CORPUS LINGUISTICS

READINGS IN A WIDENING DISCIPLINE

Edited by
Geoffrey Sampson and
Diana McCarthy

continuum
LONDON • NEW YORK

Continuum
The Tower Building, 11 York Road, London SE1 7NX
15 East 26th Street, New York, NY 10010

First published 2004 by Continuum.

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

ISBN 0–8264–6013–5 (hardback)

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

Typeset by RefineCatch Limited, Bungay, Suffolk
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wilts

SOURCES AND ACKNOWLEDGEMENTS

Our first debts of gratitude in connection with this book are to Janet Joyce, formerly commissioning editor at Continuum, for her enthusiasm, and to colleagues in the international academic community who have encouraged us to take the project forward and have helped, for instance by suggesting their personal favourite items. Many authors represented in the anthology offered cooperation going well beyond a simple reprint permission (we particularly thank Geoffrey Leech for comments on a draft of our introductory chapter); and we are also very grateful to Peter H. Fries (Mount Pleasant, Michigan), Sylviane Granger (Louvain-la-Neuve), Graeme Kennedy (Wellington, New Zealand), David Lee (Ann Arbor, Michigan), Oliver Mason (Birmingham), Vladimir Petkevič (Prague), Diana Santos (Oslo), Bilge Say (Ankara), Harold Somers (Manchester), and our Sussex colleague Desmond Watson. We apologize to anyone whose name is inadvertently omitted; and we of course take responsibility for any errors or shortcomings.

We are grateful to Michael Cotterell of Mouse Nous for producing the graphics.

In the following list, we identify the origins of successive chapters and acknowledge the respective copyright holders' permission to reprint.

Chapter 2 is excerpted from Charles Carpenter Fries, *The Structure of English: an Introduction to the Construction of English Sentences*, Harcourt Brace, 1952, and is reprinted by permission of Peter H. Fries.

Chapter 3 appeared in *College English*, vol. 26, pp. 267–73, 1965, and is © 1965 by the National Council of Teachers of English; it is reprinted by permission. The information in table 3.2 has been reorganized into a more conventional format.

Chapter 4 appeared in *Lingua*, vol. 26, pp. 281–93, 1971, and is reprinted by permission of Elsevier Science. Various numerical data have been corrected by the author.

Chapter 5 was originally presented at the 7th International Conference on English Language Research on Computerized Corpora, Amsterdam, 1986; the version reproduced here appeared in Jan Svartvik, ed., *The London–Lund Corpus of Spoken English: Description and Research*, Lund University Press, 1990, and is reprinted with the author's permission.

Chapter 6 was presented at the 2nd International Congress of the European Association for Lexicography, Zürich, 1986, and published in its Proceedings; it is reprinted with the author's permission.

An early version of chapter 7 was presented at the 8th International Conference on English Language Research on Computerized Corpora, Hanasaari, Finland, 1987. The version used here appeared in Thomas J. Walsh, ed., *Synchronic and Diachronic Approaches to Linguistic Variation*

and Change (Georgetown University Round Table on Languages and Linguistics 1988), Georgetown University Press, 1989, and is reprinted by permission of Georgetown University Press.

Chapter 8 is based on material first presented to the Council of Europe in 1987; the version reprinted here is excerpted from John Sinclair, *Corpus, Concordance, Collocation*, Oxford University Press, 1991, and reprinted with the author's permission.

Chapter 9 appeared in the *ICAME Journal*, vol. 11, pp. 5–17, 1987, and is reprinted by permission of the Norwegian Computing Centre for the Humanities.

The first version of chapter 10 was AT&T Bell Laboratories Statistical Research Report no. 90, 1989. The present version was published in Nelleke Oostdijk and Pieter de Haan, eds, *Corpus-Based Research into Language: In Honour of Jan Aarts*, Rodopi, 1994; it is reprinted by permission of Editions Rodopi B.V.

Chapter 11 appeared in *Computational Linguistics*, vol. 16, pp. 79–85, 1990, and is © the Association for Computational Linguistics and MIT Press; it is reprinted by permission of the ACL.

Chapter 12 appeared in Karin Aijmer and Bengt Altenberg, eds, *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Longman, 1991, and is © Longman Group UK Limited 1992; it is reprinted by permission of Pearson Education Limited.

