Exploring Corpus Linguistics

Routledge Introductions to Applied Linguistics is a series of introductory level textbooks covering the core topics in Applied Linguistics, primarily designed for those entering postgraduate studies and language professionals returning to academic study. The books take an innovative 'practice to theory' approach, with a 'back-to-front' structure. This leads the reader from real-world problems and issues, through a discussion of intervention and how to engage with these concerns, before finally relating these practical issues to theoretical foundations. Additional features include tasks with commentaries, a glossary of key terms, and an annotated further reading section.

Corpus linguistics is a key area of applied linguistics and one of the most rapidly developing. Winnie Cheng's practical approach guides readers in acquiring the relevant knowledge and theories to enable the analysis, explanation and interpretation of language using corpus methods.

Throughout the book practical classroom examples, concordance based analyses and tasks such as designing and conducting mini-projects are used to connect and explain the conceptual and practical aspects of corpus linguistics.

Exploring Corpus Linguistics is an essential textbook for post-graduate/graduate students new to the field and for advanced undergraduates studying English Language and Applied Linguistics.

Winnie Cheng is Professor of English in the Department of English, The Hong Kong Polytechnic University. Her publications include *Intercultural Conversation* (2003), *A Corpus-driven Analysis of Discourse Intonation* (2008), *Professional Communication: Collaboration between academics and practitioners* (2009), and *Language for Professional Communication: Research, practice & training* (2009).

www.routledge.com/cw/rial

Routledge Introductions to Applied Linguistics

Series editors:

Ronald Carter, *Professor of Modern English Language,*
University of Nottingham, UK

Guy Cook, *Professor of Language and Education*
Open University, UK

Routledge Introductions to Applied Linguistics is a series of introductory level textbooks covering the core topics in Applied Linguistics, primarily designed for those entering postgraduate studies and language professionals returning to academic study. The books take an innovative 'practice to theory' approach, with a 'back-to-front' structure. This leads the reader from real-world problems and issues, through a discussion of intervention and how to engage with these concerns, before finally relating these practical issues to theoretical foundations. Additinal features include tasks with commentaries, a glossary of key terms and an annotated further reading section.

Exploring English Language Teaching
Language in Action
Graham Hall

Exploring Classroom Discourse
Language in Action
Steve Walsh

Exploring Corpus Linguistics
Language in Action
Winnie Cheng

> 'The innovative approach devised by the series editors will make this series very attractive to students, teacher educators, and even to a general readership, wanting to explore and understand the field of applied linguistics. The volumes in this series take as their starting point the everyday professional problems and issues that applied linguists seek to illuminate. The volumes are authoritatively written, using an engaging 'back-to-front' structure that moves from practical interests to the conceptual bases and theories that underpin applications of practice.'
>
> Anne Burns, *Aston University, UK,*
> *University of New South Wales, Australia*

www.routledge.com/cw/rial

Exploring Corpus Linguistics

Language in Action

Winnie Cheng

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

www.routledge.com/cw/rial

First published 2012
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada
by Routledge
711 Third Avenue, New York, NY 10017

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

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

Library of Congress Cataloging in Publication Data
Cheng, Winnie.
Exploring corpus linguistics: language in action/Winnie Cheng.
p. cm. – (Routledge Introductions to Applied Linguistics)
Includes bibliographical references and index.
1. Corpus linguistics. 2. Language and languages – Study and teaching.
3. Applied linguistics.
I. Title.
P126.C68.C54 2011
410.1′88 – dc22 2011014235

ISBN 13: 978–0–415–58546–0 (hbk)
ISBN 13: 978–0–415–58547–7 (pbk)
ISBN 13: 978–0–203–80263–2 (ebk)

Typeset in Sabon
by Florence Production Ltd, Stoodleigh, Devon

Contents

www.routledge.com/cw/rial

Series editors' introduction

The *Introducing Applied Linguistics* series

This series provides clear, authoritative, up-to-date overviews of the major areas of applied linguistics. The books are designed particularly for students embarking on masters-level or teacher-education courses, as well as students in the closing stages of undergraduate study. The practical focus will make the books particularly useful and relevant to those returning to academic study after a period of professional practice, and also to those about to leave the academic world for the challenges of language-related work. For students who have not previously studied applied linguistics, including those who are unfamiliar with current academic study in English speaking universities, the books can act as one-step introductions. For those with more academic experience, they can also provide a way of surveying, updating and organising existing knowledge.

