EDITED BY

JOHN R. TAYLOR

The Oxford Handbook *of* THE WORD

THE OXFORD HANDBOOK OF

THE WORD

OXFORD HANDBOOKS IN LINGUISTICS

Recently published

THE OXFORD HANDBOOK OF THE HISTORY OF ENGLISH
Edited by Terttu Nevalainen and Elizabeth Closs Traugott

THE OXFORD HANDBOOK OF SOCIOLINGUISTICS
Edited by Robert Bayley, Richard Cameron, and Ceil Lucas

THE OXFORD HANDBOOK OF JAPANESE LINGUISTICS
Edited by Shigeru Miyagawa and Mamoru Saito

THE OXFORD HANDBOOK OF THE HISTORY OF LINGUISTICS
Edited by Keith Allan

THE OXFORD HANDBOOK OF LINGUISTIC TYPOLOGY
Edited by Jae Jung Song

THE OXFORD HANDBOOK OF CONSTRUCTION GRAMMAR
Edited by Thomas Hoffman and Graeme Trousdale

THE OXFORD HANDBOOK OF LANGUAGE EVOLUTION
Edited by Maggie Tallerman and Kathleen Gibson

THE OXFORD HANDBOOK OF ARABIC LINGUISTICS
Edited by Jonathan Owens

THE OXFORD HANDBOOK OF CORPUS PHONOLOGY
Edited by Jacques Durand, Ulrike Gut, and Gjert Kristoffersen

THE OXFORD HANDBOOK OF LINGUISTIC FIELDWORK
Edited by Nicholas Thieberger

THE OXFORD HANDBOOK OF DERIVATIONAL MORPHOLOGY
Edited by Rochelle Lieber and Pavol Štekauer

THE OXFORD HANDBOOK OF HISTORICAL PHONOLOGY
Edited by Patrick Honeybone and Joseph Salmons

THE OXFORD HANDBOOK OF LINGUISTIC ANALYSIS
Second Edition
Edited by Bernd Heine and Heiko Narrog

THE OXFORD HANDBOOK OF THE WORD
Edited by John R. Taylor

THE OXFORD HANDBOOK OF INFLECTION
Edited by Matthew Baerman

For a complete list of Oxford Handbooks in Linguistics please see pp. 865–866.

THE OXFORD HANDBOOK OF

THE WORD

Edited by

JOHN R. TAYLOR

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First edition published in 2015

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2015932159

ISBN 978–0–19–964160–4

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Contents

PART II WORDS AND LINGUISTIC THEORY

PART III MEANINGS, REFERENTS, AND CONCEPTS

PART IV WORDS IN TIME AND SPACE

PART V WORDS IN THE MIND

PART VI WORDS IN ACQUISITION AND LEARNING

PART VII NAMES

PART VIII FUN WITH WORDS

A FINAL WORD

List of Abbreviations

A	Agent
ABS	Absolutive case
ACC	Accusative case
ADJ	Adjective
ADV	Adverb
AFF	Affix
ANIM	Animate
ART	Article
ASP	Aspectual marker
ASS	Associative case
C	Consonant
CAUS	Causative
COND	Conditional
DEF	Definite article
DEM	Demonstrative
DET	Determiner
EMPH	Emphasis
EP	Epenthetic
ERG	Ergative case
ESS	Essive case
F	Feminine
FACT	Factive
GEN	Genitive case
HAB	Habitual
IDPH	Ideophone
IMPF	Imperfect
IND	Indicative
M	Masculine (gender)
MA	Marker
MEDIOPASS	Mediopassive
N	Noun
NCM	Noun class marker
NEG	Negative
NEUT	Neuter

NF	Non-finite
NM	Nominalizer
NP	Noun phrase
NPST	Non-past
O	Object
P	Patient argument
PASS	Passive
PC	Predicational constituent
PERF	Perfect
Pl	Plural
PLUPF	Pluperfect
POSS	Possessive
PP	Prepositional phrase
PRES	Present
PRO	Pronoun
PRT	Particle
PST	Past tense
PTC	Participle
PUNC	Punctual
PX	Possessive suffix
REMP	Remote past
REP	Reported
S	Subject
SBJV	Subjunctive
Sg	Singular
SRFL	Semireflexive
ST	Stative
SUFF	Suffix
TNS	Tense
TR	Transitive
V	Verb; Vowel
1	First person
2	Second person
3	Third person

List of Contributors

Marc Alexander is Senior Lecturer in Semantics and Lexicology at the University of Glasgow, and his work primarily focuses on digital humanities and the study of meaning in English, with a focus on lexicology, semantics, and stylistics through cognitive and corpus linguistics. He is Director of the Historical Thesaurus of English, and works mainly on applications of the Thesaurus in digital humanities, most recently through the AHRC/ESRC-funded SAMUELS and Mapping Metaphor projects. He has published, on his JISC-funded Hansard Corpus 1803–2003, a 2+ billion word corpus of political discourse over the past two centuries, and is working on enhancements to the Early English Books Online corpus. He is also Director of the STELLA Digital Humanities lab at Glasgow.

