255	46
2	make	Same	3.4691	26.0000	16.6743	140
2	get	Same	2.8717	26.0000	16.6985	143
2	if	Person	2.6992	26.0000	19.6991	106
2	which	Artist	2.6117	26.0000	25.9420	31
2	when	Person	2.3874	26.0000	20.4207	88
2	but	Main	2.1837	26.0000	20.4651	148
2	is	Ability	2.1501	26.0000	23.8869	167

Errors are frequently found in light verb constructions, e.g. in the use of the light verb construction *make decision* instead of *make a/the decision* (Fig. 12.1). As indicated by the first letters in the reference code of the corpus examples, the L1 background of the speakers in the ICLE Corpus who make this error is mainly Chinese (CN). However we also find data by, Czech (CZ), Japanese (JP) and Turkish (TR) speakers, all of which are languages without or with a different article system.

Your Query:'make decision\b' returned 9 results in ICLE_T6571_CLEAR.

|< << >> >| Show Page: 1 KWIC View Show Tags New Query Go!

No	Reference	Solutions 1 to 9 Page 1/1 Processed for gerold at 62.202.182.206
1	CNHK1170:0170.1:5	Besides the economic factors to **make decision** , the health of everyone is the most important things that need to think of.
2	CNHK1664:0664.1:5	According to, most women in middle class women believe that they have rights to **make decision** .
3	CNHK1676:0676.1:1	Anti-abortion people argue this is immoral to use abortion while supporters suggest it is women's right to **make decision** to their bodies.
4	CNHK1740:0740.1:2	They have right to live and **make decision** of their life.
5	CNUK4013:0013.4:1	So, inthe court, judger ought to know the situation, **make decision** what to do isright or not.
6	CZPR2036:0036.2:1	When the patient can not **make decision** himself.
7	JPTF1006:0006.1:3	No one else but they know what the best way is, and can **make decision** themselves on their own lives.
8	TRCU1159:0159.1:2	First of all, I believe that no one can **make decision** when to die.
9	TRKE2036:0036.2:3	Males can make their decision and they think that it is unnecessary to ask or consult to their wives incase they **make decision** .

Dependency Bank 2.0 © 2010-2017 Hans Martin Lehmann & Gerold Schneider

Fig. 12.1 Hits for *make decision*

A further frequent source or errors are generic references, as Fig. 12.2 illustrates on the basis of the use of the collocation *if (a) woman*.

Your Query:'if woman' returned 5 results in ICLE_T6571_CLEAR.

|< << >> >| Show Page: 1 KWIC View Show Tags New Query Go!

No	Reference	Solutions 1 to 5 Page 1/1 Processed for gerold at 62.202.182.206
1	CNHK1748:0748.1:1	Professionals-choice is another group which allowed abortion in most situations **if woman** decides.
2	ITB08003:0003.8:16	Man must demostrate all his qualities, even **if woman** does n`t want to be confined at home and she is becomin stronger than man, trying to show him she can do everything.
3	JPSW3023:0023.3:1	Even **if woman** gets married, we still go out the world, and she should work.
4	JPSW3023:0023.3:2	That time, **if woman** is n`t working.
5	JPSW3023:0023.3:2	In the reverse, **if woman** is working.

Dependency Bank 2.0 © 2010-2017 Hans Martin Lehmann & Gerold Schneider

Fig. 12.2 Hits for *if woman*

Not all candidates which are detected by this approach are true positives: for example, searching for the collocation *just few*, we might find both true positives and false positives. From inspection of the hits we can see though that typically *just a few* was most likely intended (Fig. 12.3).

Your Query:'just few\b' returned 7 results in ICLE_T6571_CLEAR.

|<< << >> >>| Show Page: 1 KWIC View Show Tags New Query Go!

No	Reference	Solutions 1 to 7 Page 1/1 Processed for gerold at 62.202.182.206
1	CZPR2017:0017.2:1	And there have always been just few .
2	CZPR2017:0017.2:1	And there can be just few .
3	CZUN1016:0016.1:2	During the evening they say just few words to each other.
4	ITB07002:0002.7:1	Now you can see just few ruins of its walls, walking around Via Roma and Via Castello, and even though this is the oldest point of the village, fifty years ago, it was n`t the centre of it.
5	ITB08002:0002.8:5	considering as thinking human being very few years, during which just a little part of them has decided to exercise their rifuts in every sense (they can vote but they can also be voted under candidature, but just few of them exploits or has exploited this opportunity), though the cares of it we not only due to psychological matters related to men's world.
6	ITRS2028:0028.2:7	Also the media treats the above conflicts differently from the Gulf War, as it shows us just few images of them, while during the latter we had a war-live.
7	POPZ5047:0047.5:5	Each university student should be required to fulfil just few basic courses compulsory to a particular department, and given a chance to choose a number of additional courses which he or she is interested in.

Dependency Bank 2.0 © 2010-2017 Hans Martin Lehmann & Gerold Schneider

Fig. 12.3 Hits of *just few* in the ICLE corpus

Figures 12.2 and 12.3 indicate that also speakers from languages with determiner systems may struggle with determiner choice, for instance Italian (IT) with three errors in Fig. 12.3 and one in Fig. 12.2.

We have just argued that *just few* may not constitute an error. Equally, *understand problems* could be a correct usage. Again, on inspection of the hits (see Fig. 12.4) we can see that they are errors.

Your Query:'understand problems\b' returned 2 results in ICLE_T6571_CLEAR.

|<< << >> >>| Show Page: 1 KWIC View Show Tags New Query Go!

No	Reference	Solutions 1 to 2 Page 1/1 Processed for gerold at 62.202.182.206
1	CZPU1006:0006.1:3	This small baby can not understand problems of our modern world.
2	JPSW3018:0018.3:5	For example, if you are a black person, you need black people in government, because it is easier than white governments or other racial governments that black governments understand problems of black people.

Dependency Bank 2.0 © 2010-2017 Hans Martin Lehmann & Gerold Schneider

Fig. 12.4 Hits for *understand problems* in the ICLE corpus

Also the opposite error, using a determiner in a place where native speakers most likely would not, is frequent among learners of English. Ng et al. (2014) find that determiner errors are the second most frequent error type, although they do not distinguish between

determiner insertion or omission. For each trigram in the ICLE corpus, if the middle word is a determiner (*W1 the/a W2*), I have checked if, according to BNC, the surprisal for the bigram without determiner (*W1 W2*) is smaller, and if surprisal for the trigram with the determiner is high (again using the threshold of 17). The list is sorted by O/E to give prominence to candidates that have collocational status and that are thus especially worthy of being taught as fixed expressions or idioms. The top 60 candidates are given in Table 12.3. Again, the list contains far more entries and further true positives, but the performance decreases further down in the list, it contains more false positives.

Table 12.3 Extra determiners in ICLE, top 60 sorted by decreasing collocation score (O/E)

F (ICLE)	W1	DET	W2	O/E without determiner	Surprisal with det	Surprisal w/o det	F (BNC w/o determiner)
1	committed	a	suicide	2887.851	18	13.1733	214
1	awful	a	lot	894.685	18	12.0455	661
1	higher	the	education	791.765	18	10.9299	2017
1	great	a	deal	686.695	18	10.2395	4023
1	both	the	sexes	685.011	18	13.0755	236
1	last	a	night	362.980	18	9.6054	7585
1	takes	a	place	325.948	18	11.2128	1520
4	take	the	advantage	262.959	18	11.5381	1098
1	understand	the	why	215.769	18	11.7292	907
1	taking	the	advantage	208.567	18	12.9446	269
1	makes	the	sense	159.785	18	12.3510	487
1	makes	a	sense	159.785	18	12.3510	487
1	large	a	number	133.112	18	11.0072	1867
1	difficult	the	task	128.137	18	13.2212	204
1	young	the	people	118.077	18	10.4131	3382
4	even	the	worse	110.539	18	12.3823	472
1	all	the	kinds	106.046	18	11.5564	1078
1	became	a	known	85.481	18	12.5159	413
1	make	a	sense	66.293	18	11.6955	938
1	get	the	married	62.432	18	12.3490	488

F (ICLE)	W1	DET	W2	O/E without determiner	Surprisal with det	Surprisal w/o det	F (BNC w/o determiner)
1	long	the	time	58.669	18	10.1981	4193
1	under	the	attack	56.999	18	12.9786	260
1	becoming	the	more	50.541	18	12.0952	629
1	know	the	why	48.656	18	11.1411	1633
1	no	the	matter	48.612	18	11.2552	1457
1	social	a	life	45.174	18	11.9587	721
1	give	a	evidence	43.612	18	12.7402	330
1	is	the	unclear	40.929	18	12.7402	330
1	be	a	grateful	38.506	18	12.1902	572
1	without	the	saying	38.278	18	13.0020	254
1	open	the	market	37.709	18	13.0755	236
1	one	the	thing	37.033	18	10.5797	2863
1	whole	the	world	36.768	18	12.4967	421
1	so	the	funny	35.922	18	13.0840	234
1	high	the	cost	35.775	18	13.2212	204
2	than	the	half	35.021	18	11.3965	1265
1	this	a	kind	34.392	18	10.7322	2458
1	particular	the	case	32.277	18	12.6560	359
1	this	the	context	31.851	18	11.8373	814
1	to	the	greet	31.573	18	12.6124	375
1	end	the	result	28.463	18	12.9980	255
1	be	a	distinguished	27.928	18	12.5783	388
1	no	a	real	27.548	18	11.8167	831
2	become	the	part	26.402	18	12.6700	354
1	least	a	four	26.214	18	12.8625	292
1	is	the	essentially	26.206	18	11.9313	741
3	been	a	widely	25.743	18	12.7463	328
1	be	the	reduced	24.957	18	11.5084	1131
1	be	a	prepared	24.936	18	11.2531	1460
2	not	a	sure	23.427	18	10.9230	2031

F (ICLE)	W1	DET	W2	O/E without determiner	Surprisal with det	Surprisal w/o det	F (BNC w/o determiner)
1	they	the	are	22.935	18	8.1913	31195
1	are	the	commonly	22.910	18	13.1012	230
1	one	a	person	22.516	18	11.3878	1276
1	to	the	sell	22.442	18	10.2947	3807
1	were	the	lucky	22.424	18	13.1502	219
1	such	the	things	21.935	18	11.9060	760
1	such	a	things	21.935	18	11.9060	760
2	really	a	need	21.761	18	12.3844	471
1	most	the	people	21.426	18	10.9902	1899
1	this	the	approach	21.074	18	11.6176	1014

Here, the rate of incorrect suggestions by the algorithm is a bit higher than it was for determiner omission. For example, *take the advantage* (Fig. 12.5) or *been a widely* (Fig. 12.6) could constitute correct uses. Inspecting the hits shows that *been a widely* is indeed a false positive; the uses are correct and do not contain an error.

Your Query:'take the advantage\b' returned 4 results in ICLE_T6571_CLEAR.

|< << >> >| Show Page: 1 KWIC View Show Tags New Query Go!

No	Reference	Solutions 1 to 4 Page 1/1 Processed for gerold at 62.202.182.206
1	CNHK1082:0082.1:4	Another important point which is in favour of students using credit card is that students can **take the advantage** of using credit cards if they use them properly.
2	CNHK1190:0190.1:6	Benefits of using credit cards are offer much convenience to the students, let them to learn financial management eariler and **take the advantage** of using it.
3	CNHK1575:0575.1:3	According to the students are able to **take the advantage** of the best deals on offer.
4	TSNO1250:0250.1:9	The poor installiation of street lights in some countries may be another factor that can lead to the spread of HIV / AIDS because criminals or rapist **take the advantage** to do their evil acts.

Dependency Bank 2.0 © 2010-2017 Hans Martin Lehmann & Gerold Schneider

Fig. 12.5 Hits of *take the advantage* in the ICLE corpus

Your Query:'been a widely' returned 3 results in ICLE_T6571_CLEAR.

|< << >> >| Show Page: 1 KWIC View Show Tags New Query Go!

No	Reference	Solutions 1 to 3 Page 1/1 Processed for gerold at 62.202.182.206
1	BGSU1173:0173.1:1	Equality has always **been a widely** discussed topic.
2	BGSU1263:0263.1:1	Equality of people has always **been a widely** discussed issue throughout history.
3	FIJO3010:0010.1:1	Agriculture and its importance in society has always **been a widely** discussed topic all over the world.

Dependency Bank 2.0 © 2010-2017 Hans Martin Lehmann & Gerold Schneider

Fig. 12.6 Hits of *been a widely* in the ICLE corpus

If learners use extra determiners, this is often due to hypercorrection by speakers of an L1 without determiners, such as Chinese, Japanese or Finnish. Particularly Chinese L1 speakers tend to do so: if we compare absolute numbers of determiners in ICLE, according to the L1 of the speakers, we can see that while Japanese, Tswana, Czech and Russian speakers have below average frequencies, while the frequencies of Chinese determiners are roughly average in comparison to the other learners (Fig. 12.7).