The view of applied linguistics in this series follows a famous definition of the field by Christopher Brumfit as:

> The theoretical and empirical investigation of real-world problems in which language is a central issue.
>
> (Brumfit 1995: 27)

In keeping with this broad problem-oriented view, the series will cover a range of topics of relevance to a variety of language related professions. While language teaching and learning rightly remain prominent and will be the central preoccupation of many readers, our conception of the discipline is by no means limited to these areas. Our view is that while each reader of the series will have their own needs, specialities and interests, there is also much to be gained from a broader view of the discipline as a whole. We believe there is much in common between all enquiries into language related problems in the real world, and much to be gained from a comparison of the insights from one area of applied linguistics with another. Our hope therefore is that readers and course designers will not choose only those volumes relating to their own particular interests, but use this series to construct a wider knowledge and understanding of the field, and the many cross-overs and resonances between its various areas. Thus the topics to be covered are wide in range, embracing an exciting mixture of established and new areas of applied linguistic enquiry.

The perspective on applied linguistics in this series

In line with this problem-oriented definition of the field, and to address the concerns of readers who are interested in how academic study can inform their own professional practice, each book follows a structure in marked contrast to the usual movement *from* theory *to* practice. In this series, this usual progression is presented back to front. The argument moves *from* Problems, *through* Intervention, and *only* finally to Theory. Thus each topic begins with a survey of everyday professional problems in the area under consideration, ones which the reader is likely to have encountered. From there it proceeds to a discussion of intervention and engagement with these problems. Only in a final section (either of the chapter or the book as a whole) does the author reflect upon the implications of this engagement for a general understanding of language, drawing out the theoretical implications. We believe this to be a truly *applied* linguistics perspective, in line with definition given above, and one in which engagement with real-world problems is the distinctive feature, and in which professional practice can both inform and draw upon academic understanding.

Support to the reader

Although it is not the intention that the text should be in anyway activity-driven, the pedagogic process is supported by measured guidance to the reader in the form of suggested activities and tasks that raise questions, prompt reflection and seek to integrate theory and practice. Each book also contains a helpful glossary of key terms.

The series complements and reflects the *Routledge Handbook of Applied Linguistics* edited by James Simpson, which conceives and categorises the scope of applied linguistics in a broadly similar way.

Ronald Carter
Guy Cook

Reference

Brumfit, C.J. (1995) 'Teacher professionalism and research' in G. Cook and B. Seidlhofer (eds) *Principle and Practice in Applied Linguistics*. Oxford: Oxford University Press, pp. 27–42.

Note

There is a section of commentaries on a number of the tasks, at the back of the book from p. 199. The TC symbol in the margin indicates that there is a commentary on that task.

Acknowledgement

Some of the work described in this book is substantially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No.: G-YG10).

Part I

Problems and practices

www.routledge.com/cw/rial

1 Introduction

This chapter briefly introduces the reader to corpus linguistics by answering two basic questions and explaining related concepts. The questions addressed are:

- What is a corpus?
- What is corpus linguistics?

What is a corpus?

A corpus is a collection of texts that has been compiled for a particular reason. In other words, a corpus is not a collection of texts regardless of the types of texts collected or, if a variety of text types (i.e., genres) are in the corpus, the relative weightings assigned to each text type. A corpus, then, is a collection of texts based on a set of design criteria, one of which is that the corpus aims to be representative. These design criteria are discussed in detail in Chapter 4, and so here we examine some of the wider issues that have to be thought about and decided upon when building a corpus. In this book, we are interested in how corpus linguists use a corpus, or more than one corpus (i.e., 'corpora'), in their research. This is not to say that only corpus linguists have corpora, or only corpus linguists use corpora in their research. Corpora have been around for a long time, but in the past they could only be searched manually, and so the fact that corpora are now machine-readable has had a tremendous impact on the field.

Corpora are becoming ever larger thanks to the ready availability of electronic texts and more powerful computing resources. For example, the Corpus of Contemporary American English (COCA) contains 410 million words (see http://corpus.byu.edu/coca/) and the British National Corpus (BNC) over 100 million words (see www.natcorp.ox.ac.uk/ or http://corpus.byu.edu.bnc/). Corpora are usually studied by means of computers, although some corpora are designed to allow users to also access individual texts for more qualitative analyses. It would be impossible to search today's large corpora manually, and so the development of fast and reliable corpus linguistic software has gone hand in hand with the growth in corpora. The software can do many things, such as generate word and phrase frequencies lists, identify words that tend to be selected with each other such as *brother* + *sister* and *black* + *white* (termed 'collocates'), and provide a variety of statistical functions that assist the user in deciphering the results of searches. You do not have to compile your own corpus. A number of corpora are available online, or commercially, with built-in software and user-friendly instructions.