John M. Anderson is Emeritus Professor of English Language at the University of Edinburgh, where his entire university career was spent, apart from visiting posts at other European universities. He is interested in linguistic theory, particularly in relation to English and its history. He is mainly associated with the development of dependency-based approaches to linguistic structure and of localist case grammar and notional grammar.

Benjamin Blount is a retired anthropologist who received his Ph.D. in 1969 (University of California, Berkeley) and who taught at the University of Texas Austin, the University of Georgia, and the University of Texas San Antonio. He specializes in information systems, including human cognitive models. He was the inaugural editor of the *Journal of Linguistic Anthropology* and a former Editor-in-Chief of the *American Anthropologist*. His recent publications are on the history of cognition in anthropology, cultural models of knowledge in natural resource communities, and cognition in ethnographic research.

Frank Boers' initial research areas were lexicology and semantics (e.g., studies of polysemy and metaphor from a Cognitive Linguistics perspective). His more recent research interests were sparked by his experience as a language teacher and teacher trainer. He now publishes mostly on matters of instructed second or foreign language acquisition, with a particular focus on vocabulary and phraseology. He is co-editor of the journal *Language Teaching Research.*

Geert E. Booij obtained his Ph.D. degree in 1977 at the University of Amsterdam. From 1981 to 2005, he taught linguistics at the Vrije Universiteit Amsterdam. From 2005 to 2012, he was professor of linguistics at the University of Leiden. He is founder and editor

of the book series *Yearbook of Morphology* and its successor, the journal *Morphology*. He is the author of *The Phonology of Dutch* (1995), *The Morphology of Dutch* (2002), *The Grammar of Words* (2005), and *Construction Morphology* (2010), all published by Oxford University Press, and of linguistic articles in major Dutch and international journals.

Kate Burridge is Professor of Linguistics in the School of Languages, Cultures and Linguistics, Monash University. Her research focuses on grammatical change in Germanic languages, the Pennsylvania German spoken by Amish/Mennonite communities in North America, the notion of linguistic taboo, and the structure and history of English. Recent books include *Forbidden Words: Taboo and the Censoring of Language* (with Keith Allan, 2006), *Introducing English Grammar* (with Kersti Börjars, 2010), and *Gift of the Gob: Morsels of English Language History* (2010).

G. **Tucker Childs** is Professor in Applied Linguistics at Portland State University in Oregon. Over the past fifteen years he has focused on documenting endangered languages spoken on the coasts of Guinea and Sierra Leone. Childs has worked on many non-core linguistic topics such as sound symbolism, which is particularly robust in the African word class known as ideophones, as well as on pidgins and urban slangs, and language variation in general. He is editor of *Studies in African Linguistics* and begins work on the Sherbro language of Sierra Leone in 2015.

Eve V. Clark is the Richard Lyman Professor in Humanities and Professor of Linguistics at Stanford University. She has done extensive cross-linguistic research, both experimental and observational, on children's acquisition of a first language, with particular emphasis on semantic and pragmatic development. She has also worked on the acquisition of word formation, again cross-linguistically, and on the kind of information adults provide in conversation that licenses child inferences about new word meanings. Her books include *Psychology and Language* (1977, with H. H. Clark), *The Ontogenesis of Meaning* (1979), *The Lexicon in Acquisition (*1993), and *First Language Acquisition* (2nd edn, 2009).

David Crystal is Honorary Professor of Linguistics at the University of Bangor, and works from his home in Holyhead, North Wales, as a writer, editor, lecturer, and broadcaster on linguistic topics. His main interests relate to the history and development of English, as illustrated by such works as *The Cambridge Encyclopedia of the English Language* (2nd edn 2003), *The Stories of English* (2004), *Spell It Out: The Singular Story of English Spelling* (2012), and (with Hilary Crystal) *Wordsmiths and Warriors: The English-Language Tourist's Guide to Britain* (2013).

Simon De Deyne obtained a master's degree in theoretical and experimental psychology at the University of Ghent in 2000 and received his Ph.D. in psychology on the topic of semantic vector spaces from the University of Leuven in 2008. From 2014 he has been a research associate at the University of Adelaide. His research uses a computational approach to uncover structure and dynamics in the representation of word meaning in the mental lexicon. He also coordinates the small world of words project,

a cross-disciplinary effort to map the associative structure of the mental lexicon in various languages.

Philip Durkin is Deputy Chief Editor of the *Oxford English Dictionary*, and has led the dictionary's team of specialists in etymology for the past fifteen years. His publications include *The Oxford Guide to Etymology* (2009) and *Borrowed Words: A History of Loanwords in English* (2014), and he is currently editing a handbook of lexicography for OUP. His main research interests are in etymology, language contact (especially loanwords), polysemy, homonymy, and the history of the English language.