Distribution of Query '\b(thela)\b' according to the category 'llanguage'

Tabulate: llanguage Go! Crosstabulate: X -- / Y llanguage show Count of Instances

llanguage	n	words	relative frequency per 10000 words
BULGARIAN	13831	210469	657.151
CHINESE	6419	88656	724.034
CHINESE-CANTONESE	30422	422138	720.665
CHINESE-MANDARIN	329	4577	718.811
CZECH	12487	208634	598.512
DUTCH	17702	243750	726.236
FINNISH	14116	200835	702.866
FRENCH	16225	214619	755.991
GERMAN	17319	248666	696.476
ITALIAN	18252	241118	756.974
JAPANESE	11737	210954	556.377
NORWEGIAN	14779	221578	666.989
OTHER	2059	27751	741.955
POLISH	15975	247395	645.728
RUSSIAN	15310	236885	646.305
SPANISH	16334	210562	775.734
SWEDISH	19564	290358	673.789
TSWANA	11218	208082	539.114
TURKISH	13822	206598	669.029
UNKNOWN	308	4283	719.122

Dependency Bank 2.0 © 2010-2017 Hans Martin Lehmann & Gerold Schneider

Fig. 12.7 Relative frequencies of Determiners by L1

We can compare these frequencies to native English varieties, for which we consider L1 and L2 frequencies from the ICE Corpora. Figure 12.7 shows that while Japanese, Tswana, Czech and Russian indeed use fewer determiners than L1 or L2 varieties, speakers of Romance L1s (Italian, French, Spanish) use more determiners than most L1 varieties. We can also perceive that there is a marked register variation—spoken genres in first language varieties of English contain fewer determiners than written genres (Fig. 12.8).

Crosstabulation of Query '\b(the|a)\b' according to categories 'iceCat1' and 'corpus'. Relative Frequency.

Tabulate: corpus Go! Crosstabulate: X iceCat1 / Y corpus show Count of Instances Go!

corpus/iceCat1	spoken	written	total
icecan	590.224	721.438	645.147
icegb	660.293	747.131	696.174
icehk	656.106	690.821	670.551
iceind	648.158	696.201	667.416
iceire	649.38	766.311	697.37
icejam	660.219	720.497	684.565
icenz	668.751	738.386	698.399
icephi	635.052	741.524	678.952
icesing	609.117	680.535	638.603
total	642.734	722.496	675.661

Dependency Bank 2.0 © 2010-2017 Hans Martin Lehmann & Gerold Schneider

Fig. 12.8 Frequency of determiners in spoken and written genres of the ICE corpora

12.4.2 Prepositional Constructions

Concerning the second error type, prepositional constructions, this study investigates verbs with Prepositional Phrase (PP) complements (e.g. *depend on*), adjectives with PP complements (e.g. *responsible for*), and phrasal verbs (e.g. *turn down*)—for simplicity they are all referred to as verb-PP. They largely overlap with a third error category, idiom errors. Typical verb-PP errors can be detected by comparing collocation values between the learner corpus and the native speaker corpus (BNC) because they reach high collocational status in the learner corpus, compared to a much lower collocation value in the native corpus. In Sect. 12.2.1 the use of a collocation ratio which divides the collocation value of the language learner corpus by the one of the L1 reference was therefore suggested.

This approach (cf. Schneider and Gilquin 2018) uncovers hundreds of typical verb- and adjective-PP errors: for example, learners sometimes use *replace to* instead of *replace by*, *accuse for* instead of *accuse of*, or *discuss about* instead of *discuss*. In the following, one result is shown for which the T-score collocation measure has been used. The top 30 candidates for L2 errors are given in Table 12.4, sorted by descending T ratio (first column). Column 5 gives the T-score collocation measure for ICLE, column 6 the collocation measure for the L1 reference, which is the BNC. The hits have been inspected manually. Wherever the last column (COMMENT) contains a correction (*instead*)

we have a true positive of a non-canonical learner collocation. This column contains the form that will then be taught to the students.

Table 12.4 Verb + Preposition overuse in ICLE, sorted by decreasing T ratio

T ratio	Verb	Prep	F	T (ICLE)	T (BNC)	Comment
5.9820	impose	to	10	5336.86	892.15	instead of *impose on*
3.5860	replace	to	3	1168.35	325.81	instead of *replaced by* (partly)
2.1133	accuse	for	8	5143.81	2433.98	instead of *accuse of*
2.0275	addict	on	4	3431.99	1692.68	instead of *addict to*
1.4296	better	than	87	17920.70	12535.47	–
1.3929	alarm	of	2	2691.03	1932.01	instead of *alarm about*
1.3322	handicap	after	30	10530.89	7905.03	CORPUS SELECTION essay topic
1.2812	better	for	59	14564.98	11367.88	–
1.2074	diverse	by	2	2690.71	2228.48	instead of *different according to*
1.1541	discuss	about	43	12421.43	10762.54	instead of *discuss sth*.
0.9322	consist	on	13	6290.72	6748.02	instead of *consist of*
0.9042	basic	for	2	2673.74	2957.02	–
0.8576	aim	on	2	2040.77	2379.77	instead of *aim at*
0.8351	smoke	in	1153	64641.60	77403.98	CORPUS SELECTION essay topic
0.8159	equal	than	172	25189.25	30871.17	partly CORPUS SELECTION
0.8148	helpless	for	4	3789.47	4650.78	–
0.8027	view	upon	3	3319.27	4135.30	instead of *viewed on* (archaic)
0.7813	attack	against	2	2698.64	3454.11	instead of *attack someone*
0.7328	harmful	for	55	14074.48	19207.33	–
0.7261	independent	on	6	4473.42	6160.53	instead of *independent of*
0.7166	route	through	11	6376.93	8898.68	–
0.6817	afraid	about	2	2248.11	3297.94	instead of *afraid of*
0.6645	understand	towards	2	2670.72	4019.42	–
0.6635	master	as	69	15919.97	23992.80	CORPUS SELECTION essay topic
0.6068	concentrate	to	5	2746.33	4526.23	instead of *concentrate on*
0.5894	intolerant	to	3	3289.11	5580.82	–
0.5785	speak	under	2	2533.35	4379.28	–
0.5639	reuse	of	6	4685.40	8309.02	verb instead of noun

T ratio	Verb	Prep	F	T (ICLE)	T (BNC)	Comment
0.5052	live	ago	3	3182.39	6299.41	–
0.4974	interest	about	5	4193.29	8430.47	–
0.4411	relate	with	49	13056.44	29600.00	instead of *relate to*

One example from this list which will now be explored is *discuss about*. Based on our inspection it can be concluded that this candidate, suggested by our algorithm, is a true positive of a non-canonical learner structure, i.e. what one marks as an error in a student essays. Figure 12.9 shows the uses of this prepositional construction in the learner corpus.

Your Query:'h1=discuss r1=pobj r2=prep d2=about eq2=depID=headID ' returned 43 results in ICLE_t6571.

|< << >> >| Show Page: 2 Show chunks Show Tags New Query Go!

No	Reference	Solutions 31 to 43 Page 2/2 Processed for gerold at 178.198.196.26
31	ITTO2029:0029.2:1	In an article that appeared recently in The Financial Times the journalist Joe Rogaly discussed **about the possibility** of making gun ownership illegal in every nation of the world in order to reduce and even to eliminate the opportunities to commit crimes.
32	ITTO2030:0030.2:3	If the person who shoots another is a hero or a psychopath we are not here to discuss **about this**.
33	ITVE1003:0003.1:1	In the last few years conferences and debates have been held by experts and psychologists to discuss **about the delicate issue** of artificial insemination of single women.
34	JPKO1005:0005.1:2	So I think to keep the country peacely the governments should have opportunities to explain and discuss **about the governments policies**.
35	JPKO2019:0019.2:1	I discuss **about it** the following.
36	JPKO2019:0019.2:4	Second I discuss **about** whether there **are** any relations between that we like baseball and our racial history(of our culture).
37	JPSH1001:0001.1:1	Newspapers and TV programs discussed **about the crime** for along time.
38	JPTF1032:0032.1:1	We discussed **about introducing** English education into an elementary school.
39	TRCU1137:0137.1:3	I only want to discuss **about the inequality** between these two gender.
40	TRCU1169:0169.1:1	First of all people are getting married without knowing each other very well also discussing **about small matters** triggers the couples for divorce and the most important factor of why divorce rate is increasing is that people have become less resistant to difficulties.
41	TRCU1169:0169.1:1	Then you start to discuss **about what** to do.
42	TRKE2042:0042.2:1	Especially women and men discuss **about this subject**.
43	TRME3016:0016.3:5	There is no need to explain the affect of ecomomical power in whatever subject we discuss **about education**.

BNC Dependency Bank 1.0 © 2010-2013 Hans Martin Lehmann & Gerold Schneider

Fig. 12.9 Hits for *discuss about* from ICLE

12.4.3 Pedagogical Application

The lists of typical errors that I or other developers have extracted are the base for teaching the correct idioms, forms, word sequences and article uses (see column 7, "COMMENT" in Table 12.4). This needs the small step of the developers of the materials or the teachers to provide

the correct forms—the lists and exercises are then based on them. This will allow students to be exposed to the crucial type of data, which direct DDL would have relatively small chances to offer to students even if they read and hear extensively, or which they may easily overlook.

The exercises can contain L1 real uses, cloze tests in which the students need to enter the correct preposition or determiner. Also noticing tests (Ellis 2002) are possible, but exposing students to errors instead of L1 forms runs the risk that they may remember the incorrect form.

The borderlines of the roles that material developers, teachers and students take in the application are only partly predefined. While it is too difficult and time-consuming for most teachers to calculate collocations ratios, the creation of exercises and lists of constructions can be done equally well by material developers or teachers. Querying for the right forms in L1 corpora can be done by teachers or students, as part of an exercise. While the students should mainly be exposed to the corrected forms, possibly encouraged to query for their L1 uses, advanced students might experimentally even be exposed to errors and asked to correct them to raise awareness of structure, grammar, multi-word units and constructions.

12.5 Discussion

The analysis of the learner errors reveals mechanisms of overextension and analogy at work, particularly in prepositional structures. Learners often apply subcategorization frames of nouns to verbs, like in the example of *discuss about*, based on *discussion about*, or from semantically similar words (*independent on*, which is formed in analogy to *dependent on*). As the Teddy Bear effect (Ellis 2012) predicted, already known grammatical patterns are overused by the learners. For example, the corpus data show that speakers of Romance languages tend to use the preposition *to* as a general dative case marker (*impose to*, *replace to*).[4] Finally, there are instances of L1 influence, and we noticed a general overuse of the unspecific preposition *about*.

Many determiner errors also occur in combination with idioms, particularly with light verbs, for instance *make a/the decision*, *make an*

error, *pave the way*, *commit suicide*, *take advantage*, *make sense*. As these are typical learner errors and are not intuitive, the corrected versions lend themselves particularly well to being taught explicitly. The typical errors that can be detected in this way can directly be used to create targeted teaching material for students, helping them to avoid typical pitfalls. We have also seen that while article errors are particularly frequent with speakers of L1s both without or with a different determiner system, these errors are also produced by speakers of any language background represented in ICLE.

12.6 Conclusion

This study has set out to address the guiding question of "how can learner corpora be used to detect learner errors and areas of difficulty to the benefit of language learners and teachers?" To this end, results have been presented of indirect corpus use (not in the classroom, but for material developers and software developers), which automatically detects typical errors and areas of difficulty that learners of English struggle with. In doing so, hundreds of errors of three very frequent types could be detected: determiner errors, prepositional structure errors, and fixed expressions like collocations and idioms, the latter overlapping with the first two types. These can be used to create targeted teaching material to students of a given L1.

We have further shown how an indirect approach to corpus linguistics can extract typical errors using learner corpora, enabling material developers to provide resources including exercises and cloze tests also to students who might initially be overwhelmed by direct corpus use. Once their awareness of competing constructions is developed far enough, this approach can also serve as an ice-breaker to the students' subsequent successful use of direct data-driven learning.

The use of data-driven approaches has few theoretical limitations. Comparing the language models obtained from learner corpora allows automated or semi-automated comparisons at any level. Over- and underuse of any linguistic feature can be detected and addressed, and native reference material showcasing L1 use be delivered to the learner. There are still serious practical limitations, though. The sizes of corpora needed for data-driven approaches is very large, and the calculations

too difficult for non-experts. With increasingly sophisticated linguistic methods these approaches are currently progressing fast, and will form part of language learning applications in a future that is not even so distant.

References

Alali, Fatima A., and Norbert Schmitt. 2012. Teaching Formulaic Sequences: The Same as or Different From Teaching Single Words? *TESOL Journal* 3 (2): 153–180.
[Crossref]

Altenberg, Bengt. 1998. On the Phraseology of Spoken English: The Evidence of Recurrent Word Combinations. In *Phraseology: Theory, Analysis, and Applications*, ed. Anthony P. Cowie, 101–122. Oxford: Oxford University Press.

Aston, Guy, and Lou Burnard. 1998. *The BNC Handbook. Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Boers, Frank. 2014. Idioms and Phraseology. In *The Bloomsbury Companion to Cognitive Linguistics*, ed. Jeannette Littlemore and John R. Taylor, 185–201. London: Bloomsbury.

Boulton, Alex. 2011. Data-driven Learning: The Perpetual Enigma. In *Explorations Across Languages and Corpora*, ed. Stanislaw Goźdź Roszkowski, 563–580. Frankfurt: Peter Lang.

Choueka, Yaacov. 1988. Looking for Needles in a Haystack. *Proceedings of RIAO '88*: 609–623.

Chujo, Kiyomi, Yuichiro Kobayashi, Atsushi Mizumoto, and Kathryn Oghigian. 2016. Exploring the Effectiveness of Combined Web-based Corpus Tools for Beginner EFL DDL. *Linguistics and Literature Studies* 4 (4): 262–274.
[Crossref]

Cowie, Anthony P. 1994. Phraseology. In *The Encyclopaedia of Language and Linguistics, 6*, ed. Ron Asher, 3168–3171. Oxford: Pergamon Press.

De Felice, Rachele, and Stephen G. Pulman. 2008. A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 169–176.
[Crossref]

El-Dakhs, Dina Abdel, Shazia Khalid Khan Salam, and Maram Al-Khodair. 2022. Do Foreign Language Learners Mine Input Texts for Multiword Expressions? The Case of Writing Story Retellings. *Ampersand* 9. https://doi.org/10.1016/j.amper.2021.100080. Accessed 6 March 2022.