Chapter 13 was presented at Nobel Symposium 82, Stockholm, 1991, and published in Jan Svartvik, ed., *Directions in Corpus Linguistics*, Mouton de Gruyter, 1992; it is reprinted by permission of Mouton de Gruyter.

Chapter 14 appeared in Karin Aijmer and Bengt Altenberg, eds, *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Longman, 1991, and is © Longman Group UK Limited 1992; it is reprinted by permission of Pearson Education Limited.

Chapter 15 was presented at the 13th International Conference on English Language Research on Computerized Corpora, Nijmegen, 1992, and published in Jan Aarts, Pieter de Haan, and Nelleke Oostdijk, eds, *English Language Corpora: Design, Analysis and Exploitation*, Rodopi, 1993; it is reprinted by permission of Editions Rodopi B.V.

Chapter 16 appeared in the *ICAME Journal*, vol. 16, pp. 29–50, 1992, and is reprinted by permission of the Norwegian Computing Centre for the Humanities. Our reprint has eliminated a confusion in the original about numbering of notes.

A version of chapter 17 was distributed for the Pisa Workshop on Textual Corpora, 1992. The present version appeared in *Literary and Linguistic Computing*, vol. 8, pp. 243–57, 1993, and is reprinted by permission of the Association for Literary and Linguistic Computing.

Chapter 18 appeared in Mona Baker, Gill Francis, and Elena Tognini-Bonelli, eds, *Text and Technology: in Honour of John Sinclair*, John Benjamins, 1993, and is reprinted by permission of John Benjamins Publishing Company.

Chapter 19 appeared in *Computational Linguistics*, vol. 19, pp. 103–20, 1993, and is © the Association for Computational Linguistics and MIT Press; it is reprinted by permission of the ACL.

Chapter 20 appeared in Mona Baker, Gill Francis, and Elena Tognini-Bonelli, eds, *Text and Technology: in Honour of John Sinclair*, John Benjamins, 1993, and is reprinted by permission of John Benjamins Publishing Company. A long Appendix to the original article, and a paragraph referring to it in the main text, are omitted from this reprint.

The poems *Days* and *First Sight* are reprinted from *Collected Poems* by Philip Larkin. Copyright © 1988, 1989 by the Estate of Philip Larkin. Reprinted by permission of Faber and Faber Ltd and of Farrar, Straus & Giroux LLC.

Chapter 21 appeared in *Computational Linguistics*, vol. 19, pp. 313–30, 1993, and is © the Association for Computational Linguistics and MIT Press; it is reprinted by permission of the ACL.

Chapter 22 was presented at the 1st International Conference on Teaching and Language

Corpora, Lancaster, 1994, and published in its Proceedings; it is reprinted with the authors' permission. The present version incorporates passages redrafted by agreement between authors and editors.

Chapter 23 was presented at the 4th International Workshop on Parsing Technologies, Prague and Karlovy Vary, 1995, and published in its Proceedings; it is reprinted by permission of the Caroline University, Prague. A brief remark is added from the version reprinted in Jenny Thomas and M.H. Short, eds, *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*, Longman, 1996, and some rephrasing has been agreed between authors and editors.

Chapter 24 appeared in Sidney Greenbaum, ed., *Comparing English Worldwide: The International Corpus of English*, Clarendon Press, Oxford, 1996, and is reprinted by permission of Oxford University Press.

Chapter 25 appeared as research report CS–96–02 of the Brown University Department of Computer Science, 1996, and is reprinted with the author's permission. Some rephrasing has been adopted from the report of a revised version of the experiment in the Proceedings of the 13th American Association for Artificial Intelligence Annual Conference, Portland, Oregon.

Chapter 26 appeared in Jenny Thomas and M.H. Short, eds, *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*, Longman, 1996, and is © Addison Wesley Longman Limited 1996; it is reprinted by permission of Pearson Education Limited.

Chapter 27 appeared as ILLC Research Report LP–96–13, Institute for Logic, Language and Computation, University of Amsterdam, 1996, and is reprinted with the authors' permission.

Chapter 28 was presented at the 17th International Conference on English Language Research on Computerized Corpora, Stockholm, 1996, and published in M. Ljung, ed., *Corpus-Based Studies in English*, Rodopi, 1997; it is reprinted by permission of Editions Rodopi B.V.