Christiane Fellbaum is a Senior Research Scientist at Princeton University, where she earned her Ph.D. in linguistics. Her work focuses on lexical semantics, computational linguistics, the syntax–semantics interface and multi-word expressions. She is one of the original developers of the WordNet lexical database and currently directs the WordNet project, for which she was awarded, together with the late George A. Miller, the 2006 Antonio Zampolli Prize. She is a co-founder and co-President of the Global WordNet Association, and supports the developments of cross-lingual lexical databases.

Dirk Geeraerts is Professor of Linguistics at the University of Leuven and founder of the research group Quantitative Lexicology and Variational Linguistics. He is the author of *The Structure of Lexical Variation* (1994), *Diachronic Prototype Semantics* (1997), *Words and Other Wonders* (2006), and *Theories of Lexical Semantics* (2010), and the editor, along with Hubert Cuyckens, of *The Oxford Handbook of Cognitive Linguistics* (2007).

Nikolas Gisborne is Professor of Linguistics in the School of Philosophy, Psychology and Language Sciences at the University of Edinburgh. He is interested in syntax and semantics, with a research focus on the interaction of subsystems in the grammar. He is the author of *The Event Structure of Perception Verbs*, published by Oxford University Press in 2010.

Cliff Goddard is Professor of Linguistics at Griffith University, Australia. He is a leading proponent of the Natural Semantic Metalanguage approach to semantics and its sister theory, the cultural scripts approach to ethnopragmatics. His major publications include the edited volumes *Ethnopragmatics* (2006, Mouton de Gruyter), *Cross-Linguistic Semantics* (2008, John Benjamins), and *Semantics and/in Social Cognition* (2013, special issue of *Australian Journal of Linguistics*), the textbook *Semantic Analysis* (2nd edn, 2011, Oxford University Press), and *Words and Meanings: Lexical Semantics Across Domains, Languages and Cultures* (co-authored with Anna Wierzbicka; Oxford University Press, 2014). He is a Fellow of the Australian Academy of Humanities.

Katharine Graf Estes is a member of the Psychology Department at the University of California, Davis. She received her Ph.D. in developmental psychology from the University of Wisconsin-Madison in 2007. Her research investigates the processes underlying early language acquisition. She has received funding from the National Institutes of Health and the National Science Foundation.

Anthony P. Grant is Professor of Historical Linguistics and Language Contact at Edge Hill University, Ormskirk. Having studied at York under Robert Le Page, he continued work on creolistics; his Ph.D. (Bradford, 1995) explored issues in agglutinated nominals in Creole French, and he has published over fifty articles and chapters on Native North American languages, Romani, Austronesian historical linguistics, pidgins, creoles, mixed languages, English dialectology and etymology, and lexicostatistical methods.

Peter Grzybek works at the Slavic Department of Graz University in Austria. After his MA thesis on 'Neurosemiotics of Verbal Communication' (1984) and his Ph.D. dissertation on 'The Notion of Sign in Soviet Semiotics', he qualified as a professor in 1994 with his 'Slavistic Studies on the Semiotics of Folklore'. His major research fields are linguistics and semiotics, literary and cultural theory, phraseology and paremiology. In his study of text and language, his particular focus is on exact and quantitative methods, attempting to apply statistical methods to the modelling of text structures and processes.

Reese M. Heitner teaches applied linguistics at Drexel University in Philadelphia. His interest in the developmental bootstrapping relationship between basic-level object categorization and phonemic word categorization and the experiment outlined in his chapter were inspired by Roger Brown's 'Original Word Game' approach to word learning.

Kristine A. Hildebrandt received her Ph.D. in Linguistics at the University of California Santa Barbara in 2003. She is currently an Associate Professor in the department of English Language and Literature at Southern Illinois University Edwardsville. Her research interests include phonetics–phonology interfaces, prosodic domains, the phonetic dimensions of tone, and language documentation and description.

Andrew Hippisley is Professor of Linguistics and Director of the Linguistics Program at the University of Kentucky, where he also serves as Chair of University Senate Council. He is author of *Network Morphology: A Defaults-Based Approach to Word Structure* (Cambridge University Press, 2012; with Dunstan Brown) and has published numerous articles on morphology in such outlets as *Yearbook of Morphology, Linguistics, Studies in Language, Natural Language Engineering* as well as chapters in books such as *Variation and Change in Morphology* (Benjamins, 2010), and *Handbook of Natural Language Processing* (Taylor & Francis, 2010). He is co-editor of *Deponency and Morphological Mismatches* (Oxford University Press, 2007), *Cambridge Handbook of Morphology* (forthcoming), and *Defaults in Morphological Theory* (Oxford University Press, forthcoming).

Michael Hoey is a Pro-Vice Chancellor and Emeritus Professor of English Language at the University of Liverpool, with interests in discourse analysis, lexicography, and corpus linguistics. His book *Patterns of Lexis in Text* won the English Speaking Union Duke of Edinburgh Award for best book in applied linguistics 1991, and his book *Lexical Priming* was short-listed for the BAAL Award for best book in applied linguistics 2005. He was chief consultant to Macmillan for their award-winning *Macmillan English Dictionary*, aimed at advanced learners of English. He is an academician of the Academy of Social Sciences.