Ellis, Rod. 2002. The Place of Grammar Instruction in the Second/Foreign Language Curriculum. In *New Perspectives on Grammar Teaching in Second Language Classrooms*, ed. Sandra Fotos and Eli Hinkel, 17–34. Mahwah, NJ: Lawrence Erlbaum.

Ellis, Nick C. 2012. Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual Review of Applied Linguistics* 32: 17–44.
[Crossref]

Ellis, Nick C., and Fernando Ferreira-Junior. 2009. Constructions and Their Acquisition: Islands and the Distinctiveness of Their Occupancy. *Annual Review of Cognitive Linguistics* 7: 187–220.
[Crossref]

Evert, Stefan. 2009. Corpora and Collocations. In *Corpus Linguistics. An International Handbook*, ed. Anke Lüdeling and Merja Kytö, 1212–1248. Berlin: Mouton de Gruyter.

Forsberg, Fanny. 2010. Using Conventional Sequences in L2 French. *International Review of Applied Linguistics in Language Teaching* 48 (1): 25–51.
[Crossref]

Gardner, Dee, and Mark Davies. 2007. Pointing out Frequent Phrasal Verbs: A Corpus-Based Analysis. *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect* 41 (2): 339–359.
[Crossref]

Gilquin, Gaëtanelle, and Sylviane Granger. 2011. From EFL to ESL: Evidence from the International Corpus of Learner English. In *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*, ed. Joybrato Mukherjee and Marianne Hundt, 55–78. Amsterdam: John Benjamins.
[Crossref]

Granger, Sylviane. 2001. Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In *Phraseology: Theory, Analysis, and Applications*, ed. Anthony P. Cowie, 145–160. Oxford: Oxford University Press.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English v2 (Handbook + CD-Rom)*. Louvain-la-Neuve: Presses universitaires de Louvain.

Gries, Stefan Th., and Stefanie Wulff. 2005. Do Foreign Language Learners Also Have Constructions? Evidence from Priming, Sorting, and Corpora. *Annual Review of Cognitive Linguistics* 3: 182–200.
[Crossref]

———. 2009. Psycholinguistic and Corpus Linguistic Evidence for L2 Constructions. *Annual Review of Cognitive Linguistics* 7: 163–186.
[Crossref]

Han, Na-Rae, Joel Tetreault, Soo-Hwa Lee, and Jin-Young Ha. 2010. *Using an Error-Annotated Learner Corpus to Develop an ESL/EFL Error Correction System*. Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10).

Langacker, Ronald W. 1990. *Foundations of Cognitive Grammar, Volume 2: Descriptive Applications*. Stanford, CA: Stanford University Press.

Lapata, Mirella, and Frank Keller. 2005. Web-based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing* 2 (1): 1–31.
[Crossref]

Laufer, Batia, and Tina Waldman. 2011. Verb-noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Language Learning* 61 (1): 647–672.
[Crossref]

Lehmann, Hans Martin, and Gerold Schneider. 2011. A Large-scale Investigation of Verb-attached Prepositional Phrases. In *Studies in Variation, Contacts and Change in English (VARIENG)*, ed. Sebastian Hoffmann, Paul Rayson, and Geoffrey Leech, Vol. 6. Methodological and Historical Dimensions of Corpus Linguistics.

Levy, Roger, and T. Florian Jaeger. 2007. Speakers Optimize Information Density Through Syntactic Reduction. *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, Vancouver, Canada.

Li, Jie, and Norbert Schmitt. 2010. The Development of Collocation Use in Academic Texts by Advanced L2 Learners: A Multiple Case Study Approach. In *Perspectives on Formulaic Language: Acquisition and Command*, ed. David Wood, 22–46. London: Continuum.

Luo, Qinqin. 2016. The Effects of Data-driven Learning Activities on EFL Learners' Writing Development. *Springerplus* 5 (1): 1255. https://doi.org/10.1186/s40064-016-2935-5.
[Crossref]

Master, Peter. 1990. Teaching the English Articles as a Binary System. *TESOL Quarterly* 24 (3): 461–478. https://doi.org/10.2307/3587230.
[Crossref]

McEnery, Tony, and Richard Xiao. 2010. What Corpora Can Offer in Language Teaching and Learning. In *Handbook of Research in Second Language Teaching and Learning 2*, ed. Eli Hinkel, 364–380. London & New York: Routledge.

Nelson, Gerald, Sean Wallis, and Bas Aarts 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Varieties of English Around the World: G29, Amsterdam: Benjamins.

Ng, Tou Hwee, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Hendy Raymond Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, Baltimore, Maryland, 1–14.

Nicholls, Diane. 2003. The Cambridge Learner Corpus: Error Coding and Analysis for Lexicography and ELT. *Proceedings of the Corpus Linguistics 2003 Conference*, 572–581.

Pawley, Andrew, and Frances Hodgetts Syder. 1983. Two Puzzles for Linguistic Theory: Native-like Selection and Native-like Fluency. *Language and Communication*, 191–226.

Peters, Elke. 2014. The Effects of Repetition and Time of Post-test Administration on EFL Learners' form Recall of Single Words and Collocations. *Language Teaching Research* 18 (11): 75–94.

[Crossref]

Schmidt, Richard. 1990. The Role of Consciousness in Second Language Learning. *Applied Linguistics, 11*, 129–158.

Schmidt, Richard. 2010. Attention, awareness, and individual differences in language learning. In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan, & I. Walker, *Proceedings of CLaSIC 2010, Singapore, December 2–4*, 721–737. Singapore: National University of Singapore, Centre for Language Studies.

Schneider, Gerold. 2008. *Hybrid Long-distance Functional Dependency Parsing*. Doctoral Thesis, University of Zurich, Faculty of Arts. https://doi.org/10.5167/uzh-7188.

Schneider, Gerold, and Gaëtanelle Gilquin. 2018. Detecting Innovations in a Parsed Corpus of Learner English. In *Linguistic Innovations: Rethinking Linguistic Creativity in Non-native Englishes*, ed. Sandra C. Deshors, Sandra Götz, and Samantha Laporte, 47–74. Amsterdam: Benjamins.

Schneider, Gerold, and Lena Zipp. 2013. Discovering New Verb-preposition Combinations in New Englishes. *Studies in Variation, Contacts and Change in English (VARIENG)*, Vol. 13. http://www.helsinki.fi/varieng/series/volumes/13/schneider_zipp.pdf.

Shannon, Claude E. 1951. Prediction and Entropy of Printed English. *The Bell System Technical Journal* 30: 50–64.
[Crossref]

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.
[Crossref]

Yoon, Hyunsook, and Jung Won Jo. 2014. Direct and Indirect Access to Corpora: An Exploratory Case Study Comparing Students' Error Correction and Learning Strategy Use in L2 Writing. *Language Learning & Technology* 18 (1): 96–117.

Footnotes

1 O/E can be calculated as follows.

$$p(x) = \frac{f(x)}{N}$$

$$O = p\left(x, y\right) \cong p(x)p(y) = E$$

The independent probability of generating *x* is its frequency in the corpus divided by corpus size; and for *y* analogously. The probability of *x* and *y* in combination, in other words the observed value (O), is the frequency of *x* and *y* in combination (e.g. the first word in the bigram is *x*, the second *y*) divided by the corpus size.

$$p(x) = \frac{f(x)}{N}; p(y) = \frac{f(y)}{N}; p\left(x, y\right) = O = \frac{f\left(x, y\right)}{N}$$

If co-occurrence of *x* and *y* is due to chance, i.e. if there is no collocational force, then the independent probability of seeing both Expected (E) and Observed (O), the joint probability of seeing the combination, are roughly equal:

$$O = p(x, y) \cong p(x)p(y) = E$$

O/E, Observed divided by Expected, is then:

$$\frac{O}{E} = \frac{p(x, y)}{p(x)p(y)} = \frac{\frac{f(x,y)}{N}}{\frac{f(x)}{N}\frac{f(y)}{N}} = \frac{f(x, y)\,NN}{f(x)f(y)N} = \frac{f(x, y)\,N}{f(x)f(y)}$$

2 if $c_{\mathrm{L1}}(a,b)$ is a collocation measure *c* for L1 of words w_1 and w_2, then

$$\text{Collocation ratio} = c_{\mathrm{L2}}(w_1, w_2) / c_{\mathrm{L1}}(w_1, w_2)$$

The collocation ratio is a measure of overuse, of "overcollocability". The O/E ratio is itself an O/E measure, in which O = O/E(L2); and E = O/E(L1). O/E is an information theoretic measure of surprise (Shannon 1951)—the interpretation of the O/E ratio is equally straightforward; it is also a measure of surprise.

3 In Bayesian statistics this is p(word|context). It is usually expressed as a logarithm to give an information-theoretic value, the surprise in bits in the sense of Shannon (1951) for seeing a new word in the given context. While the detailed definitions can vary, we are using a simple operationalization: the probability of a word linearly combined with the probability of transition to the next word, or $p(w_k|w_{k-1})$.

The definition is thus:

$$bigram\ surprisal = \log \frac{1}{p(w_k)} + \log \frac{1}{p(w_k|w_{k-1})}$$

4 Errors of this type, particularly *participate to* are frequently discussed in teaching aids for English learners with Italian or French as native language, for instance: http://macmillandictionaries.com/MED-Magazine/February2006/35-Phrasal-Verbs-Learners.htm; https://studentlanguages.com/common-english-mistakes-italians-make-19-4-18/.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_13

13. The Potential Impact of EFL Textbook Language on Learner English: A Triangulated Corpus Study

Elen Le Foll[1] ✉
(1) Osnabrück University, Osnabrück, Germany

✉ **Elen Le Foll**
Email: elefoll@uni-osnabrueck.de

Keywords Causative constructions – English as a Foreign Language – Materials design – Input – Textbook corpus – Learner corpus – Construction learning

13.1 Introduction

In instructed foreign language learning contexts and, in particular, at lower secondary school level, a considerable proportion of learners' language input comes from their textbooks (Volkmann 2010: 235). Given their relevance, it will come as no surprise that there is now a growing body of corpus-linguistic studies analysing what will be referred to here as "Textbook English". Most studies to date have compared Textbook English to that of naturally-occurring or "authentic" English as a Native Language (ENL) and many have reported major discrepancies in the use of individual lexico-grammatical features such as future constructions (e.g., Mindt 1987), the progressive (e.g., Römer 2005), and the phraseological patterns of high-frequency verbs (e.g., Le Foll 2022a). Far fewer studies, however, have explored potential correlations between Textbook English and the language produced by learners of English as a Foreign Language (EFL) as the recipients of these textbooks. Exceptions include triangulated investigations of *if*-conditionals (e.g., Gabrielatos 2013; Möller 2020; Winter and Le Foll 2022), which have, on the one hand, shown that the representations of *if*-conditionals in EFL textbooks are far removed from naturalistic ENL use and, on the other, that EFL learners are influenced by the types of *if*-sentences featured in such pedagogical materials, even at higher levels of English proficiency.

Since they entail the comparison of three corpora, such triangulated, Textbook English vs. ENL vs. Learner English, studies are rarely conducted. The present study shows how the results of existing EFL learner vs. ENL corpus research may be drawn upon to explore whether and, if so, how textbook language input may influence learners' syntactic and lexical choices. As an example of this influence, the current study investigates the potential impact of Textbook English on EFL learners' use of causative constructions. Causative constructions were chosen as an example of a construction that straddles textbooks' traditional binary distinctions between vocabulary and grammar and because they have been the focus of a number of learner corpus studies that have pointed to persistent issues in learners' use of this construction (e.g., Wong 1983; Altenberg and Granger 2001; Liu and Shaw 2001; Gilquin 2012, 2016).

The following section introduces the theoretical background and summarises insights from learner English corpus studies on causatives. Section 13.3 presents the textbook corpus data queried as part of this investigation, the procedure adopted to process the data, and the method employed for the phraseological analysis. Section 13.4 begins with the results of the qualitative analysis of the descriptions of causative constructions featured in the textbook grammar sections and continues with the corpus-based analysis of the causatives implicitly featured within the texts of the same textbooks. The representations of causative constructions in textbooks are then compared to the results of previous corpus-based learner English studies with the aim of teasing out the potential impact of Textbook English on EFL learners' language production, before showing how corpora can be used to supplement textbook materials.

13.2 Theoretical Background

13.2.1 Causatives in the Construction Grammar Framework

The present study focuses on EFL textbooks' representations of periphrastic causative constructions with non-finite verb complements such as:

(1) Has this lesson *made you consider* changing the way you use your mobile phone in public? <TEC: Achievers B2>[1]

These constructions consist of a causative verb and a non-finite verb complement clause. The causative verb (e.g., in example (1), *make*) is governed by the causer (*this lesson*) and controls the complement clause which defines both the causee (*you*) and the effect exerted by the causer (*consider changing the way you use your mobile phone in public*). The most frequent causative verbs are MAKE (1), CAUSE (2), GET (3) and HAVE (4):

(2) What situations *cause people to make* faces like that? <TEC: *Green Line New* 5>
(3) Josh *got me to try* bungee-jumping. <TEC: *Piece of Cake* 3e>
(4) She *had her old phone repaired*, even though she expected to get a new one. <TEC: *Solutions intermediate*>

Causative constructions in English have been extensively described within a number of theoretical frameworks (for an overview, see Gilquin 2016: 115). The present study adopts a Construction Grammar (e.g., Goldberg 1995, 2006) approach. Within this framework, constructions are, at the most basic level, understood as "conventionalized pairings of form and function" (Goldberg 2006: 3) and are seen as the basic units of language. Traditional English grammars have usually presented the different types of English causative constructions as more or less synonymous (see Gilquin 2010: 98–99 for overview). However, within a Cognitive Linguistics (e.g., Langacker 2008) and, more specifically, Construction Grammar framework, the principle of "one form, one meaning" prevails. In other words, "[i]f two constructions are syntactically distinct, they must be semantically or pragmatically distinct" also (Goldberg 1995: 67; see also Croft's 2001: 111–112, "Principle of Contrast"). By combining the analysis of corpus and elicitation data, Gilquin (2010: Ch. 5) demonstrates that this principle of non-synonymy applies to periphrastic causative constructions. Though many syntactic and semantic features are distributed in significantly different ways across different causative constructions, the factor that best predicts the choice of one causative construction over another is a lexical one: namely the verb used to describe the EFFECT in the non-finite complement clause (Gilquin 2010: 138–139), e.g., DO in (5), taken from the British National Corpus.