Chapter 29 appeared in *Computational Linguistics*, vol. 22, pp. 249–54, 1996, and is © the Association for Computational Linguistics and MIT Press; it is reprinted by permission of the ACL.

Chapter 30 appeared in Susan C. Herring, ed., *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*, John Benjamins, 1996, and is reprinted by permission of John Benjamins Publishing Company. Some complexities relating to choice and numbering of examples in the original printing have been altered by agreement between author and editors.

An initial version of chapter 31 was presented at an ACL SIGLEX workshop on Semantic Tagging held in conjunction with the Applied Natural Language Processing conference in Washington, DC, 1997; the revised and extended version included here appeared in *Natural Language Engineering*, vol. 5, pp. 113–33, 1999, and is reprinted by permission of Cambridge University Press.

Chapter 32 appeared in the *Journal of Second Language Writing*, vol. 6, pp. 183–205, 1997, and is reprinted by permission of Elsevier Science.

Chapter 33 was presented at the American Association for Artificial Intelligence Spring Symposium on applying machine learning to discourse processing, Stanford, California, 1998, and published in its Proceedings; it is reprinted with the author's permission.

Chapter 34 was presented at the 19th International Conference on English Language Research on Computerized Corpora, Newcastle, Co. Down, Northern Ireland, 1998, and published in John M. Kirk, ed., *Corpora Galore*, Rodopi, 2000; it is reprinted by permission of Editions Rodopi B.V.

A version of chapter 35 appeared as Research Paper HCRC/RP–95 of the Human Communication Research Centre, University of Edinburgh, 1998, and is reprinted with the

author's permission; the present version incorporates passages redrafted by agreement between author and editors.

Chapter 36 was presented at the 1st International Conference on Language Resources and Evaluation, Granada, 1998, and published in its Proceedings; it is reprinted with the author's permission.

Chapter 37 was presented at the 'Journées *ATALA* sur les corpus annotés pour la syntaxe' workshop, Paris, 1999, and published in its Proceedings; it is reprinted with the authors' permission.

A version of chapter 38 was the keynote address at the 'Journées *ATALA* sur les corpus annotés pour la syntaxe' workshop, Paris, 1999. The version printed here was presented at the Corpus Linguistics 2001 conference, Lancaster.

Chapter 39 was published on the website of the School of Cognitive Science, University of Edinburgh, and is printed with the authors' permission.

Chapter 40 appeared in *ELSNews*, vol. 10, Spring 2001, pp. 9–10, and is reprinted by permission of ELSNET, the European Network in Human Language Technologies.

A version of chapter 41 was presented at the Corpus Linguistics 2001 conference, Lancaster, and published in its Proceedings; the version included here has been revised by the authors and is printed with their permission.

A version of chapter 42 was presented at the Corpus Linguistics 2001 conference, Lancaster, and published in its Proceedings; the version included here has been revised by the author and is printed with his permission.

Chapter 43 was presented at the Speech Prosody 2002 Conference, Aix-en-Provence, and is published in its Proceedings; it is reprinted with the authors' permission. Some numerical data have been corrected by the authors.

ABBREVIATIONS USED IN THIS BOOK

The following list includes widely used abbreviations and symbols with special relevance to corpus linguistics. Occasional uses of more local abbreviations, and generally recognized acronyms such as IBM, are not listed.

A	adverbial
ACL	Association for Computational Linguistics
ADD	ATR Dialogue Database
ADJP	adjective phrase
ADVP	adverb phrase
AFOSR	(US) Air Force Office of Scientific Research
ANLT	Alvey Natural Language Tools
AP	Associated Press, or adjective phrase
APB	average parse base
ARO	(US) Army Research Office
ARPA	Advanced Research Projects Agency
ASCII	American Standard Code for Information Interchange
ASR	automatic speech recognition
ATALA	Association pour la Traitement Automatique des Langues
ATIS	Air Travel Information System
ATS	analytic tree structure
bboard	bulletin board
BNC	British National Corpus
BoE	Bank of English
Brown	Brown University Corpus
C	complement, or consonant
CA	conversation analysis
CAI	computer-aided instruction
CCITT	International Telephone and Telegraph Consultative Committee
CDIF	Corpus Document Interchange Format
CEC	Corpus of English Conversation
CED	*Collins English Dictionary*
CF, CFG	context-free, context-free grammar
CLAWS	Constituent-Likelihood Automatic Word-tagging System
CNC	Czech National Corpus