Carole Hough is Professor of Onomastics at the University of Glasgow, where she has worked since 1995. She is President of the International Council of Onomastic Sciences, President of the International Society of Anglo-Saxonists, and Vice-President of the Society for Name Studies in Britain and Ireland. Her research interests focus particularly on the interaction between names and other areas of language, and she has published extensively on Anglo-Saxon studies, historical and cognitive linguistics, and onomastics.

Christian Kay is an Honorary Professorial Research Fellow at the University of Glasgow. She was an editor of the *Historical Thesaurus of the Oxford English Dictionary* (Oxford University Press, 2009) and founded the Scottish Corpus of Texts and Speech (SCOTS). She has written on historical semantics and lexicography and is currently working on two projects: 'Mapping Metaphor with the Historical Thesaurus of English' and 'SAMUELS (Semantic Annotation and Mark-Up for Enhancing Lexical Searches)'.

Robert Kennedy is a Continuing Lecturer at University of California, Santa Barbara. His research interests include phonology, phonetics, reduplication, accents of English, naming practices, and the linguistics of sports.

Adam Kilgarriff is Director of Lexical Computing Ltd. He has led the development of the Sketch Engine, a leading tool for corpus research used for dictionary-making at Oxford University Press, Cambridge University Press, and by many universities and publishers worldwide. Following a Ph.D. on polysemy from Sussex University, he worked at Longman Dictionaries, Oxford University Press, and the University of Brighton prior to starting the company in 2003.

Marie-Claude L'Homme is Professor in the Department of Linguistics and Translation of the University of Montreal, where she teaches terminology. She is also the director of the Observatoire de linguistique sens-texte (OLST), a research group investigating various theoretical, methodological, and applied aspects related to the lexicon (general and specialized). Her main research interests are lexical semantics and corpus linguistics applied to terminology. She develops, along with researchers in linguistics, terminology, and computer science, lexical resources in the fields of computing and the environment.

Barbara C. Malt is a Professor of Psychology at Lehigh University. Her research focuses on thought, language, and the relation between the two. She is especially interested in how objects and actions are mentally represented, how monolingual and bilingual children and adults talk about these objects and actions using the tools available in their language(s), and what influence, if any, the different ways of talking have on non-linguistic representations. She is an associate editor for *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Asifa Majid is Professor of Language, Communication, and Cultural Cognition at the Centre for Language Studies at Radboud University Nijmegen. Her work is interdisciplinary, combining standardized psychological methodology, in-depth linguistic studies, and ethnographically-informed description. This coordinated approach has been used

to study domains such as space, event representation, and more recently the language of perception.

Rosamund Moon is a Senior Lecturer in the Department of English at the University of Birmingham. She was previously a lexicographer, working on the Cobuild Dictionary Project at Birmingham (1981–90, 1993–9), and also at Oxford University Press (1979–81, 1990–93). Her main research areas are lexis and phraseology, lexicography, figurative language, and corpus linguistics; her publications include *Fixed Expressions and Idioms in English: A Corpus-Based Approach* (1998, Oxford University Press), and (with Murray Knowles) *Introducing Metaphor* (2006, Routledge).

Paul Nation is Emeritus Professor of Applied Linguistics in the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand. His books on vocabulary include *Teaching and Learning Vocabulary* (1990) and *Researching and Analysing Vocabulary* (2011) (with Stuart Webb) both from Heinle Cengage Learning. His latest book on vocabulary is *Learning Vocabulary in Another Language* (2nd edn, Cambridge University Press, 2013). Two books strongly directed towards teachers appeared in 2013 from Compass Media in Seoul: *What Should Every ESL Teacher Know?* and *What Should Every EFL Teacher Know?* He is also co-author, with Casey Malarcher, of *Reading for Speed and Fluency* (Seoul: Compass Publishing, 2007).

Victor Raskin the founder of the dominant linguistic theory of humour, is a theoretical and computational semanticist who earned his degrees in mathematical, structural, and computational linguistics from Moscow State University, USSR (Ph.D., 1970). Besides his alma mater, he taught at the Hebrew University of Jerusalem (full time) and Tel Aviv University (part time) in 1973–8. At Purdue University since 1978, he is now Distinguished Professor of English and Linguistics, with courtesy affiliations in Computer Science and Computer and Information Technology. He is the Founding Editor-in-Chief (1987–99) of *Humor: International Journal of Humor Research*, now into its 27th volume, the author of a grossly overrated and over-cited *Semantic Mechanisms of Humor* (1985, Reidel), a charter Board member of the International Society of Humor Studies, and its first elected academic President in 2000.

Nick Riemer works on lexical semantics and the history and philosophy of linguistics at the University of Sydney, Australia, and at the Laboratoire d'histoire des théories linguistiques, Université Paris-Diderot, France.