(5) yeah yeah it is a long operation […]
of course he won't be able to work for weeks and weeks
no this is it that's why [he] hasn't *had it done* he *could have had it done*
I thought he did
he *had it done* once but he's got more joints now <BNC1994: SQS3>

\r\n

Construction Grammar assumes that speakers' language knowledge consists of "a network of learned pairings of form and function, or constructions" (Goldberg 2005: 17). These range from morphemes to individual words, to collocations and verb-argument structures. Whilst the idea that language can be broken down to constructions that span right across the traditional grammar-lexicon divide is no longer new, it remains controversial in many a linguistic circle (Langacker 2008: 22). In foreign language pedagogy, however, approaches advocating remarkably compatible ideas go back a long way (e.g., Hausmann 1984). Yet such approaches have yet to be reflected in most of the EFL textbooks currently in use in secondary schools, which still tend to apply a strict (but often largely arbitrary) distinction between grammar units on the one hand and vocabulary on the other.

13.2.2 Causatives in English Language Teaching and Learner English

Periphrastic causative constructions are both very frequent and of high communicative value; however, they form complex syntactic patterns which often prove problematic for learners of English. Gilquin (2012) shows with data from the International Corpus of Learner English (ICLE; Granger et al. 2009) that even proficient learners of English frequently confuse the different verb complements of causative verbs (e.g., using a *to*-infinitive instead of the bare infinitive with MAKE as in (6)) or struggle with the placement of the causee (e.g., (7)).

(6) It is good practice to use English every day. That *makes us not to forget* English grammar or words. <ICLE: Japanese L1>

(7) Compensation orders *could make feel parents* more responsible for their sons and daughters. <ICLE: Italian L1>

In addition to syntactic issues, EFL learners tend to significantly over- or underuse certain types of causative constructions as compared to ENL speakers (Wong 1983; Altenberg and Granger 2001; Liu and Shaw 2001; Gilquin 2012, 2016). These same studies have also shown that learners frequently make infelicitous or non-idiomatic lexical choices when producing causative constructions. For example, Gilquin (2016) compared the lexical choices of native English (ENL), English as a Second Language (ESL) and English as a Foreign Language (EFL) speakers for the non-finite verb slot of causative constructions with MAKE. The ESL vs. EFL learner comparison is particularly interesting in the context of the present study because, whilst most EFL users will have, at least at first, learnt English as a school subject with few additional language inputs other than their textbooks, ESL users have usually completed large parts of their education in English and/or live in countries where English is an official language (e.g., India, Nigeria, Singapore) and are therefore exposed to much more non-textbook materials than most EFL learners.

Gilquin (2012) identifies three possible causes for learners' unidiomatic use of certain periphrastic causative constructions: (a) lack of register awareness, (b) negative L1 transfer, and (c) the inadequacy of teaching materials. The influence of this third potential factor (also identified by Altenberg & Granger (2001: 184) as a potential source of learners' misuse of MAKE causatives) was further explored in Gilquin (2016), whose results suggest that different types of language input may indeed be partially responsible for the differences observed between EFL and ESL learners' use of causatives. Gilquin's (2016) results support her input-dependent L2 acquisition hypothesis which states that:

> considering that (i) EFL learners learn English mainly in an instructional setting and receive relatively little exposure to the target language, and (ii) ESL learners, in addition, also acquire English in a natural setting and thus receive more exposure to the target language, it can be expected that ESL learners will have been exposed to more (authentic) instances of a given construction than EFL learners and will thus have better integrated its schema, resulting in a better, more native-like command of the construction. (Gilquin 2016: 119)

\r\n

This hypothesis, which raises the question of the (in)adequacy of teaching materials in EFL instructional settings, is at the heart of the present study. It draws on a large corpus of EFL textbooks to explore the distributional patterns of the syntactic, semantic, and phraseological characteristics of the causative constructions to which learners are frequently exposed to via their textbooks. Moreover, it explores whether textbooks provide learners with adequate input that will help them produce these constructions idiomatically in context.

13.3 Data and Methodology

13.3.1 Corpus Data

The Textbook English Corpus (TEC; cf. Le Foll 2021a, 2022a, 2022b) aims to capture the language input that lower secondary school pupils in France, Germany and Spain obtain from their English textbooks. The corpus consists of nine series of best-selling EFL textbooks from eight publishers. The textbooks have been manually annotated to enable fine-grained comparisons between learner levels, learner L1 and text registers (see Appendix 1). Seven register subcorpora of the TEC were included in the present analysis: Informative, Instructional, Narrative, Other, Personal, Poetry and Spoken (see Appendix 2). Together, these subcorpora comprise 1.8 million words (hereafter referred to collectively as the TEC-T).

13.3.2 Data Extraction and Processing

This section describes how the causative constructions of the TEC-T were extracted for the second part of the analysis. In the context of a pedagogically-oriented study, it made sense to focus on the most frequent causative verbs: MAKE, CAUSE and GET.[2] The first step was thus to locate and analyse all the causative constructions of the TEC-T with these three verbs. To this end, the corpus was uploaded onto the web-based corpus analysis tool Sketch Engine (Kilgarriff et al. 2014). Sketch Engine provides a layer of automatic linguistic analysis which makes it possible to query the data on the basis of part-of-speech information (e.g., whether the word *make* is a noun, a verb in infinitive form, or present tense) and lemmas (e.g., *make*, *makes*, *making* and *made* are all forms of the lemma MAKE). Due to the high frequency of the verbs MAKE and GET,

the extraction of these constructions was divided into two phases: search queries were used to automatically extract all possible instances of the constructions and these were then manually sorted to narrow the results down to actual occurrences of causative constructions. Thus, for the first phase, it was necessary to find queries that would be precise enough to considerably reduce the second time-consuming phrase but, at the same time, would have the highest possible recall rate so as not to miss any relevant constructions. After some experimentation, the following queries were chosen for the automatic extraction phase:

(8) [lemma="make" & tag="V.*"] [tag!="SENT|,"]{0,5} [tag="V.*"]
(9) [lemma="get"] [tag!="IN|V.N|V.G" & word!="into|to"] [tag="N.*|PP.?"]{1,5} [tag="V.*"]{1,7}
(10) [lemma="cause" & tag="V.*"]

\r\n

The queries are formulated in Corpus Query Language (CQL). The first, (8), returns all occurrences of the verb MAKE followed by another verb form (operationalised by [tag="V.*"]) with up to five intervening tokens in between to allow for longer nominal phrases as direct objects. A token is a word-like unit which, however, can also include punctuation marks which is why query (8) specifies that the five intervening tokens may be of any kind except end-of-sentence punctuation or commas. The second query for GET causatives, (9), is more complex because it attempts to exclude a number of other frequent GET constructions such as passives (e.g., *He got fined*) and constructions of intransitive motion (e.g., *How did she get into this?*). As CAUSE is a relatively infrequent verb, its CQL query, (10), is much simpler: it simply retrieves occurrences of CAUSE as a verb and, since they are only 105 in the TEC-T, these were then manually disambiguated.

All causative constructions were subsequently manually annotated for their intermediate-level construction following Gilquin's (2010: 222) active-passive classification (see Table 13.1 for subtypes of MAKE constructions), as well as for the lemma of the verb occurring in the non-finite verb slot (e.g., APPEAR, HEAR, KEEP and KNOW in the examples of Table 13.1).

Table 13.1 Intermediate-level causative constructions with MAKE

Intermediate-level construction	Example
[MAKE$_{act}$ V$_{act}$]	This is often done deliberately to *make* a product *appear* more popular. <TEC: Achievers B1+>
[MAKE$_{act}$ V$_{pas}$]	How can you *make* yourself *heard*? <TEC: Piece of Cake 4^{e}>
[MAKE$_{pas}$ V$_{act}$]	It was *made to keep* the devil away. <TEC: Join the Team 5^{e}>
[MAKE$_{pas}$ V$_{pas}$]	There is nothing secret that will not be *made known*. <BNC1994: A18>

13.3.3 Collostructional Analysis

Collostructional analysis explores the interactions between grammatical constructions and the content words (lexemes) associated with them (cf. Stefanowitsch and Gries 2003). As such, the method also follows a construction-based approach to language since it assumes that grammar, just like the lexicon, consists of pairings of linguistic forms with linguistic meaning. Collostructional analysis takes a construction as a starting point and measures which lexemes are attracted to or repelled by the construction. To do so, it evaluates whether lexemes occur more or less frequently than their expected frequency on the basis of the total lexeme and construction counts in the corpus under study. The lexemes associated with a particular construction are referred to as its collexemes (Stefanowitsch and Gries 2003: 215).

The analysis was carried out using the Coll.analysis 3.2a package for R (Gries 2007). As recommended by Stefanowitsch and Gries (2003: 218), the association measure adopted here is Fisher's exact test because it does not rely on or make any distributional assumptions (which are frequently not met when working with corpus data), nor does it impose specific sample size demands. Nonetheless, like all phraseological measures, collostructional analysis requires relatively large sample sizes to yield significant results. The causative constructions data extracted from the TEC-T only provided enough data to perform a collexeme analysis on the most frequent causative construction: [X MAKE Y V$_{inf}$].

13.4 Causatives in EFL Textbooks

13.4.1 Causatives in Textbook Grammars

The term *causative* is mentioned in five out of the nine textbook series of the TEC (see Appendix 1). The term is found in the following volumes: *Piece of Cake 4e*, *New Bridges 2e*, *New Missions 2e*, *English in Mind 3* and *Solutions B2 Intermediate Plus*. The first three represent one volume from each of the three French textbook series of the TEC. This is because, in France, causative constructions are expected to be explicitly taught as part of the lower secondary school EFL curriculum.

New Bridges 2e (see extract (11) from p. 196; emphasis original) focuses on potential syntactic difficulties when translating the French construction *faire faire* and, at the end of the explanation, points to a semantic difference between MAKE and HAVE, arguing that MAKE implies more of a constraint than HAVE.

(11) **« faire faire »**
⦿ **MAKE ou HAVE + base verbale**
The police made the terrorist talk. […]
Juliet had us talk all night. […]
→ **MAKE** exprime plus la contrainte que **HAVE**.
⦿ **HAVE + participle passé**
We've had the lock changed. […]
→ On ne nous dit pas qui a changé la serrure mais seulement que la serrure a été changée. Le participe passé a un sens passive. MAKE est impossible dans ce cas.
⦿ **Traduction de FAIRE FAIRE**
Il faut se demander si l'énoncé a un **sens passif** ou **actif**.
■ **Sens passif : HAVE + participe passé** […]
» *She's had some food delivered*.
■ **Sens actif (quelqu'un accomplit une action à la demande de quelqu'un d'autre): MAKE ou HAVE + base verbale** […]
» *She made / had them go out of the room*.

In its grammar point on causative constructions, *New Missions 2e* also takes the translation of *faire faire* as its starting point. It, too, focuses on the differences in the syntactic patterns of the passive and active forms. In addition, it draws a semantic continuum across MAKE, HAVE and LET from the most constraint to the least (where "*authorisation*" is considered to be the least constrained—though it may be argued that *authorisation* does not necessarily imply causation) (see Table 13.2). All of its example sentences are drawn from the semantic field of law.

Table 13.2 Extract from *New Missions 2e*, p. 208 (emphasis original)

verbe	construction (base verbale)	degré de pression
make	*someone* do *something*	*Her lawyer* made her admit *the crime*. Forte contrainte
have		*He* had the prisoner pay *for his crime*. Pression moins forte
let		*He* let his lawyer talk. Autorisation

New Missions 2e (p. 208) also lists *allow*, *ask*, *cause*, *encourage/incite/urge*, *force/oblige*, *get*, *order* and *persuade* as causative verbs. However, these are not further explained or illustrated with any example sentences.

Piece of Cake 4e (p. 166) features a relatively detailed description of causative constructions with the verbs HAVE, MAKE and LET. Aside from analysing their syntactic structures, it also contrasts a number of semantic aspects. In particular, the authors claim that HAVE causatives are used to delegate actions, whilst MAKE causatives imply that a (forced) action or an emotion is provoked.

The two non-French textbooks that explicitly mention causative constructions only do so fleetingly. *English in Mind 3* (pp. 49 and 53) focuses on HAVE constructions in the sense of *HAVE (something) done* and its exercises on the subject suggest that the semantics of the construction are limited to services provided to individuals (the highlighted prototypical example being: *having your hair done*), whilst *Solutions B2 Intermediate Plus* (p. 130) simply lists GET and HAVE as verbs entering causative constructions in the following pattern *verb + object + past participle* without providing any further information.

This section has provided an overview of how causative constructions are explicitly presented in the grammar sections of five EFL textbook series. The following sections investigate the nature of the MAKE, CAUSE and GET causatives that learners are exposed to implicitly when learning English with the textbooks of the TEC. The results are then compared to the results of existing EFL learner corpus research in Sect. 13.5.