Corpus linguists are researchers who derive their theories of language from, or base their theories of language on, corpus studies. As a result, one basic consideration when collecting spoken or written texts for a corpus is whether or not the texts should be naturally occurring. Most corpus linguists are only interested in corpora containing texts that have been spoken or written in real-world contexts. This, therefore, excludes contrived or fabricated texts, and texts spoken or written under experimental conditions. The reason for this preference is that corpus linguists want to describe language use and/or propose language theories that are grounded in actual language use. They see no benefit in examining invented texts or texts that have been manipulated by the researcher. Another consideration when collecting texts for a corpus is whether only complete texts should be included or if it is acceptable to include parts of texts. This can become an issue if, for example, the corpus compiler wants each text to be of equal length, which almost certainly means that some texts in the corpus are incomplete. Some argue that there are advantages when comparing texts to have them all of the same size, while others argue that cutting texts to fit a size requirement impairs their authenticity and possibly removes important elements, such as how a particular text type ends. The consensus, therefore, is to try to collect naturally occurring texts in their entirety. Another reason for carefully planning what goes into a corpus is to maintain a detailed record of each text and its context of use – when it happened, what kind of text it is, who the participants are, what the communicative purposes are and so on. This information is then available to users of the corpus, and is very useful in helping to interpret and explain the findings.

There are many different kinds of corpora. Some attempt to be representative of a language as a whole and are termed 'general corpora' or 'reference corpora', while others attempt to represent a particular kind of language use and are termed 'specialised corpora'. For example, the 100 million-word British National Corpus (BNC, see http://corpus.byu.edu/bnc/) contains a wide range of texts which the compilers took to be representative of British English generally, whereas the Michigan Corpus of Academic Spoken English (MICASE, see http://micase.elicorpora.info/) is a specialised corpus representing a particular register (spoken academic English) that can also be searched based on more specific text types (genres) such as lectures or seminars. The latter corpus is also special in the sense that it is comprised only of spoken language. Spoken language is generally massively underrepresented in corpora, a problem for those corpora that aim to represent general language use, for example. The logistics and costs of collecting and transcribing naturally occurring spoken data are the reasons for this, whereas the sheer ease and convenience of the collection of electronic written texts has led to the compilation of numerous written corpora. This imbalance needs to be borne in mind by users of corpora because what one finds in spoken and written corpora may differ in all kinds of ways.

Corpora are typically described in terms of the number of words that they contain and this raises another set of considerations because of the basic question: what is a word? When you count the number of words you have typed on your computer, the number of words is not based on the number of words, but on the number of spaces in the text and this is also how some corpus linguistic software packages arrive at the number of words in a corpus. However, what about something such as *haven't*? Should this be counted as one word or two (*have* + *n't*)? Or what about *PC* (as in 'personal computer')? Is this a word or two words or something else? All of these issues, of course, have to be resolved and made clear to the users of the corpus. The words in a corpus are often further categorised into 'types' and 'tokens'. The former comprise all of the unique word types in a corpus, excluding repetitions of the same word, and the latter are made up of all the words in a corpus, including all repetitions.

The 'type' category raises yet another issue. What constitutes a type? For example, *do*, *does*, *doing* and *did*. Each of these words share the same 'lemma' (i.e., they are all derived from the same root form: *DO*), but should they be counted as four different words (i.e., four 'types') in a word frequency list, or as one word based on the lemma and not listed separately? Most corpus linguistic software lists them as separate types. Similarly, if you search for one of these four words, do you want the search to include all the other forms as well? Some software packages allow the user to choose. Again, these are things to think about for corpus compilers, corpus linguistic software writers and corpus users. Counting words, categorising words and searching for words in a corpus all raise issues that corpus linguists have to address. An option for corpus compilers is to add additional information to the corpus, such as identifying clauses or word classes (e.g., nouns and verbs) by means of annotation (i.e., the insertion of additional information into a corpus), which enables the corpus linguistic software to find particular language features.