CNRS	Centre National de la Recherche Scientifique
COBUILD	Collins Birmingham University International Language Database
COCOSDA	International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques
COLT	Bergen Corpus of London Teenage Language
CPU	central processing unit
CULD	*Chambers Universal Learners' Dictionary*
d.f.	degrees of freedom
DAMSL	Dialogue Act Markup in Several Layers
DARPA	Defense Advanced Research Projects Agency (at some periods called ARPA)
DCG	definite clause grammar
DGA	Délégation Générale pour l'Armement
DOP	data-oriented parsing/processing
DRI	Discourse Research Initiative
DTD	document type definition
EAGLES	Expert Advisory Group on Language Engineering Standards
EFL	English as a Foreign Language
EM	expectation maximization
EPSRC	Engineering and Physical Sciences Research Council (UK)
ESCA	European Speech Communication Association (now ISCA)
ESL	English as a Second Language
ESRC	Economic and Social Research Council (UK)
$f(x)$	frequency of x
F_0	fundamental frequency
FLOB, Frown	1990s corpora parallel to LOB and Brown, compiled at Freiburg-im-Breisgau
GATE	General Architecture for Text Engineering
GCE	General Certificate of Education
GEIG	Grammar Evaluation Interest Group
HMM	Hidden Markov Model
HTML	Hypertext Markup Language
$I(x, y)$	mutual information between x and y
ICAME	International Computer Archive of Modern (later, Modern and Mediaeval) English
ICE	International Corpus of English
ICLE	International Corpus of Learner English
IE	Indo-European
IEA	International Association for the Evaluation of Educational Achievement
iff	if and only if
IP	intonation phrase
IRC	Internet Relay Chat
IRV	International Reference Version
ISO	International Organization for Standardization
ISP	Internet service provider
ITU	International Telecommunication Union
IViE	Intonation Variation in English
KWIC	key word in context
L1, L2	first language, second language
LA	lexical association
LALR	lookahead LR
LCA	Lancaster Corpus of Abuse

LDC	Linguistic Data Consortium
LDEI	*Longman Dictionary of English Idioms*
LDOCE	*Longman Dictionary of Contemporary English*
LFG	lexical functional grammar
LL	London–Lund Corpus
LOB	Lancaster–Oslo/Bergen Corpus
LR	left-to-right scan
LT NSL	Language Technology group Normalized SGML Library
MAP	maximum a posteriori
MATE	Multilevel Annotation, Tools Engineering
MBE	modern British English
MLE	maximum likelihood estimator
MOO	MUD, object oriented
MOS	mean opinion score
MUC	Message Understanding Conference
MUD	multiple user dungeons/dialogue
N	noun
N	sample size
$\overline{\text{N}}$, NBAR	(in X-bar syntax) a tagma containing a noun and complement
n-gram	word-sequence of length n
NI	non-Indo-European
NIML	non-indigenous minority languge
NLP	natural language processing
NNS	non-native-speaker
NP	noun phrase
NP-hard	nondeterministic polynomial-time hard
NS	native-speaker
NSF	National Science Foundation
O	object
OALD	*Oxford Advanced Learners' Dictionary of Current English*
ODCIE	*Oxford Dictionary of Current Idiomatic English*
OED	*Oxford English Dictionary*
OUCS	Oxford University Computing Services
$P(x)$	probability of x
$P(x \mid y)$	probability of x given y
PCFG	probabilistic context-free grammar
PDT	Prague Dependency Treebank
PoS	part of speech
PP	prepositional phrase
RMS	root mean square
RP	Received Pronunciation
S	clause (sentence), or subject
s.v.	under the word
SAG	Svenska Akademiens grammatik (Swedish Academy Grammar)
SAM	Speech Assessment Methods
SBAR	(in X-bar syntax) a tagma containing a complementizer and an S
SCFG	stochastic context-free grammar (synonym for PCFG)
SEC	Spoken English Corpus
SED	Survey of English Dialects
SGML	Standard Generalized Markup Language