13.4.2 Syntactic Analysis of Causative Constructions in the TEC-T

Table 13.3 displays the raw and relative frequencies (per million words, hereafter pmw) of the causative constructions extracted from the TEC-T with the combination of the automatic corpus queries and manual sorting (see Sect. 13.3.2). A cursory glance suffices to conclude that one construction is featured much more prominently than any other: [X MAKE Y V_{inf}], e.g., (1). It accounts for 97% of MAKE causative constructions in the TEC-T. Whilst some of the textbooks focus their grammar sections on causatives on the syntactic difficulty involved in the use of the causative verb MAKE in the passive (see Sect. 13.4.1), the [X BE made $V_{to\text{-}inf}$] construction is largely absent from the actual language input delivered by the textbooks: it is featured just once in two textbook series (*Access* and *Join the team*) and twice in *Green Line New*. No single occurrence of [$MAKE_{pas}$ V_{pas}] was observed across the entire TEC-T.

Table 13.3 Causative constructions in the TEC-T

Construction	Raw frequency in the TEC-T	Relative frequency in the TEC-T (pmw)
[X CAUSE Y $V_{to\text{-}inf}$]	8	5.98
[X GET Y V_{pp}]	13	9.71
[X GET Y V_{prp}]	2	1.49
[X MAKE Y V_{inf}]	289	215.88
[X MAKE Y V_{pp}]	5	3.73
[X BE made $V_{to\text{-}inf}$]	4	2.99

13.4.3 Collostructional Analysis of the [X MAKE Y V_{inf}] Construction in the TEC-T

Table 13.4 lists the collexemes of the [X MAKE Y V_{inf}] construction with the strongest collostructional strength (see Appendix 3 for full results). The collostructional strength column indicates whether the co-occurrence frequency of a verb collexeme in the construction is significantly different to what we would expect under the assumption of no association given the overall frequencies of the construction and the collexeme within the TEC-T (these correspond to the expected frequencies). The higher the value, the more strongly this verb is associated with the construction. The verb most strongly associated to the [X MAKE Y V_{inf}] construction in the TEC-T is, by far, FEEL, e.g.:

Table 13.4 Most significant results from the collexeme analysis of [X MAKE Y V_{inf}] in the TEC-T (for full results see Appendix 3)

Collexeme lemma	Lemma frequency in the TEC-T	Observed frequency in [X MAKE Y V_{inf}]	Expected frequency	Relation	Collostructional strength
FEEL	1412	65	1.27	attraction	88.74
LAUGH	252	14	0.23	attraction	20.36
SOUND	344	10	0.31	attraction	11.90
THINK	5432	25	4.88	attraction	10.34
PROMISE	87	4	0.08	attraction	5.87
WONDER	133	4	0.12	attraction	5.14
SEEM	301	5	0.27	attraction	5.04
SNEEZE	14	2	0.01	attraction	4.14
CONSIDER	122	3	0.11	attraction	3.71
CRY	139	3	0.12	attraction	3.54
UNDERSTAND	654	5	0.58	attraction	3.46
WANT	3280	10	2.94	attraction	3.05
DO	19,247	6	17.30	repulsion	−2.88

Collexeme lemma	Lemma frequency in the TEC-T	Observed frequency in [X MAKE Y V_{inf}]	Expected frequency	Relation	Collostructional strength
BE	73,499	1	66.06	repulsion	−30.80

(12) Hobbies are excellent ways to *make* you *feel* happy and they can also be a way of making friends. <TEC: Hi There 5e>

\r\n

Other strongly associated verbs include more verbs of emotion (CRY, LAUGH) and verbs of cognition (THINK, WONDER, PROMISE, CONSIDER, UNDERSTAND), e.g.:

(13) She's the funniest person in our family—she can really *make* you *laugh*. <TEC: English in Mind 1>
(14) Do you think the people in the photo spend a lot of time at the gym? What *makes* you *think* so? <TEC: Solutions B1 Intermediate>

\r\n

13.5 The Potential Impact of EFL Textbooks on EFL Writing

In this section, the relative frequencies of causative constructions in the TEC-T are compared to the relative frequencies found in corpora of EFL and ESL learner essay writing.

13.5.1 EFL Learners' Choice of Causative Constructions

Gilquin (2016: 127) reports that the relative frequency of the [X MAKE Y V_{inf}] construction is above 400 pmw in advanced EFL speakers (cf. Gilquin 2012: 8) and around 125 pmw in advanced ESL speakers. At 216 pmw, the relative frequency of this particular causative construction in the TEC-T is more than four times higher than in Gilquin's (2012: 8) reference ENL corpus (50.77 pmw). It is also considerably higher than in her ESL corpus data.

Only 13 instances of the [X GET Y V_{pp}] construction were found in the TEC-T. Interestingly, the construction's relative frequency in the TEC-T (9.71 pmw) is noticeably closer to that found in Gilquin's (2012: 8) EFL data (7.73 pmw in ICLE) than in the reference ENL data (3.80 pmw). The [X GET Y V_{prp}] construction has a similarly low frequency across all corpora. By contrast, Gilquin (2012) reported a significant underuse of the [X GET Y $V_{to\text{-}inf}$] construction by EFL learners and, indeed, not a single occurrence of this construction is found in the textbooks under study.

The second most frequently used causative construction in Gilquin's (2016) ENL data is [X CAUSE Y $V_{to\text{-}inf}$]. This construction is considerably less frequent in the TEC-T than in the ENL data. This difference may well be due to text register differences and would warrant further analysis with a comparable reference corpus. At the same time, however, it is worth noting that, according to Gilquin (2016), the [X CAUSE Y $V_{to\text{-}inf}$] construction is the only causative construction which ESL learners significantly overuse compared to ENL writers. By contrast, EFL learners use it to approximately the same extent as ENL writers (cf. Gilquin 2012: 48). The verb CAUSE is not infrequent in the TEC-T, but only 8% of occurrences are in periphrastic constructions, compared to 44% in British ENL data (Gilquin 2010: 37).

Although the learners of the learner corpus studies cited above presumably did not learn with exactly the textbooks of the TEC-T, they can be assumed to have been exposed to very similar pedagogical materials; thus, these results suggest that how frequently a construction is presented in typical textbook materials influences learners' choices in their own language production (see also Winter and Le Foll 2022). Indeed, by far the most frequent causative construction in the TEC-T, [X MAKE Y V_{inf}], is seemingly overused by EFL learners, whilst constructions that are underused either have a relative frequency that is considerably lower in the TEC-T than in ENL writing, or are entirely absent from the 43 textbooks of the TEC.

13.5.2 Unidiomatic Syntax in EFL Writing

Not only does textbook-based language input appear to influence EFL learners' choice of construction, the results also suggest that low exposure from textbooks corresponds to an increase in EFL users' production of unidiomatic syntax. Whereas periphrastic constructions with CAUSE occur more than 30 times pmw in

ENL and EFL writing (Gilquin 2012: 48), in Sect. 13.4.2 we saw that they are very rare in the TEC-T. This finding sheds new light on Gilquin's (2012: 50) conclusion that the causative constructions which trigger the most syntactic errors in EFL writing are those involving the verb CAUSE: nearly a quarter of CAUSE constructions in ICLE are reported to feature verb- and/or causee-related syntactic errors such as those illustrated in (17) and (18) (taken from Gilquin 2012: 49–50).

(15) The ease payment of credit card *cause the students buying* too much and too quickly. <ICLE: Chinese L1>
(16) *It causes to lose* their good emotions such as love, mercy and the other good ones. <ICLE: Turkish L1>

Furthermore, Gilquin (2016: 127–128) demonstrates that ESL speakers produce CAUSE constructions which follow the prototypical syntactic patterns used by ENL speakers significantly more frequently than EFL learners. Gilquin interprets this finding as a confirmation of the input-dependent L2 acquisition hypothesis because ESL speakers are likely to have much richer, extra-curricular language input than EFL students. The results of the present study support this hypothesis by pointing to the extremely low frequency of CAUSE constructions in school EFL textbooks. Moreover, CAUSE is not featured in the explicit treatment of causative constructions in the nine textbook series investigated (except for *New Missions 2*[e] where it is mentioned once, but with no example sentence or additional explanation, see Sect. 13.4.1).

Inversely, the results of the present corpus-based analysis of Textbook English suggest that high exposure through textbook input may have a positive impact on learners' production. Active [X MAKE Y V_{inf}] constructions have been shown to be considerably overused by EFL learners (Gilquin 2012, 2016) and, indeed, this study shows (see Sect. 13.4.3) that they are, by far, the most frequent causative constructions in the nine textbook series under study. In addition, Gilquin's (2016: 128–137) syntactic error analysis in ESL and EFL essays reveals that, with [X MAKE Y V_{inf}] constructions, EFL learners generally fare better than ESL students. Together, these findings suggest that EFL learners' generally good assimilation of the syntactic peculiarity of the [X MAKE Y V_{inf}] construction—which is unusual in that it takes a bare infinitive—may be, at least partially, the result of high exposure to this construction in EFL teaching materials.

The passive [X BE made $V_{to\text{-}inf}$] construction, on the other hand, was shown to be largely absent from the TEC-T. If the syntactically unusual active [X MAKE Y V_{inf}] construction is featured so prominently in textbooks, but the passive construction is more or less ignored, this is likely to have an impact on EFL learners' idiomaticity when they do attempt the passive construction. Indeed, Gilquin (2016: 131) reports that whilst ESL students who are likely exposed to more naturalistic language input use native-like syntax in 100% of attested cases, some EFL learners, on the other hand, seemingly overgeneralise the rule from the active construction and thus produce sentences such (17) (from Gilquin 2016: 133).

(17) Czech students majoring in languages, for example, are often *made believe* that grammar is the most important part of language […]. <ICLE: Czech L1>

\r\n

Gilquin (2016: 136) hypothesises that MAKE, often perceived as the most prototypical causative verb (see also Altenberg 2002: 99), may "serve as a pathbreaking verb in instruction, that is, a verb that is used to introduce the general characteristics of the construction, before the construction is extended to other verbs." Whilst this pedagogical strategy appears to bear fruit for the active [X make Y V_{inf}] construction, it may be misleading some learners into assuming that all causative constructions follow this syntactic pattern with a bare infinitive. The present results thus underline the inseparability of grammar and vocabulary in language teaching and learning: each causative construction needs to be taught as a lexico-grammatical unit with its own syntactic patterns, semantics, and, as will be discussed in the following section, frequent lexical associations.

13.5.3 EFL Learners' Use of the [X MAKE Y V_{inf}] Construction

The collexeme analysis in Sect. 13.4.3 shows that [X MAKE Y FEEL] is by far the most salient instantiation of causative constructions in the TEC-T. Interestingly, both Liu and Shaw (2001) and Gilquin (2012, 2016) report that FEEL is used significantly more frequently by EFL than ENL users. In addition, BE and BECOME are also overused in this construction by learners (Liu and Shaw 2001: 179, Gilquin 2016: 139). In both cases, however, such an overuse of BE and BECOME is likely to be the result of learners using a verb complement where an adjective complement would have sufficed, see example (20) from Gilquin (2012: 14), and, as such, we would not expect textbooks to feature such infelicitous constructions.

(18) That will *make it be* more popular. <ICLE: Tswana L1>

Indeed, the results in Table 13.4 show that BE is the verb lemma most strongly repelled by the [X MAKE Y V_{inf}] construction. Thus, unlike the use of FEEL, learners' overuse of BE and BECOME in this construction does not appear to be textboook-induced. This finding also supports the input-dependent L2 acquisition hypothesis as Gilquin (2016: 139–140) observes an overuse of BE and BECOME for both ESL and EFL learners, whereas the overuse of FEEL is seemingly exclusive to EFL learners who can be assumed to rely much more on textbook input than ESL users.

13.6 Pedagogical Applications

In response to the results presented in this chapter, teachers may want to supplement textbook materials with their own corpus-informed materials. This could be achieved by using the web-based corpus platform Sketch Engine to create data-driven learning activities. Figure 13.1 shows how to search for MAKE causative constructions in the British National Corpus (BNC). Under the "Advanced" tab (a), Sketch Engine's concordance function has a CQL "Query type" option (b). Having selected this option, the CQL queries formulated in Sect. 13.3.2 can be copied into the search field (c). If desired, the "Text types" drop-down menu (d) can be used to select a register subcorpus of interest, e.g., "Spoken demographic" (e) for everyday conversation.

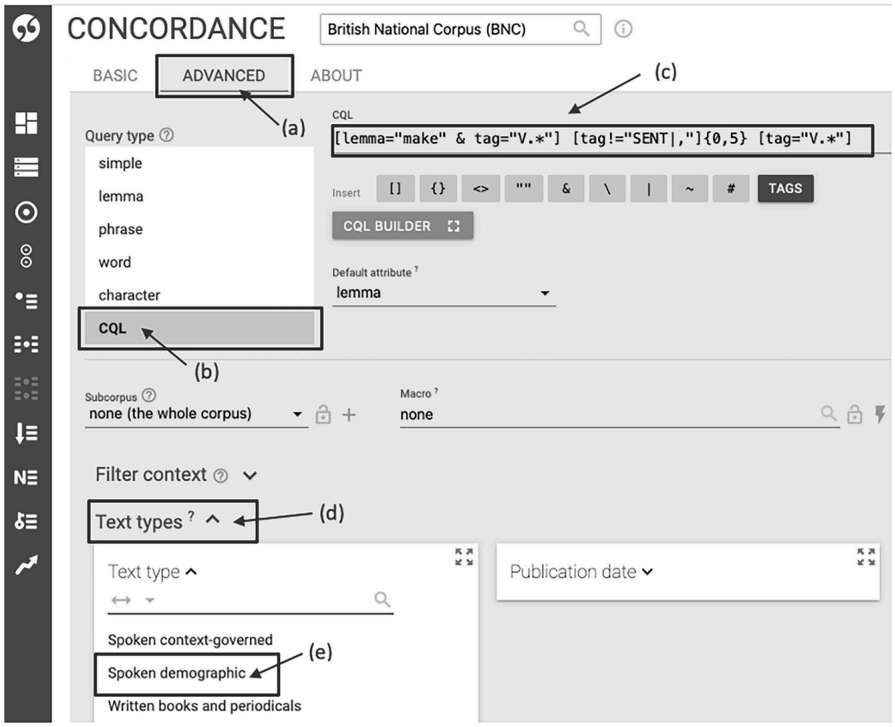

Fig. 13.1 Screenshot of the advanced concordance search function on sketchengine.eu

Relevant concordance lines can then be selected and downloaded for students to explore causatives in real-life English (see Fig. 13.2). Alternatively, tasks could involve students exploring a set of unsorted concordance lines and identifying which excerpts represent causative constructions (for detailed instructions and further ideas on how to use Sketch Engine to develop corpus-informed materials for the EFL classroom, see Le Foll 2021b: Chaps. 2, 5, 8, 9 and 15).