Niels O. Schiller is a professor of psycho- and neurolinguistics. He is interested in the cognitive system underlying language processing and its neural substrate. In particular, he investigates syntactic, morphological, and phonological processes in language production and reading aloud. Furthermore, he is interested in articulatory-motor processes during speech production, language processing in neurologically impaired patients (aphasia), and forensic phonetics. Schiller makes use of behavioural as well as neurophysiological (EEG/ERP) and neuroimaging (fMRI) methods. He has published widely in international peer-reviewed journals in his field.

Mark C. Smith is currently a lecturer in the Department of Psychology at the Open University. He obtained his Ph.D. in experimental psychology at the University of Birmingham. He has published on a wide variety of topics including linguistics, psychology, and art history. At present, his main topic of research is the problem of propositional unity.

Joseph Sorell holds a Ph.D. in Applied Linguistics from Victoria University of Wellington, New Zealand, and an MA in TESOL from Michigan State University. His research interests are in vocabulary learning, corpus linguistics, and cross-cultural communication. He has taught EFL, literature, linguistics, and computer literacy in Taiwan, Saudi Arabia, and Abu Dhabi and has worked or studied in Germany, Israel, the UK, and the USA.

Gert Storms obtained a master's degree in social and clinical psychology in 1983 and received his Ph.D. in mathematical psychology from the University of Leuven in 1990. He is currently a full professor at the laboratory for experimental psychology at the University of Leuven. Using both modelling and correlational and experimental procedures, he has been doing research in the areas of category learning, concept representation, psychosemantics, and psychological scaling.

Dennis Tay is Assistant Professor at the Department of English, Hong Kong Polytechnic University. He has been working on the application of cognitive linguistic theory to the analysis of discourse, particularly in mental health contexts. He has authored a monograph, *Metaphor in Psychotherapy: A Descriptive and Prescriptive Analysis* (John Benjamins, 2013), published articles in discourse analysis and counselling journals, and co-edited a volume (with Masataka Yamaguchi and Benjamin Blount), *Approaches to Language, Discourse, and Cognition* (Palgrave, 2014).

John R. Taylor obtained his Ph.D. in 1979. He is the author of *Possessives in English* (1996), *Cognitive Grammar* (2002), *Linguistic Categorization* (3rd edn, 2003), and *The Mental Corpus* (2012), all published by Oxford University Press, and co-editor of the *Bloomsbury Companion to Cognitive Linguistics* (2014). He is a managing editor for the series *Cognitive Linguistics Research* (Mouton de Gruyter) and an Associate Editor of the journal *Cognitive Linguistics*.

Rinus G. Verdonschot is currently a JSPS post-doctoral fellow at Nagoya University, Japan. His research uses behavioural and neuro-correlational methods, and focuses on language production, language comprehension, bilingualism, and music cognition.

Henk J. Verkuyl is Emeritus Professor of Linguistics at Utrecht University. His main research interest has been the semantics of tense and aspect resulting in work including *On the Compositional Nature of the Aspects* (1972), *A Theory of Aspectuality* (1993), and *Binary Tense* (2008). He also hides behind the pseudonym 'Dr. Verschuyl' (quite literally, 'Dr. Hyde', because the Dutch verb *verschuilen* = *hide* in English) with his *Cryptogrammatica* (Cryptogrammar), a book about the linguistic principles behind the

Dutch crossword puzzle (7th edn, 2005). In 2014 his most recent work under this pseudonym, a crossword dictionary, appeared as *Groot Puzzelwoordenboek* (1,399 pages).

Cynthia Whissell teaches psychology at Laurentian University (Ontario, Canada) with a focus on psycholinguistics, emotion, statistics, and research methodology. She teaches methodology in an interdisciplinary doctoral programme involving both the Humanities and the Social Sciences. Most of her research addresses the quantification of emotion expressed in the words and sounds of the English language. This gives her the excuse to study entertaining works of literature as well as trends in onomastics.

John N. Williams is Reader in Applied Psycholinguistics at the University of Cambridge. He is co-editor of *Statistical Learning and Language Acquisition* (Mouton de Gruyter, 2012) and area editor for the cognitive section of the *Encyclopedia of Applied Linguistics* (Blackwell, 2012). His research on cognitive aspects of second language learning and processing has appeared in *Studies in Second Language Acquisition, Language Learning, Second Language Research, Applied Psycholinguistics, Bilingualism: Language and Cognition*, and *Lingua*.

Margaret E. Winters is Professor of French and Linguistics in the Department of Classical and Modern Languages, Literatures, and Cultures at Wayne State University, where she is currently Provost and Senior Vice President for Academic Affairs. Her research interests are in historical semantics and the history of the Romance languages, both within the framework of Cognitive Linguistics. She has published in these fields in a variety of scholarly journals in North America and Europe and in volumes of collected papers. She has also published two editions of Old French courtly romances and has co-edited two volumes, one of papers in applied linguistics with Geoffrey Nathan, also at Wayne State University, and the other a co-edited book of papers on semantic change. She is working currently on a textbook of historical linguistics and papers both on semantics and the history of linguistic theory.