SIG	Special Interest Group
STSG	stochastic tree-substitution grammar
SUSANNE	Surface and Underlying Structural Analyses of Natural English
t-test	test using the Student's *t* statistic
TEI	Text Encoding Initiative
TESS	Text Segmentation for Speech
TFA	topic-focus articulation
TGTS	tectogrammatical tree structure
TIMIT	Texas Instruments–MIT corpus of read speech
ToBI	Tones and Break Indices
TOEFL	Test of English as a Foreign Language
TREC	Text Retrieval Conference
TTS	text-to-speech
UCREL	Unit for Computer Research on the English Language, Lancaster
URL	universal resource locator
V	verb, or vowel
VB	verb
VP	verb phrase
W3C	World Wide Web Consortium
WH-item	interrogative or relative word beginning *wh-*
WHIZ deletion	deletion of relative pronoun followed by *is, are*, etc.
WSD	word sense disambiguation
WSJ	*Wall Street Journal*
XML	Extensible Markup Language
XSL	Extensible Stylesheet Language
XSLF	XSL Formatting
XSLT	XSL Transformations

CONTENTS

1

INTRODUCTION

1 Background

A *corpus*, for people who study language and languages, is a collection of specimens of a language as used in real life, in speech or writing, selected as a sizeable 'fair sample' of the language as a whole or of some linguistic genre, and hence as a useful source of evidence for research on the language. Corpus linguistics is the kind of research, carried out in university departments of linguistics, computer science, and related subjects (and nowadays often in industrial research labs too), which makes crucial use of language corpora.

(The word *corpus* is Latin for 'body', used here as in 'body of evidence'. In Latin the plural of *corpus* is *corpora*; most corpus linguists, ourselves included, use that as the English plural also. It is quite permissible to Anglicize the plural and write *corpuses* – some corpus linguists use that form: we prefer *corpora* because *corpuses* sounds like 'corpses'.)

Naturally, people have studied languages via corpora for a long time. More than a century ago, Wilhelm Kaeding and an army of helpers processed a corpus of almost eleven million words of German by hand to extract statistics for use in improving shorthand systems (Hausser 1998). In the eighteenth century, Dr Johnson based his famous English dictionary in part on a collection of over 150,000 quotations chosen to illustrate the use of words by reputable authors – a collection of that kind is too selective to be a typical language corpus in the modern sense, but it was certainly a corpus of a sort. There were earlier scholars whose work might arguably be included under the same heading. (W.N. Francis 1992 gives a historical survey.)

But modern corpus linguistics depends very heavily on the computer. Only electronic processing allows one to search for some form or structure of interest in a large collection of language samples with confidence that one has extracted *all* relevant instances, rather than just picking out a subset that strike one's eye. For that matter, only computing technology allows one to make multiple copies of a large standard corpus and distribute them to separate research sites. Corpora compiled by linguists in the decades immediately 'B.C.' (before computers), notably the Survey of English Usage at University College London, existed in unique hard copies only. Scholars wishing to exploit the information in the Survey of English Usage had to travel to London to work with its banks of typewritten paper slips (until the material was eventually computerized). Because the Survey coverage, particularly of spontaneous speech, was hard to match elsewhere, many scholars did make such trips, but obviously this mode of data dissemination was extremely restrictive by comparison with modern electronic techniques.

Consequently 'corpus linguistics' nowadays is usually understood to mean 'electronic corpus linguistics'. For the sake of historical background our selection of readings in this volume

does include two examples of corpus linguistics 'B.C.', but otherwise they are all drawn from the period since computers became available in practice to academic researchers on human language. For the more favourably placed, this period began some time in the mid-1960s; by about 1980, any linguist who wanted to use a computer could easily get access to one.

(It is because of the central role of computing in modern corpus linguistics that we were careful, in the preceding paragraph, to refer to the topic of corpus linguistics as '*human* language'. To computer scientists, 'language' refers by default to programming languages like C or Java, and languages like English or Chinese are distinguished as 'human languages' or 'natural languages'. But we shall be discussing only human languages, not programming languages, so from now on we shall call the former simply 'languages'.)