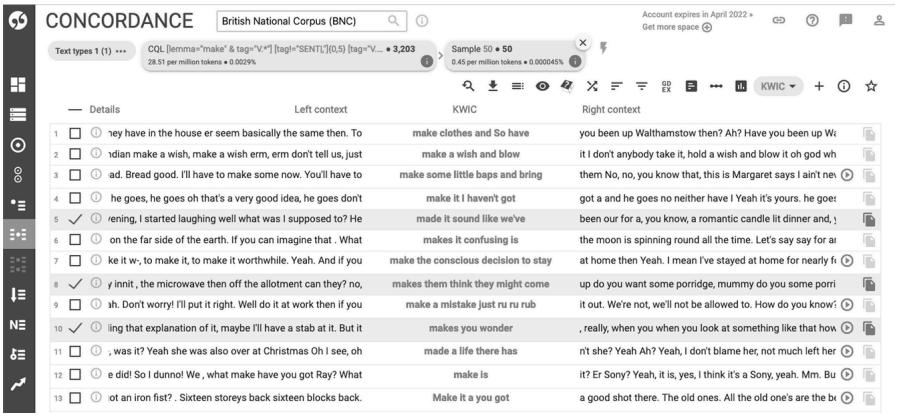

Fig. 13.2 Screenshot of a random sample of query results with three selected concordance lines

In addition to ENL corpora such as the British National Corpus (BNC), Sketch Engine also hosts the Open Cambridge Learner Corpus (OCLC), which textbook authors, editors and EFL teachers can query to gain a better understanding of typical learner difficulties with constructions such as causatives. Figure 13.3 shows how, under the "Text types" (a) options of the OCLC, it is possible to narrow down the search to learners of a specific L1, e.g., Spanish (b), education level (c), CEFR level (d), etc.

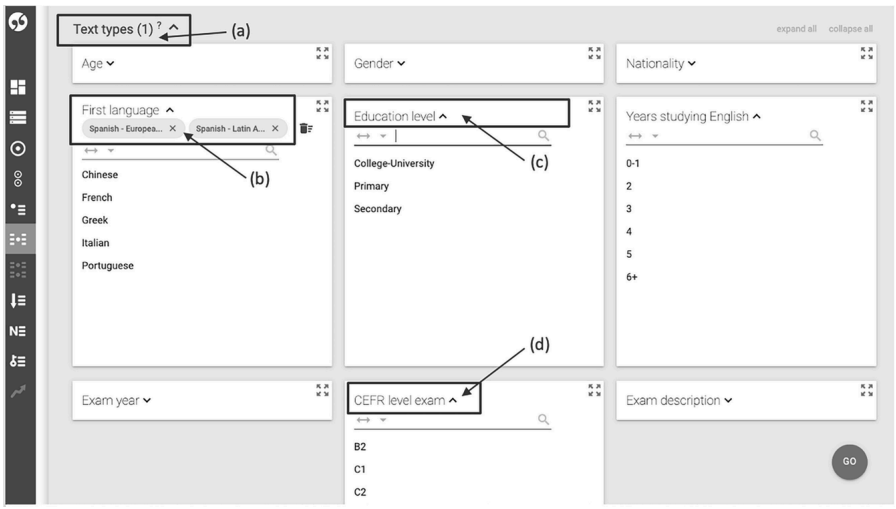

Fig. 13.3 Screenshot of the text type options when querying the OCLC

Thus, different types of corpora can be used to design materials that supplement traditional textbook materials. In particular, corpus-informed materials can raise awareness of the most frequent semantic and lexical associations of each construction and any register-specific constraints. They can help pre-empt learners' frequent syntactic errors and unidiomatic lexical and semantic associations.

13.7 Conclusions

This study set out to examine the potential impact of EFL textbook language on EFL learner output. For this purpose, representations of causative constructions in school EFL textbooks were examined, with specific focus on periphrastic constructions involving three of the most frequent causative verbs: CAUSE, GET and MAKE. The results have been compared to the findings of previous learner corpus studies with the aim of further exploring the input-dependent L2 acquisition hypothesis (Gilquin 2016) which suggests that EFL learners' use of specific constructions is likely to be influenced by the distributions of these same constructions in their main source of L2 input, namely their textbooks. Consequently, EFL teachers may want to draw on ENL corpus data to expose their students to constructions that are underrepresented in teaching materials (e.g., GET causatives). Additionally, learner corpora can be used to raise awareness of frequent difficulties that learners of specific L1s face in attempting to produce particular constructions.

As in all corpus-based investigations, the observations made in this study are limited to the data included in the TEC-T and may only very tentatively be generalised to the entire population, in this case, of all lower secondary school EFL textbooks. Even more tentative are the hypotheses drawn from comparisons between this study's results and those obtained from learner corpora because they are based on different text registers. The learner corpora queried in Gilquin (2012, 2016), Altenberg and Granger (2001) and Liu and Shaw (2001) all consist of academic essays, a register which is not widely represented in secondary school EFL textbooks. Medium and register, however, have been shown to have a major influence on the distribution and phraseological patterns of causative constructions in English. For instance, the [X MAKE Y V_{inf}] construction has been shown to be about four times more frequent in spoken than in written language (Gilquin 2010: 226). Moreover, Gilquin (2012: 58) points out that [X MAKE Y FEEL] is highly frequent in spoken English, and demonstrates using distinctive collexeme analysis that the frequency of use of [X MAKE Y FEEL] in EFL academic essays is closer to its frequency in ENL speech than in ENL essay writing. It may therefore be tempting to conclude that school EFL textbooks are more representative of spoken than written language; however, this hypothesis does not entirely hold, since it would also imply that causative constructions with GET ought to be far more frequent than they were shown to be in Sect. 13.4.2. Register-based variation ought to be considered when teaching causatives. Since current EFL textbooks do not do so adequately (Le Foll 2021a, 2022a, 2022b), Sect. 13.7 showed how textbook authors and teachers can use Sketch Engine to explore how causative constructions vary in different registers, e.g., academic writing vs. conversation.

Appendix 1

To best compare pedagogical materials used in different educational systems, each textbook was labelled for proficiency level on a universal scale of A to E: level A textbooks correspond to the first year of EFL instruction at secondary level, in other words, beginner level to roughly A1 on the CEFR scale, whilst level E corresponds to the fifth year (CEFR B1–B2). Note that French textbook series only cover the first four years of secondary school (which take place at *Collèges*), which is why, whenever possible, a textbook from the same publisher corresponding to the fifth year of instruction (the first year of *Lycée*) was added. At the time of corpus compilation, Le Livre Scolaire did not produce any textbooks for *Lycées* (Table 13.5).

Table 13.5 Composition of the Textbook English Corpus (TEC)

Country of use	Publisher	Textbook series	Volume	Level	Publication date
France	Bordas	Hi There	6ème	A	2012
			5ème	B	2013
			4ème	C	2014
			3ème	D	2015
		New Mission	2nde	E	2014
	Nathan	Join the Team	6ème	A	2010
			5ème	B	2011
			4ème	C	2012
			3ème	D	2013
		New Bridges	2nde	E	2010
	Le Livre Scolaire	Piece of Cake	6ème	A	2017
			5ème	B	2017
			4ème	C	2017
			3ème	D	2017
Germany	Klett	Green Line	1	A	2006
			2	B	2006
			3	C	2007

Country of use	Publisher	Textbook series	Volume	Level	Publication date
			4	D	2008
			5	E	2009
	Klett	Green Line New	1	A	2014
			2	B	2015
			3	C	2016
			4	D	2017
			5	E	2018
	Cornelsen	Access	1	A	2013
			2	B	2014
			3	C	2015
			4	D	2016
			5	E	2017
Spain	Richmond	Achievers	A1+	A	2015
			A2	B	2015
			B1	C	2015
			B1+	D	2015
			B2	E	2015
	Cambridge University Press	English in Mind	Starter	A	2010
			1	B	2010
			2	C	2010
			3	D	2011
			4	E	2011
	Oxford University Press	Solutions	Elementary	A	2014
			Pre-Intermediate	B	2016
			Intermediate	C	2017
			Intermediate Plus	D	2017

The full bibliographic metadata is available on doi.org/10.5281/zenodo.4922819.

Appendix 2

Register subcorpora	Example texts	Number of tokens	% of annotated tokens
Isolated individual words, phrases or sentences	Exercises, vocabulary lists, gap-filling or phrase matching activities	914,083	34.09%
Instructional language	Instructions, grammar explanations	592,061	22.08%
Conversation	Dialogues, audio and video transcripts, speech bubbles	510,954	19.05%
Informative text	Fact boxes, newspaper articles, reports, webpages	302,916	11.30%
Narrative writing	Short stories, extracts of novels, story-like introductions to dialogues	253,888	9.47%
Personal communication	Blog and diary entries, personal letters and e-mails	67,098	2.50%
Poetry	Poems, songs, rhyme	26,194	0.98%
Other	Questionnaires, advertisements, formal letters, jokes, recipes	14,394	0.54%
	Total annotated tokens	**2,681,588**	

Appendix 3

word.freq: frequency of the word in the corpus

obs.freq: observed frequency of the word with/in causative_MAKE

exp.freq: expected frequency of the word with/in causative_MAKE

faith: percentage of how many instances of the word occur with/in causative_MAKE
relation: relation of the word to causative_MAKE
delta.p.constr.to.word: delta p: how much does the word/construction help guess the word?
delta.p.constr.to.word: delta p: how much does the construction help guess the word/construction?
coll.strength: index of collocational/collostructional strength: -log10 (Fisher-Yates exact, one-tailed), the higher, the stronger

words	Word freq	Obs. freq	Exp. freq	Relation	Faith	Delta.p. constr.to.word	Delta.p. word.to.constr	Coll.strength
feel	1412	65	1.2691	attraction	0.046	0.2278	0.0453	88.7419
laugh	252	14	0.2265	attraction	0.0556	0.0492	0.0547	20.355
sound	344	10	0.3092	attraction	0.0291	0.0346	0.0282	11.8997
think	5432	25	4.8822	attraction	0.0046	0.0719	0.0038	10.3363
promise	87	4	0.0782	attraction	0.046	0.014	0.0451	5.8727
wonder	133	4	0.1195	attraction	0.0301	0.0139	0.0292	5.1389
seem	301	5	0.2705	attraction	0.0166	0.0169	0.0157	5.0425
sneeze	14	2	0.0126	attraction	0.1429	0.0071	0.142	4.1383
consider	122	3	0.1097	attraction	0.0246	0.0103	0.0237	3.7079
cry	139	3	0.1249	attraction	0.0216	0.0103	0.0207	3.5415
understand	654	5	0.5878	attraction	0.0076	0.0158	0.0068	3.4614
want	3280	10	2.948	attraction	0.003	0.0252	0.0022	3.0519
cringe	1	1	0.0009	attraction	1	0.0036	0.9991	3.0463
empathise	1	1	0.0009	attraction	1	0.0036	0.9991	3.0463
wince	1	1	0.0009	attraction	1	0.0036	0.9991	3.0463
stop	826	5	0.7424	attraction	0.0061	0.0152	0.0052	3.0072
happen	1387	6	1.2466	attraction	0.0043	0.0170	0.0034	2.759
itch	3	1	0.0027	attraction	0.3333	0.0036	0.3324	2.5696
sweat	3	1	0.0027	attraction	0.3333	0.0036	0.3324	2.5696
come	2638	8	2.371	attraction	0.003	0.0201	0.0022	2.5318
throw up	5	1	0.0045	attraction	0.2	0.0036	0.1991	2.3481
work	768	4	0.6903	attraction	0.0052	0.0118	0.0043	2.2696
ache	6	1	0.0054	attraction	0.1667	0.0036	0.1658	2.2692
suspect	6	1	0.0054	attraction	0.1667	0.0036	0.1658	2.2692
poke	7	1	0.0063	attraction	0.1429	0.0036	0.142	2.2024
stand out	7	1	0.0063	attraction	0.1429	0.0036	0.142	2.2024
disappear	136	2	0.1222	attraction	0.0147	0.0067	0.0138	2.1659
change	843	4	0.7577	attraction	0.0047	0.0116	0.0039	2.13
realize	168	2	0.151	attraction	0.0119	0.0066	0.011	1.9899
appear	172	2	0.1546	attraction	0.0116	0.0066	0.0107	1.9704
stretch	18	1	0.0162	attraction	0.0556	0.0035	0.0547	1.7944
appreciate	19	1	0.0171	attraction	0.0526	0.0035	0.0517	1.7711
flow	20	1	0.018	attraction	0.05	0.0035	0.0491	1.749
sign	22	1	0.0198	attraction	0.0455	0.0035	0.0446	1.708
react	242	2	0.2175	attraction	0.0083	0.0064	0.0074	1.6909
regret	24	1	0.0216	attraction	0.0417	0.0035	0.0408	1.6706
correspond	31	1	0.0279	attraction	0.0323	0.0035	0.0314	1.5608
grow	285	2	0.2562	attraction	0.007	0.0062	0.0061	1.5595
dream	47	1	0.0422	attraction	0.0213	0.0034	0.0204	1.3832
bite	54	1	0.0485	attraction	0.0185	0.0034	0.0176	1.3242
tidy	55	1	0.0494	attraction	0.0182	0.0034	0.0173	1.3164