To summarise, a corpus is a collection of texts that has been compiled to represent a particular use of a language and it is made accessible by means of corpus linguistic software that allows the user to search for a variety of language features. The role of corpora means that corpus linguistics is evidence-based and computer-mediated. While not unique to corpus linguistics, these attributes are central to this field of study. Corpus linguistics is concerned not just with describing patterns of form, but also with how form and meaning are inseparable, and this notion is returned to throughout this book. The centrality of corpora-derived evidence is perhaps best encapsulated in the phrase 'trust the text' (see, for example, Sinclair 2004), which underscores the empirical nature of this field of language study.

What is corpus linguistics?

Corpus linguists compile and investigate corpora, and so corpus linguistics is the compilation and analysis of corpora. This all seems reasonably straightforward, but not everyone engaged in corpus linguistics would agree on whether corpus linguistics is a methodology for enhancing research into linguistic disciplines such as lexicography, lexicology, grammar, discourse and pragmatics, or whether it is more than that and is, in effect, a discipline in its own right. This debate is explored later in this book, and is covered elsewhere by, for example, Tognini-Bonelli (2001) and McEnery *et al.* (2006). The distinction is not unimportant because, as we shall see, the position one takes is likely to influence the approach adopted in a corpus linguistic study. Simply put, those who see corpus linguistics as a methodology (e.g., McEnery *et al.*, 2006, 7–11) use what is termed the 'corpus-based approach' whereby they use corpus linguistics to test existing theories or frameworks against evidence in the corpus. Those who view corpus linguistics as a discipline (e.g., Tognini-Bonelli, 2001; Biber, 2009) use the corpus as the starting point for developing theories about language, and they describe their approach as 'corpus-driven'. These approaches and their differences are examined in detail later in this book. For now, it is sufficient to understand that there is not one shared view of exactly what corpus linguistics is and what its aims are. In other words, even though the two main groupings both compile and investigate corpora, they adopt very different approaches in their studies because one sees corpus linguistics as a tool and the other as a theory of language. The author, it should be noted, subscribes to the latter view, and this will be foregrounded as the book unfolds.

As mentioned above, the fact that corpora are machine-readable opens up the possibility for users to search them for a multitude of features. The frequencies of types and tokens can be generated all but instantly, along with lists of key words. Key words are those that are either unique to, or have a higher frequency in, one corpus or text compared with a reference corpus. Lists of key words can tell us about what is termed the 'keyness' of texts and corpora (see Bondi and Scott, 2010). Also, how a word or phrase is distributed through a corpus, or each text in a corpus, can be displayed (see Scott, 2008) to enable the user to see if it is found throughout the text or corpus, or confined to a particular section. Another type of search can find sequences of words, such as *a lot of* and *there is*, which make up a sizeable amount of the language patterning, or phraseology, found in language. These sequences are variously termed 'lexical bundles', 'clusters', 'chunks' and 'n-grams'. More recently, the investigation of the extent of phraseological variation has received increased attention (see Cheng *et al.*, 2006 and Cheng *et al.*, 2009), in addition to the study of the fixed phraseologies found in n-gram searches. The ways in which words tend to collocate (form patterns of association) are a major focus of corpus linguistics, and associations based on structure and grammatical categories (termed 'colligation') are another important focus. The extent of patterning in the language is termed its phraseological tendency by Sinclair (1987), and corpus linguistics has shown how meaning is created not by meaning residing in single words, but by patterns of word co-selections. This process of meaning creation and the retrieval of the resultant phraseologies is discussed at length in this book.

In order to view the results of searches, corpus linguists have devised ways of displaying the results of searches, and the concordance is probably the best known of these. Below, a concordance for the search item *services* in a corpus of political speeches (see http://rcpce.engl.polyu.edu.hk/policy_addresses/default.htm) is shown (Figure 1.1).

A concordance is a display of all of the search items in a corpus and is usually presented on the computer screen in the KWIC format (i.e., Key Word in Context), which centres the search item and provides its immediate co-text to the left and right. Figure 1.1 is not the full concordance, but is a sample to illustrate what a concordance looks like. Each concordance line has been sorted based on the first word to the left of *services*, which means that the concordance is displayed based on the alphabetical order of the first word to the left. In the case illustrated here, such a search option allows the user to easily identify the words that are used to modify the search item *services* (the software used here is ConcGram 1.0, written by Greaves, 2009). It is also possible to sort to the right of the search item and to sort based on the second, third, fourth word and so on. Thus, the ways in which search outputs can be configured and displayed play an important role in helping the user to investigate the corpus. The inclusion in most corpus linguistic software packages of statistical

```
1  re on medical and health services from the present 15% to
2  r upgrade our healthcare services to benefit the community
3  nd efficient immigration services are essential. Hong Kong
4  e, a hub for information services and logistics and a prem
5   provision of integrated services will also strengthen Hong
6   and comprehensive legal services for dispute resolution ar
7 to support school library services and the Chinese and Engl
8  development of mediation services. On many occasions, inter
9  ct to strengthen medical services for residents. We expect
10  e delivery of municipal services. More importantly, it wil
```

Figure 1.1 Sample concordance lines for *services* sorted one word to the left

measures is another example of how the functionality available to the user can extend the kinds of study undertaken, and these measures are summarised later in the book.