Alison Wray is a Research Professor in Language and Communication at Cardiff University. She gained a BA (1st class) from the University of York, UK, in linguistics with German and Hindi, and a D.Phil. in psycholinguistics from the same institution. After a postdoctoral position and a lectureship in York, she became Assistant Director of the Wales Applied Language Research Unit, University of Swansea, before being appointed senior research fellow at Cardiff University in 1999. Since 2004 she has been Director of Research for Cardiff's School of English, Communication and Philosophy. Her main research area is theoretical explanations for formulaic language (recurrent patterns in language output), extending across adult native speaker language, first and second language acquisition, the evolutionary origins of language, and language disorders, particularly attrition and compensation in the language of people with dementia. She has also contributed to researcher development agendas by means of textbooks, training materials, and research coaching.

INTRODUCTION

JOHN R. TAYLOR

1 Introduction

Words are the most basic of all linguistic units, and the ones which speakers of a language are most likely to be aware of and to talk about. Newspapers carry articles listing the new words which have made it into the dictionaries; parents identify the onset of speech by their child's first words; improving proficiency in one's own language is often thought of as a matter of increasing one's vocabulary; learning a foreign language is associated, above all, with learning the words; important aspects of a culture can be encapsulated in key words; languages are perceived to be related on the basis of similarities between words; prior to the 20th century, with its focus on syntax, linguistic description was mainly an account of words, their meaning, their pronunciation, their history, their structure, and the relations they contract with each other. One of the most striking facts about words—and one which is often overlooked, probably because it is so obvious—is their sheer number; for Carstairs-McCarthy (1999: 10–12) the abundance of words is one of the features which distinguish human languages from all animal communication systems. Practically all the major subdivisions of linguistic study have something to say about words. In the case of morphology and syntax this is self-evident, but it is no less true of phonology, historical linguistics, sociolinguistics, psycholinguistics, and language acquisition research.

What, though, are words, how is 'word' to be defined, and how do we identify the words in the language around us? Before approaching these questions, we need to clear up some ambiguities in the use of the term.

A first distinction is between **word token** and **word type**. The word count facility on your word processing package counts the number of word tokens, usually defined, for this purpose, as anything that occurs between spaces and/or punctuation marks (though apostrophes and hyphens are typically ignored; *mother-in-law's*, on the programme I am currently using, is counted as one word). When a publisher specifies that a manuscript should come in at a certain number of words (800, 8000, 150,000, or whatever),

it is word tokens that are at issue. Even here, though, we encounter some problems of definition and identification. A word processing package might identity such things as numerals, bullet points, listing devices such as '(a)' and 'i.', and even dashes (if surrounded by spaces) as words. These kinds of elements are not normally thought of as words, and authors submitting an 800-word article would not be expected to include them in the word count.

In any text (barring the very shortest), some word tokens, however they are identified, will occur more than once. These tokens are instances of the same word type. Inevitably, then, the number of word types in a text is going to be smaller (or, at least, cannot be greater) than the number of word tokens. It is word types that we have in mind when we speak of some words being more frequent in the language than others. Word types are also at issue when we enquire into vocabulary size. How many words did Shakespeare use? How many words does an average 10-year-old know? How many words do you need to know to get the gist of a newspaper article, a scientific paper, or a weather report? What are the most frequent words in English? Does English have more words than French?

The notion of word type hides some further distinctions. *Catch, catches, catching* are three different **word forms**. Yet if we were interested in stating the size of a person's vocabulary, we would probably want to regard the three forms as instances of the same word, or **lexeme**. The rationale for this is simple: the three word forms do not have to be individually learned. Once you have learned any one of the three forms (and provided that you know the rules for inflecting the verb and the conditions for the use of the different forms), you automatically have access to the other two forms. For this reason, a dictionary would list only one form, in this case, the 'basic' uninflected form *catch*. For English the matter is relatively trivial; regularly inflected verbs have four distinct forms: *talk, talks, talking, talked*, while nouns have only two: *dog, dogs*. For languages with more complex inflectional systems, the number of distinct forms can be quite large. For regularly inflecting verbs in Italian, Spanish, and Latin, the number of distinct forms can approach the high double digits, while nouns and adjectives in languages such as Russian and (again) Latin can have up to a dozen different forms. An Italian or Spanish speaker who learns a new (regular) verb has immediate access to (i.e. can produce and can understand) several score word forms. For heavily inflecting languages there is also the question of what the 'basic' form might be, i.e. the one that is to be listed in a dictionary and from which all others can be created. (The listed form is sometimes referred to as a **lemma**.) Often, more than one basic form is required. For example, Latin nouns, even those which are fully regular, are usually listed in dictionaries in both the nominative singular and the genitive singular forms.