words	Word freq	Obs. freq	Exp. freq	Relation	Faith	Delta.p. constr.to.word	Delta.p. word.to.constr	Coll.strength
stand up	63	1	0.0566	attraction	0.0159	0.0034	0.015	1.259
pay	420	2	0.3775	attraction	0.0048	0.0058	0.0039	1.2562
go	7419	11	6.6681	attraction	0.0015	0.0155	0.0006	1.1301
look	5761	9	5.1779	attraction	0.0016	0.0137	0.0007	1.106
rise	95	1	0.0854	attraction	0.0105	0.0033	0.0096	1.0868
believe	555	2	0.4988	attraction	0.0036	0.0054	0.0027	1.0475
cut	127	1	0.1141	attraction	0.0079	0.0032	0.007	0.9668
get up	146	1	0.1312	attraction	0.0068	0.0031	0.006	0.9099
hurt	160	1	0.1438	attraction	0.0062	0.0031	0.0054	0.8728
teach	169	1	0.1519	attraction	0.0059	0.003	0.005	0.8507
buy	773	2	0.6948	attraction	0.0026	0.0047	0.0017	0.813
improve	186	1	0.1672	attraction	0.0054	0.003	0.0045	0.8123
fly	198	1	0.178	attraction	0.0051	0.0029	0.0042	0.7874
prefer	219	1	0.1968	attraction	0.0046	0.0029	0.0037	0.7476
decide	1119	2	1.0057	attraction	0.0018	0.0036	0.0009	0.5745
bring	388	1	0.3487	attraction	0.0026	0.0023	0.0017	0.5306
forget	409	1	0.3676	attraction	0.0024	0.0023	0.0015	0.5116
fall	420	1	0.3775	attraction	0.0024	0.0022	0.0015	0.5021
hope	424	1	0.3811	attraction	0.0024	0.0022	0.0015	0.4987
love	1297	2	1.1657	attraction	0.0015	0.003	0.0006	0.488
keep	580	1	0.5213	attraction	0.0017	0.0017	0.0008	0.3907
wear	585	1	0.5258	attraction	0.0017	0.0017	0.0008	0.3879
sit	595	1	0.5348	attraction	0.0017	0.0017	0.0008	0.3823
stay	617	1	0.5546	attraction	0.0016	0.0016	0.0007	0.3705
spend	628	1	0.5644	attraction	0.0016	0.0016	0.0007	0.3647
guess	737	1	0.6624	attraction	0.0014	0.0012	0.0005	0.3143
run	758	1	0.6813	attraction	0.0013	0.0011	0.0004	0.3057
become	931	1	0.8368	attraction	0.0011	0.0006	0.0002	0.246
watch	974	1	0.8754	attraction	0.001	0.0004	0.0001	0.2336
help	2365	2	2.1256	repulsion	0.0008	−0.0004	−0.0001	−0.1921
listen	1587	1	1.4264	repulsion	0.0006	−0.0015	−0.0003	−0.2348
take	3177	2	2.8555	repulsion	0.0006	−0.0031	−0.0003	−0.3418
see	4185	2	3.7614	repulsion	0.0005	−0.0063	−0.0004	−0.5637
say	6719	3	6.039	repulsion	0.0004	−0.0109	−0.0005	−0.8394
use	5548	2	4.9865	repulsion	0.0004	−0.0107	−0.0005	−0.9086
make	4831	1	4.3421	repulsion	0.0002	−0.0119	−0.0007	−1.1678
read	5070	1	4.5569	repulsion	0.0002	−0.0127	−0.0007	−1.2452
get	5105	1	4.5883	repulsion	0.0002	−0.0128	−0.0007	−1.2566
do	19247	6	17.299	repulsion	0.0003	−0.0404	−0.0006	−2.8828
be	73499	1	66.0602	repulsion	0	−0.2326	−0.0012	−30.7971

In order to determine the degree of repulsion of verbs that are not attested with/in causative_MAKE, the following table gives the collocational/collostructional strength for all verb frequencies in orders of magnitude the corpus size allows for.

Absents.words	Absents.obs.freqs	Absents.exp.freqs	Relation	Absents.delta.p.constr.to.word	Absents.delta.p.word.to.constr	Absent
a	10	0.009	repulsion	0	−0.0009	0.5527
b	100	0.0899	repulsion	−0.0003	−0.0009	2.14
c	1000	0.8988	repulsion	−0.0032	−0.0009	0.0984

Absents.words	Absents.obs.freqs	Absents.exp.freqs	Relation	Absents.delta.p.constr.to.word	Absents.delta.p.word.to.constr	Absent
d	10000	8.9879	repulsion	−0.0321	−0.0009	2.9323
e	100000	89.879	repulsion	−0.3213	−0.0013	0.1344

If your collostruction strength is based on *p*-values, it can be interpreted as follows: Coll.strength>3; $p<0.001$; coll.strength>2; $p<0.01$; coll.strength>1.30103; $p<0.05$.

References

Altenberg, Bengt. 2002. Causative Constructions in English and Swedish: A Corpus-based Contrastive Study. In *Lexis in Contrast: Corpus-based Approaches*, ed. Bengt Altenberg and Sylviane Granger, 97–116. Amsterdam: John Benjamins.
[Crossref]

Altenberg, Bengt, and Sylviane Granger. 2001. The Grammatical and Lexical Patterning of MAKE in Native and Non-native Student Writing. *Applied Linguistics* 22 (2): 173–195.
[Crossref]

Croft, William. 2001. *Radical Construction Grammar: Syntactic Theory in typological Perspective* (pp. 111–112). Oxford, New York: Oxford University Press.

Gabrielatos, Costas. 2013. *If*-conditionals in ICLE and the BNC: A Success Story for Teaching or Learning? In *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead*, ed. Sylviane Granger, Fanny Meunier, and Gaëtanelle Gilquin, 155–166. Presses Universitaires de Louvain.

Gilquin, Gaëtanelle. 2010. *Corpus, Cognition and Causative Constructions (Studies in Corpus Linguistics 39)*. Amsterdam: John Benjamins.
[Crossref]

———. 2012. Lexical Infelicity in Causative Constructions. Comparing Native and Learner Collostructions. In *Analytical Causatives: From 'Give' and 'Come' to 'Let' and 'Make'*, ed. Jaako Leino and Ruprecht von Waldenfels, 41–64. München: Lincom Europa.

———. 2016. Input-dependent L2 Acquisition: Causative Constructions in English as a Foreign and Second Language. In *Applied Construction Grammar*, ed. Sabine De Knop and Gaëtanelle Gilquin, 198–259. Berlin: De Gruyter.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

———. 2005. Argument Realization: The Role of Constructions, Lexical Semantics and Discourse Factors. In *Construction Grammars: Cognitive Grounding and Theoretical Extensions*, ed. Jan-Ola Östman and Mirjam Fried, 17–43. Amsterdam: John Benjamins.
[Crossref]

———. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English*. Louvain-la-Neuve: UCL Presses Universitaires de Louvain.

Gries, Stefan Th. 2007. *Coll.analysis 3.2a Package for R*. http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/index.html Accessed 14 February 2019.

Hausmann, Franz Josef. 1984. Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts* 31 (4): 395–406.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography* 1 (1): 7–36.
[Crossref]

Langacker, Ronald. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press.
[Crossref]

Le Foll, Elen. 2021a. Register Variation in School EFL Textbooks. *Register Studies* 3 (2). https://doi.org/10.1075/rs.20009.lef.

———. 2021b. *Creating Corpus-Informed Materials for the English as a Foreign Language Classroom: A Step-by-Step Guide for (Trainee) Teachers Using Online Resources* (Open Educational Resource). https://elenlefoll.pressbooks.com. Accessed 30 July 2021.

———. 2022a. MAKING Tea and Mistakes: The Functions of MAKE in Spoken English and Textbook Dialogues. In *Multifunctionality in English: Corpora, Language and Academic Literacy Pedagogy*, ed. Zihan Yin and Elaine Vine, 157–178. Oxon: Routledge.
[Crossref]

———. 2022b. Textbook English: A Corpus-Based Analysis of the Language of EFL textbooks used in Secondary Schools in France, Germany and Spain. Osnabrück, Germany: Osnabrück University. Unpublished PhD thesis.

Liu, Eric T.K., and Philip M. Shaw. 2001. Investigating Learner Vocabulary: A Possible Approach to Looking at EFL/ESL Learners' Qualitative Knowledge of the Word. *International Review of Applied Linguistics in Language Teaching* 39 (3): 171–194. https://doi.org/10.1515/iral.2001.001.
[Crossref]

Mindt, Dieter. 1987. *Sprache, Grammatik, Unterrichtsgrammatik: futurischer Zeitbezug im Englischen (Schule und Forschung)*. Frankfurt am Main: Diesterweg.

Möller, Verena. 2020. From Pedagogical Input to Learner Output: Conditionals in EFL and CLIL Teaching Materials and Learner Language. *Pedagogical Linguistics* 1 (2): 95–124. https://doi.org/10.1075/pl.00001.mol.
[Crossref]

Römer, Ute. 2005. *Progressives, Patterns, Pedagogy: A Corpus-Driven Approach to English Progressive Forms, Functions, Contexts, and Didactics (Studies in Corpus Linguistics 18)*. Amsterdam: John Benjamins.
[Crossref]

Stefanowitsch, Anatol, and Stefan Th. Gries. 2003. Collostructions: Investigating the Interaction of Words and Constructions. *International Journal of Corpus Linguistics* 8 (2): 209–243.
[Crossref]

Volkmann, Laurenz. 2010. *Fachdidaktik Englisch: Kultur und Sprache*. Tübingen: Narr.

Winter, Tatjana, and Elen Le Foll. 2022. Testing the Pedagogical Norm: Comparing *If*-conditionals in EFL Textbooks, Learner Writing and English Outside the Classroom. *International Journal of Learner Corpus Research* 8 (1): 31–66.

Wong, S.C. 1983. Overproduction & Unidiomatic Usage in the Make Causatives of Chinese Speakers: A Cause for Flexibility in Interlanguage analysis. *Language Learning & Communication* 2 (2): 151–165.

Footnotes

1 In this and the following examples, the code between angle brackets denotes the source of the example, i.e. the abbreviation of the corpus followed by the relevant subcorpus (e.g., textbook volume, learner L1 language) or document ID. Unless otherwise stated, emphasis was added.

2 The extraction of causative HAVE constructions was not attempted because these are relatively rare and their extraction would have required a lot of manual work. Indeed, Gilquin (2010: 37) reported that only ca. 0.35% of all occurrences of HAVE are instances of causative constructions.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
K. Harrington, P. Ronan (eds.), *Demystifying Corpus Linguistics for English Language Teaching*
https://doi.org/10.1007/978-3-031-11220-1_14

14. Conclusion

Patricia Ronan[1] ✉ and Kieran Harrington[1] ✉
(1) Faculty of Cultural Studies, TU Dortmund University, Dortmund, Germany

✉ **Patricia Ronan (Corresponding author)**
Email: patricia.ronan@tu-dortmund.de

✉ **Kieran Harrington**
Email: kieran.harrington@tu-dortmund.de

Keywords Corpus linguistics – Corpora in the classroom – Data driven learning – Direct corpus use – Indirect corpus use – TEFL – Second Language Acquisition – Four skills

14.1 Demystifying Corpus Linguistics for English Language Teaching

The contributions to this volume show clearly how language teaching in the (English-) language classroom can benefit from the use of corpus linguistic approaches. Corpus-based, data-driven learning approaches can be used profitably to improve students' skills in writing tasks (**Friginal** et al. 2022, this volume). In particular, students and pupils can benefit from corpus resources to improve their use of vocabulary and phrases (**Selivan** 2022, this volume, Schneider 2022, this volume). Corpus linguistic approaches are also beneficial for grammar instruction (**Jones** 2022, this volume) and they can help students to become more aware of specific phrases and constructions (**Le Foll** 2022, this volume). In addition to advancing these structural skills, the

receptive task of reading can be improved by using corpus linguistics (**Curado Fuentes** 2022, this volume), and positive effects can also be determined in speaking and listening abilities (**McCarthy and McCarten** 2022, this volume).

But not only can pupils' and students' competences in the four skills (reading, writing, listening and speaking) profit from the use of corpus linguistics in the classroom, the awareness of pragmatic and, following from this, cultural differences between home culture and target language and culture on the one hand, and between different varieties of the target language and culture on the other hand, can also be improved (**Harrington** 2022, this volume). Especially for furthering an understanding of differences between different Englishes, be they learner varieties or other international varieties (**Buschfeld and Weidle** 2022, this volume, **Riegler** 2022, this volume), the use of corpora in the classroom can provide instructors with many powerful and easily accessible options, for example in the form of the International Corpus of English (ICE) Corpora (Nelson et al. 2002) or the Corpus of Global Web-Based English (GloWbE; Davies 2013) or the Vienna-Oxford International Corpus of English (VOICE 3.0 2021).

Yet in order to exploit the full potential of corpus linguistic approaches, teacher training still needs to provide more tuition to trainee teachers in using corpus linguistics for the classroom (**Ronan** 2022, this volume) and materials producers could also turn to corpus resources to a larger extent to produce materials based on authentic language use (**Schneider** 2022, this volume, Le Foll 2022, this volume).