At its core, language is all about creating meaning, and so uncovering how this is realised through the co-selections of words and structures by speakers and writers is of central interest in corpus linguistics. The evidence coming out of corpus linguistics has contributed hugely to a better understanding of both the extent of language patterning and the ways in which these patterns are the product of the co-selections made to create meanings. The forms of patterning, and the kinds of co-selections to be found by means of corpus linguistics, are a major focus in this book, along with lots of opportunities for the reader to practise finding these and other language features across a wide range of corpora in the many activities provided in each chapter.

Aims and structure of the book

This book concentrates on aspects of corpus linguistics deemed to be both important and valuable to learners, teachers and researchers who wish to become competent and reflective language users and researchers. It discusses, with many illustrative examples, a wide range of questions that corpus linguistics is able to answer more accurately and effectively, and other questions that only corpus linguistics can answer. Findings from corpus linguistics have challenged many established assumptions in linguistic research. Examples of such findings are that different word forms of a lemma often have different patterns of meaning, that collocation is a good guide to meaning and that combinations of words generate context-specific meanings.

The book adopts an inductive approach in both design and structure. It is organised into three parts: it begins by identifying some major issues and problems in the study of language structure and use in various contexts of communication (Part I), followed by interventions that suggest and exemplify the use of a wide range of language resources and corpus linguistic methods to address the issues and deal with the problems (Part II), and, finally, a systematic description of some major concepts and models of corpus linguistics that underpin research studies in corpus linguistics (Part III). In addition, in both Parts I and II in Chapters 2–5, a large number of activities are included for readers to consolidate their knowledge of corpus linguistics and to try out for themselves different corpus linguistic search methods and functions.

Part I begins with an overview of corpus linguistics and then outlines a number of linguistic inquiries in both speaking and writing in major fields of linguistics and applied linguistics, including lexis, grammar, register, conversation analysis, genre analysis, pragmatics and discourse intonation, across a number of domains of language use and communicative contexts of situations such as academic, business, social and professional contexts. Part I aims to raise the awareness of the reader by highlighting the wide-ranging inquiries that have examined language structure and use, as well as possible and alternative ways of conducting linguistic inquiries.

Part II describes the rationale for studying corpus linguistics as a discipline and using corpus linguistics as a method of linguistic inquiry. It describes different types of corpora and their specific uses, the mechanics of corpus design and construction, and various corpus applications. It also describes basic corpus search functions, major functions in corpus linguistic software and methods of analysis used in corpus linguistics. Building on Part I, Part II describes and exemplifies a wide range of corpus studies in linguistics in areas that have been introduced in Part I, with specific focuses on major contributions made by corpus linguistics in linguistic inquiries, particularly lexical phraseology and the notion of the lexical item. Part II ends with a description of the rationale, procedure and assessment of a corpus-driven language project for those new to corpus linguistics. The project advocates data-driven learning (Johns, 1991a, 1991b) that encourages the learner to take on the role of the researcher and, at the same time, aims to enhance the learner's problem-solving, critical analysis and independent learning capabilities.

Part III returns to the main questions, features and phenomena relating to various linguistic inquiries raised in Part I, and then addressed by means of corpus linguistic approaches and methods in Part II, by describing some of the main concepts and models that underpin corpus linguistic research and illustrating these concepts and models with some significant research findings from corpus linguistic studies across a range of linguistic fields.

This book, through its innovative three-part 'problems-interventions-theories' inductive structure and the large number of guided practical tasks, coupled with detailed commentaries, hopes to dispel any apprehensions that some students, teachers and researchers might have about corpus linguistics. It is also hoped that the best possible arrangements can be made to introduce corpus linguistics to learners, whether as a subject on its own or incorporated into other linguistics subjects, with due consideration being given to the availability of computer and language resources, teacher training, curriculum space and the computer literacy of the learners.