The picture is complicated by the existence of irregular forms. Past tense *caught*, being irregular, does have to be learned. Even so, *caught* would (probably) not be regarded as a distinct lexeme, additional to *catch*, and (probably) would not be taken into consideration in statements of vocabulary size. It is important, however, to distinguish between word forms that *have to be* learned (irregular past tenses and irregular plurals are cases in point) and those which *have been* learned, and which are stored as such in the speaker's mental grammar. There is psycholinguistic evidence that perfectly regular forms, such as English plurals, may indeed be stored as such, alongside their base forms,

especially when the plurals are of high frequency vis-à-vis the singulars. One source of evidence is performance on lexical decision tasks. Here, you are shown a string of letters on a screen and must decide as quickly as possible whether the string constitutes a word or not. One factor which influences the speed of your response is the frequency of the word form in the language. High-frequency plurals tend to elicit shorter response times than the corresponding lower-frequency singulars, suggesting that language users have registered the plural forms in their mental grammar (see e.g. Sereno and Jongman 1997).

Not all word forms need to be learned, of course, and many are surely not learned. English speakers will have no hesitation in declaring *portcullises* to be an English word, even though few will ever have had occasion to speak of more than one portcullis. Consider the case raised by George Miller:

> For several days I carried in my pocket a small white card on which was typed UNDERSTANDER; on suitable occasions I would hand it to someone. 'How do you pronounce this?' I asked.
> He pronounced it.
> 'Is it an English word?'
> He hesitated. 'I haven't seen it used very much. I'm not sure.'
> 'Do you know what it means?'
> 'I suppose it means "one who understands."'
>
> (Miller 1967: 77–78)

Is *understander* an English word? No instances are recorded in the 100-million-word British National Corpus (BNC: Davies 2004–), though five tokens are found in the 450-million-word Corpus of Contemporary American English (COCA: Davies 2008–).[1] An example like the following, from the 1.9-billion-word Corpus of Global Web-based English (GloWbE: Davies 2013) is unlikely to raise any eyebrows:

> I'm no great understander of women.

The case of *understander* is crucially different from the case of *portcullises*. *Portcullises* is an inflected form of a familiar, if somewhat infrequent word. In general, inflectional processes have the property of not allowing 'gaps' in their paradigms; every singular noun has a plural form (even if irregular, as with *ox~oxen* and *sheep~sheep*), which we are able to create should the need arise, and every present-tense verb form has a corresponding past-tense form (even if irregular, as with *catch~caught* and *put~put*). *Understander*, in contrast, is a derived form. Whereas inflection creates variants of a word (lexeme), derivation creates new words, often of a different lexical category (part of speech), and

[1] How does 100 million words relate to a person's linguistic experience? Obviously individuals differ enormously with respect to the amount of language (spoken and written) that they are exposed to (and attend to). Assuming, however, an average speaking rate of 120 words per minute, at ten hours per day without breaks, it would take a person almost 4 years to read out loud the total content of the BNC (Taylor 2012: 16). It seems fair to conclude that 100 million words corresponds to a substantial chunk of a person's lifelong linguistic experience.

with sometimes specialized or unpredictable meanings. Derivational processes tend to be less than fully productive and subject to all kinds of restrictions—semantic, phonological, or simply idiosyncratic. We have *length*, not *longness; goodness*, not *goodth; ethnicity*, rather than *ethnicness; childhood* is a common word, while *infanthood, babyhood*, and *teeenagehood* are not. This is not to say that *infanthood* etc. are not English words; like *understander*, they are readily understood if encountered, but unlikely to be listed in a dictionary, and unlikely to be found in any but the largest of corpora.

A second distinction is between **actual** words and the **potential** words in a language. Actual words are those that have been attested; potential words are those which have not been attested but which could be created by one of the word-formation processes operative in the language and which could, therefore, become part of a language's vocabulary.

The notion of actual word, however, is far from unproblematic. Enumerating the existing words of a language looks straightforward enough and, one might suppose, could be accomplished on the basis of a very large and representative corpus of texts. Even so, no corpus is able to deliver a complete, definitive list of all the words of a language. Increasing the size of even a very large corpus will result in ever more potential words making their appearance; we saw this in the example of *understander* (this word is absent from a 100-million-word corpus, but is attested in a corpus four and a half times larger). Neither would it be possible to enumerate the potential (but not yet actualized) words of a language. To take just one example. How many potential words are there which take the suffix *-hood*? There are about a dozen words in *-hood* in common use, and a further couple of dozen which are somewhat rare but still understandable in their context. Linguists might talk of wordhood, and even sentencehood, texthood, and phonemehood.[2] But to make a list of all the not-yet-existing words in *-hood* would be an impossible task.