In addition to showing the value of corpus linguistic approaches in the classroom, various hands-on corpus-based activities are also suggested by the contributors to this volume. Many of these can either be directly employed in language classrooms or they can be adapted to language teachers' needs regarding specific class-levels or target groups. By doing so, the use of corpus linguistic approaches is facilitated and powerful corpus tools can be harnessed for everyday classroom activities which allow students or pupils and their teachers and instructors alike to profit from the multiple and extensive corpus resources that are at our fingertips.

14.2 Where Do We Go from Here

14.2.1 At Teacher Education Level

The contributors to this volume are joining the call for more extensive corpus use in language classrooms. Though constraints to and restrictions of corpus use in the classroom have been pointed out (see in particular Friginal et al. 2022, this volume), a number of additional benefits can be brought to the language classrooms. In particular these have been argued to be the possibility to provide pupils with authentic language data, to engage the pupils with their own small-scale research projects or to exploit the pupils' interest in digital media (for example Curado Fuentes 2022, Friginal et al. 2022, Selivan 2022, all this volume).

As various studies have shown (for example Mukherjee 2004; Farr 2008; Leńko-Szymańska 2014; Callies 2019), providing trainee teachers with access to corpora and introducing them to corpus tools is the most important step in this direction. Yet once the first steps are taken, more education will not necessarily lead to more willingness to use corpora in teaching (Ronan 2022, this volume) as other constraints may stop teachers from using corpora. In particular, these are time constraints either within the curriculum or in the preparation of learning materials or in classroom management. These constraints can be overcome by universities or teacher training colleges taking a more active role in helping the trainee teachers preparing targeted, corpus-based teaching materials using indirect corpus methods. To do so would not only alleviate the teachers' time constraints in preparing teaching materials but would also remove the concerns about potential technical problems that pupils or teachers might encounter in direct corpus use in the classroom, as well as the fear that the corpus work would take too much time.

On the other hand, investing additional time into direct corpus use in the classroom, especially with advanced students, will bring the added benefit that any type of digital competencies will be highly valuable to pupils and students after and outside school. So will be the ability to evaluate and collect materials independently. As shown by Templeton and Timmis (2022, this volume), especially in times where digital literacy has been becoming a more and more crucial tool to help

societies grapple not only with a deluge of (digitally mediated) information, but also with information that is not quality controlled anymore, digital literacy and an ability to spot alternative possibilities in source materials are highly valuable additional skills that corpus use can teach us today.

Instructors or researchers working in teacher education should thus show the trainee teachers how to use direct corpus methods in order to make the trainee teachers more comfortable with using corpora. But in addition to teaching preparing future teachers for direct corpus use, trainee teachers will profit from being familiarized with the use of indirect methods to target specific pupil needs.

14.2.2 At Teacher Level

For the teachers themselves, strong cases have also been made in favour of extending corpus use in the classrooms. On important argument is that corpora can allow teachers to complement scarce teaching resources. When teachers identify gaps in teaching materials, such as indicated by Le Foll (2022, this volume), turning to corpus resources for direct or indirect corpus use can be a very good approach. Students with lower-level competencies can profit from indirect corpus use to see how target structures are used in everyday language. Students with advanced level competencies can profit from direct corpus use where they can obtain an overview of different types of structures that can be employed to encode a target expression. Even lower-level students may profit from obtaining an overview over a wider variety of expressions. They may pick their favourite structure and thus adopt their own "teddy bear" (see Schneider 2022, this volume).

Corpora can further complement teaching materials on international Englishes, which may form part of school curricula, but resources could not possibly be available for any given variety to the extent that everyone would wish for. Here, teachers, or indeed their pupils, might select corpora of the variety or varieties of English they are most interested in and identify and describe features of that variety on the basis of these corpus materials. As international Englishes are gaining importance not only in international communication as shown by various authors (Buschfeld and Weidle 2022, this volume; Riegler

2022, this volume), their introduction into the classrooms can be facilitated by corpus linguistic approaches. For advanced pupils or students, the use of learner corpora can also be very valuable to alert them to the question which language specific items merit attention.

But not only are corpora valuable for teaching language structures, they also provide an excellent basis for studying cultural differences that are encoded in language use (Harrington 2022, this volume). Corpus linguistics will thus provide an excellent resource for exploring these differences. These specific benefits of corpus work apart, a further important point speaks in favour of the use of corpus linguistic approaches in the classroom. This is the fact that in addition to teacher centred activities which have traditionally been found in language classrooms, as pointed out for example by Callies (2019), corpus linguistic approaches provide teachers with the opportunity to include more learner-centred activities in their instruction.

For all these endeavours, either more confidence in the use of corpus tools is necessary to allow teachers to introduce them into their classrooms or the availability of—and indeed familiarity with—easy-to-use tools which are appropriate for direct corpus use in language classrooms.

14.2.3 At Developer Level

As various contributions to this volume have shown, one restricting factor in direct corpus use, both for teachers and for pupils, is availability of and access to user-friendly corpus tools. It has thus become time to provide a community of interested users with easy-to-use corpus tools which will help to demystify corpus use and allow its broader accessibility not only for linguists who specialize in the use of corpus-based approaches, but also for a large number of language practitioners and emergent language users: teachers and their pupils. For this, a larger amount of user-friendly corpus interfaces is needed as argued for example by Selivan (2022, this volume) and Curado Fuentes (2022, this volume). As shown by these authors, Lextutor (Cobb 2021) is a valuable new resource that fulfils these requirements. In addition to being a powerful tool for vocabulary learning, it is easy to use and attractively designed. The design and creation of similarly powerful and attractive tools for further teaching needs, such as grammar

(Jones 2022, this volume) is an important desiderate that would bring new impulses to language instruction in the digital age.

In addition to the design of teacher and learner-centred corpus tools, corpus linguistic tools can also be used to bring forward the creation of targeted teaching materials. Schneider (2022, this volume) shows how materials that reflect authentic language use on phrasal and collocational levels can be compiled. While the approach presented in his study is likely to go beyond individual users' resources, they offer a useful approach to authors and publishers of teaching resources. Various contributors to this volume have pointed to the benefits in supplementing extant commercially published teaching resources with authentic materials (Harrington 2022, Jones 2022, Le Foll 2022, Selivan 2022, this volume), and the approach shown by Schneider illustrates a viable approach for doing so.

References

Callies, Marcus. 2019. Integrating Corpus Literacy into Language Teacher Education. The Case of Learner Corpora. In *Learner Corpora and Language Teaching*, ed. Sandra Götz and Joybrato Mukherjee, 245–263. Amsterdam: Benjamins.

Cobb, Tom. 2021. *Compleat Lextutor Tutor*. Lextutor 2021. https://www.lextutor.ca/. Accessed 1 March 2022.

Curado Fuentes, Alejandro. 2022. Corpus Affordances in Foreign Language Reading Comprehension. In *Demystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington, and Patricia Ronan.

Davies, Mark. 2013. Corpus of Global Web-Based English: 1.9 Billion Words from Speakers in 20 Countries (GloWbE). Available Online at https://corpus.byu.edu/glowbe/. Accessed 20 March 2022.

Farr, Fíona. 2008. Evaluating the Use of Corpus-Based Instruction in a Language Teacher Education Context: Perspectives from the Users. *Language Awareness* 17 (1): 25–43. [Crossref]

Friginal, Eric, Cox, Ashleigh, and Rachelle Udell. 2022. Corpus Linguistics and Writing Instruction. In *Demystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington, and Patricia Ronan.

Harrington, Kieran. 2022. Culture in English Language Teaching: Let the Language Do the Talking. In *Demystifying Corpus Linguistics for English Language Teaching,* ed. Kieran Harrington, and Patricia Ronan.

Jones, Chris. 2022. Corpus Linguistics and Grammar Teaching. In *Demystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington, and Patricia Ronan.

Le Foll, Elen. 2022. Causatives in School EFL Textbooks. In D*emystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington, and Patricia Ronan.

Leech, Geoffrey. 1997. Teaching and Language Corpora: A Convergence. In *Teaching and Language Corpora*, ed. Anne Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles, 1–23. London: Longman.

Leńko-Szymańska, Agnieszka. 2014. Is This Enough? A Qualitative Evaluation of the Effectiveness of a Teacher-Training Course on the Use of Corpora in Language Education. *ReCALL* 26 (2): 1–19. https://doi.org/10.1017/S095834401400010X. [Crossref]

McCarthy, Michael, and Jeanne McCarten. 2022. Speaking and Listening: Two Sides of the Same Coin. In *Demystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington, and Patricia Ronan.

McEnery, Tony, and Richard Xiao. 2010. What corpora can offer in language teaching and learning. In *Handbook of Research in Second Language Teaching and Learning*, ed. Eli Hinkel, vol. 2, 364–380. London & New York: Routledge.

Mukherjee, Joybrato. 2004. Bridging the Gap between Applied Corpus Linguistics and the Reality of English Language Teaching in Germany. In *Applied Corpus Linguistics*, *A Multidimensional Perspective Language and Computers*, ed. Ulla Connor and Thomas A. Upton, vol. 52, 239–250. Amsterdam: Rodopi.

Nelson, Gerald, Sean Wallis, and Bas Aarts, eds. 2002. *Exploring Natural Language Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

O'Keeffe, Anne, Michael McCarthy, and Ronald Carter. 2007. *From Corpus to Classroom*. Cambridge: Cambridge University Press. [Crossref]

Reppen, Randi. 2010. *Using Corpora in the Language Classroom*. Cambridge: Cambridge University Press.

Riegler, Stefanie. 2022. Annotating VOICE for Pedagogic Purposes – The Case for a Mark-Up Scheme of Pragmatic Functions in ELF Interactions. In *Demystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington, and Patricia Ronan.

Ronan, Patricia. 2022. Learning to Teach English as a Foreign Language with Corpus Linguistic Approaches: A Survey of Teacher Training Students' Attitudes. In *Demystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington, and Patricia Ronan.

Schneider, Gerold. 2022. Detecting and Analysing Learner Difficulties Using a Learner Corpus Without Error Tagging. In *Demystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington, and Patricia Ronan.

Selivan, Leo. 2022. Corpus Linguistics and Vocabulary Teaching. In *Demystifying Corpus Linguistics for English Language Teaching*, ed. Kieran Harrington, and Patricia Ronan.

VOICE. 2021. The Vienna-Oxford International Corpus of English (Version 3.0 Online). https://voice3.acdh.oeaw.ac.at. Accessed 7 October 2021.

Index

A

Annotation
  corpus
  pragmatic
Antconc
Attitudes
  corpus use
  student and teacher attitudes
Autonomy

B

Bank of English Corpus
BNC, *see* British National Corpus
British National Corpus (BNC)
Brown Corpus

C

Cambridge and Nottingham Corpus of Discourse English
Cambridge English Corpus (CEC)
Causative constructions
CHEsS corpus
Chunk(s)
Claws
COCA, *see* Corpus of Contemporary American English
COHA, *see* Corpus of Historical American English
Collins Birmingham University International Language Database (COBUILD)
Collocation fork
Collocation(s)
Collostructional analysis
Communication strategies
Compleat Lextutor
Concordances
Concordancing
Concordancing tools

AntConc
Wordsmith
Consciousness raising
Constituent Likelihood Automatic Word-tagging System (CLAWS)
Construction Grammar
Conversation
Corpora and writing instruction
Corpus
annotation
interfaces
linguistics
query
Corpus-based approach
Corpus-driven approach
Corpus of Contemporary American English (COCA)
Corpus of Historical American English (COHA)
Corpus query language (CQL)
Corpus use
advantages
direct
disadvantages
indirect
CQL, *see* Corpus Query Language
Culture
and language
Curriculum design and materials development

D

Data-driven
Data-driven approaches
Data-driven learning (DDL)
hard version
soft version
Development, teaching materials
Digital competencies
Digital literacy
Digital natives

Digital tools
Direct data-driven learning

E

English as a lingua franca (ELF)
English for Specific Purposes (ESP)
English language teaching (ELT)
  classroom
  in Germany
  practical considerations
Error tagging

F

Form
Framework
Frequency
Function

G

Galloway
Global Web-Based Corpus of English (GloWbE)
Grammar
Grammatical parsing

I

ICLE, *see* International Corpus of Learner English
Ideal
Indirect data-driven learning
Input
Intermediate levels
International Corpus of English (ICE)
  corpus
International Corpus of Learner English (ICLE)

K

Keyness
Key words
Key words in context (KWIC)
KWIC, *see* Key words in context

L

L2 writing pedagogy
Lancaster Oslo Bergen Corpus (LOB)
Language and culture
Language pedagogy
Learner corpora
Learner corpus
Learner corpus research (LCR)
Learning
Lemma
Lexical approach
Lexical bundles
Lexico-grammar
Lextutor
Listening skills

M

Multiword expression
Multi-word unit

N

Native speaker
  ideal
  new
Naturalness
New
Normativity
Nottingham Corpus of Discourse English

P

Parser
Phraseology
Pos tagging
Practical considerations
Pragmatic annotation
Pragmatic practices
Proximal development principle

Q
Questionnaire studies

R
Reading comprehension
Response

S
Sketch Engine
Skills
  listening
  reading
  speaking
  writing
Sociopragmatics
Sorry
Speaking skills
Student and teacher attitudes
Surprisal

T
Tagging
Teacher development
Teaching linguistic variation
Teddy bear effect
Text-based approach
Textbooks
Theory-practice divide
Token
Turn-taking
Type/token ratio

V
Varieties of English
Vienna International Corpus of English (VOICE)
Vienna-Oxford International Corpus of English (VOICE)
Vocabulary

W

Word count
Word frequency listing
Wordlists
Wordsmith
Writing
Writing skills
Written academic discourse