2 *Language engineering and lexicography*

Widening access to computers, and growth in computer power, during the latter half of the twentieth century did not only make it easier for humanistic scholars to use language corpora in order to pursue their existing interests. It also called into being a new, technical field often called 'language engineering', which aims to develop computer systems that execute practical, economically useful tasks related to language, and which depends heavily on analysis of language corpora.

This contrast between humanities-based linguistic scholarship and economically useful language engineering is not intended to suggest that humanistic corpus researchers have no practical concerns. Often they have. For instance, much of the impetus for creation and analysis of English-language corpora, on the European side of the Atlantic, has come from Continental experts on teaching English as a foreign language, who need precise information about how the English language is being used at the present time. But 'language engineering' (Cunningham 1999) refers to development of sophisticated software allowing computers themselves to execute language-related tasks, as opposed to the use of computers simply to register and sort language data for examination by human researchers.

One language-engineering application which has been in the public eye in recent years, for instance, is automatic speech recognition – systems, like Dragon Naturally Speaking and IBM's ViaVoice, that translate a user's spoken dictation into written words. The relationship between acoustic signals and the words they represent is so loose in practice that speech-recognition systems cannot succeed merely by analysing the physical sound waves; they also need language models telling them which sequences of words are plausible and which are not, so that tentative identifications of some words can be used by the computer to make inferences about neighbouring words. Language models in this sense are created by processing large corpora.

In the early years of computers, the public thought of these machines as intended exclusively for numerical calculations in scientific and commercial environments, but in reality language-related applications had been in the minds of their inventors from the start. Within weeks of the world's first run of a stored-program digital computer (which occurred in Manchester on 21 June 1948), Alan Turing drew up a memo (quoted in Hodges 1985: 382–3) listing potential uses for the novel machine, including for instance automatic translation between human languages.

Quite a lot of work was in fact done on developing machine translation systems in the 1950s and 1960s, largely motivated by Cold War anxieties and focused on the language-pair Russian to English. But computer systems in the early decades were very limited in storage and processing capacity, which made most corpus-based techniques difficult or impossible. Relative to typical scientific or commercial applications, language processing makes heavy demands on computer capacity. The 'atoms' of a language are its words, but rather than the hundred-odd types of atom recognized by a chemist, even a small English dictionary will list many tens of thousands of word-types, each having its own special properties. The smallest corpora of raw,

unannotated language samples which have been widely used each contain about a million word-tokens. (The valuable distinction between *type* and *token* was defined by C.S. Peirce; thus, in the Gertrude Stein quotation *a rose is a rose is a rose*, there are eight word-tokens but only three word-types, namely *a, rose*, and *is*.) It was not until the 1980s that computer power evolved to the point where figures like these ceased to look daunting.

Since then, though, there has been an explosive growth of interest in corpus-based language-engineering techniques, for machine translation and many other applications. Society has come to appreciate the importance of making computers interact with people in modes that are natural for people, and the favourite communication medium for our species is obviously language. Language corpora have become a crucial resource for developing and testing many different aspects of natural language processing ('NLP') technology. (A leading recent textbook on this field is Jurafsky and Martin 2000.)[1]

Another use of corpora which is practical rather than (or as well as) scholarly is lexicography. Compiling dictionaries may be a specialized activity, but since Dr Johnson's day dictionary publishing has become quite a large-scale business. In the case of the English language, markets include not only English-speaking countries but the vast numbers of people elsewhere in the world who want to learn and use English as a second language. Dictionary publishers aim to keep their dictionaries up to date by tracking new words and phrases and the changing usage of existing words: so they need access to large samples of real-life usage.

The early, million-word electronic corpora were not very interesting from a lexicographic viewpoint (except for 'grammatical words' like *if* or *should*, vocabulary items typically do not recur often enough in a million-word sample to yield a representative picture of their usage). But dictionary publishers were leading players in the more recent development of much larger corpora. Of the four leading British dictionary publishers, Collins collaborated with Birmingham University to produce an electronic 'Bank of English' [boe][2] whose size at different stages has varied from tens to hundreds of millions of words, while Oxford University Press, Longman, and Chambers collaborated with Lancaster and Oxford Universities and the British Library to compile the carefully balanced 100,000,000-word British National Corpus, published in 1995 ([bnc]: see chapter 15 below), which is currently the most important single corpus resource for the English language.

3 Generative versus corpus-based linguistics

Although the initial impetus to create corpora stemmed from linguistics as a humanities discipline, pure academic linguistics was surprisingly slow to exploit corpus data. The first electronic corpus of English (the 'Brown Corpus', or more formally the 'Brown University Standard Corpus of Present-Day American English' [bro]) was put into circulation in 1964, and many corpus-research techniques relevant for pure linguistics require only simple processing methods that were within the capabilities of computers even at that early period. But the 1960s and 1970s were a time when academic linguistics, particularly but by no means only in the USA, was heavily influenced by the intuition-based 'generative' theory. Many American and other linguists believed that it was not necessary to test one's language descriptions against laboriously-gathered electronic samples of usage, because it was far easier just to ask a speaker of the language what he would or would not say; if a linguist was describing his own language, he could simply consult his personal intuitions – external evidence of any kind was irrelevant. Noam Chomsky's doctrine of 'competence' and 'performance' (Chomsky 1965: 4) suggested that linguistic intuitions were not just adequate as the basis for linguistic theorizing but, in a sense, were *better* evidence than samples of real-life usage would be.

Generative linguists were sometimes remarkably blunt in rejecting the value of corpus work. Douglas Biber and Edward Finegan (1991: 204) quote a conversation from the early 1960s

between the generative linguist R.B. Lees and W. Nelson Francis, one of the two creators of the Brown Corpus mentioned above:

> Lees asked Francis what he was up to at the time, and Francis replied that he had a grant to compile a computerized corpus of English. When Lees asked 'Why in the world are you doing that?', Francis answered that he wanted to uncover the 'true facts of English grammar'. As Francis recalls the incident, Lees then looked at him 'in amazement' and exclaimed: 'That is a complete waste of your time and the government's money. You are a native speaker of English; in ten minutes you can produce more illustrations of any point in English grammar than you will find in many millions of words of random text'.

Attitudes like Lees's persisted for a long time. Twenty years on, Jan Aarts and Theo van den Heuvel (1985: 303–5) analysed the hostility still felt by many linguists toward corpus evidence, which was 'stigmatized as "degenerate" '. In some quarters this view can be encountered even today. But it is difficult to agree that intuitive data are reliable enough to base a scientific subject on. The truth is that speakers' 'intuitions' about their native language are influenced in real life by many factors apart from the properties which that language actually has; sometimes their intuitions are just wrong. Adults' ideas about their language are moulded, for instance, by what they were taught about language at school, and by their awareness of other languages or other, perhaps higher-prestige dialects of their own language. William Labov (1975: 106–7) has an anecdote about a speaker of a regional dialect of American English who assured Labov's researchers that he had never heard a special usage which is characteristic of that dialect, and had no idea what it meant – but was then overheard using it spontaneously in the way which is normal for his region. Geoffrey Pullum and Barbara Scholz (2002) have documented the way in which far-reaching claims by generative linguists about language structure being innate in the human mind have depended to a large extent on those linguists' intuitive beliefs that certain grammatical patterns are never used, although empirical evidence shows that the patterns in question are actually used quite heavily.

The implications of this situation have increasingly been accepted by academic linguists, so that in this discipline too, despite the slow start, corpus-based research is coming to be the usual thing. In the USA this development will be given a large boost when the currently planned American National Corpus [anc] is completed, enabling US linguists routinely to test their ideas against tens of millions of words of written and spoken American English, as British linguists have been doing with their national variety of the language since publication of the BNC in 1995.

4 Aims of this volume

The consequence of these various developments is that many people have been drawn in recent years into one branch or another of corpus linguistics, without having much prior knowledge of where the subject as a whole has come from or the range of directions in which it is developing. Because corpus linguistics has grown rapidly from small beginnings, important publications of a few years back sometimes appeared in low-circulation journals or hard-to-get-hold-of conference proceedings, making it difficult to develop background knowledge. One of our aims in assembling this collection is to give newcomers to corpus linguistics a handle on the domain which they have entered, introducing readers with a humanities background to basic technical aspects of the subject, and readers who are chiefly computing specialists to the arts and social-science aspects which have been a principal motive for the creation of language corpora. We hope that readers will come away from the volume with a sense of what has been attempted in corpus linguistics (and what has not yet been tried), what areas of knowledge are seen as prerequisites and what resources are available to the researcher,