The question of potential words becomes especially acute when we consider two further processes of word creation, in addition to derivation, namely, compounding and blending. Compounding is an extremely productive process in English. In principle, just about any two randomly selected nouns can come together in a noun–noun compound; the process is also recursive, in that a compound can be built out of already existing compounds. Is *airport* one word or two? The orthography suggests that it is one word, as does the phonology (the compound has only one primary stress) and the semantics (an airport is not really a kind of port). But what about *seaport*? *Bus route*? *Airport bus route*? *Airport bus route management company*? Since nominal compounding is recursive, the number of noun–noun compounds in English is truly open-ended, and any attempt to list all the not-yet-actualized compounds would be futile.

Blending is another word-formation process with open-ended outputs. Blends are created by combining the first part of one word with the second part of another word (where 'part', in the limiting case, can comprise the whole word). Often-cited examples

[2] In my first draft of this paragraph, I used *texthood* and *phonemehood* as examples of non-existing potential words in *-hood*. A subsequent Google search showed that these words were indeed attested in linguistics texts.

include *brunch* (*br*eakfast + l*unch*) and *smog* (*sm*oke + f*og*). Many blends are, in fact, compressed syntagms and are subject only to the ingenuity and creativity of speakers (which is not to deny that phonological and other constraints may not be relevant: see Kelly 1998 and Gries 2006). Readers may be familiar with the term *Brexit*—referring to the possibility (or desirability) of a *Br*itish *exit* from the European Union. A couple of decades or so ago, when the concept of Brexit had not yet crystallized, the word would not even have been deemed to be potential.

2 Identifying words

Identifying the words in an utterance might seem a trivial matter (even given the distinctions discussed above). In most cases, it is easy. But sometimes it is not. It is to these problematic cases that we now turn.

To illustrate some of the issues, consider the tag question below. How many words are there in the following, and what are they?

Isn't it?

One answer would be 'two', separated by a blank space. On the other hand, one might argue that *isn't* is a shortened form of two words, *is* and *not*. This is the analysis supported by the Corpus of Contemporary American English. If you search the corpus for the form *isn't*, you will be instructed to insert a word space before contracted *n't*; that is, *isn't* is taken to be *is* + *n't*. Similarly with *wasn't, aren't*, and *don't*. This procedure guarantees that occurrences of the forms *isn't, aren't*, and *don't* contribute to the frequency count of the word forms *is, are*, and *do*. There are, however, a number of problems. First, *Is not it?* is not a usual sequence. If the 'component words' of *isn't it* are spelled out in full, the accepted form would be *Is it not?* Second, application of the procedure to the forms *won't, can't*, and *shan't* attributes word status to *wo, ca*, and *sha*. In fact, these 'words' will turn out to have quite a high frequency of occurrence in the language. We might want to say that *wo, ca*, and *sha* are forms of the lexemes WILL, CAN, and SHALL. How, then, do we handle the form *ain't*? This can be a contracted form of *am/is/are not*, as well as of *has/have not*. Then there is the question of how to deal with the orthographic rendering of *isn't it* as *innit*, sometimes written as *ennit*. Is this one word (no internal spaces), two words (if so, what are they?), or a contracted form of three words?

The question of word identity also arises in connection with the following example (Lakoff 1987: 562):

There's a man been shot.

Suppose we say that *there's* is a contracted form of two words. What is the second of the two words? *Is* or *has*? Note that *There is a man been shot* and *There has a man been shot*

are both of dubious acceptability. The choice of a tag might suggest that the contracted item should be construed as *has*.

There's a man been shot, hasn't there? /*isn't there?

But try putting the sentence into the plural:

?There've two men been shot.
There's two men been shot.

The second example seems preferable, suggesting that *there's*, pronounced [ðəz], is a unique word form, specific to the presentational construction. This supposition is supported by the following examples (sourced from the Internet), where the form *there's* appears to be insensitive to possible paraphrases with *is*/*are*/*has*/*have*:

There's lots of people been saying it's dangerous.
There's someone been looking for scapegoats.
But there's some people been waiting two hours.
There's some people been here longer than you.
There's someone been on my mind lately.
There's somebody been asking around about you.

Decisions on these matters are of vital importance to anyone studying the statistical properties of words in text, such as their frequency or their length. How we handle a relatively frequent form such as *isn't* will impact on frequency measures for *is* and *not*. If *isn't* is treated as one word, the frequency profile of *is* (and of the lexeme BE) will be lowered. A similar situation arises in connection with conventions for the use of the word space. Older texts—the novels of Charles Dickens are an example—have *some where, some one, every one*, whereas the modern practice is to join up the two components: *somewhere, someone, everyone*. We write *indeed* (one word) but *in fact* (two words), *perchance* but *by chance*. *Of course* distributes in the language as if it were a single word (cf. German *sicher, natürlich, selbstverständlich*; French *naturellement*); treating it as two words increases the frequency count for *of* and for *course*. The consequences are not insignificant. A search of the BNC shows that of the 48,654 occurrences of the noun *course*, 29,429—about 60 per cent—are in the phrase *of course*. Treating *of course* as two words more than doubles the frequency count for the noun *course* in the language.

There is a common theme here. The dubious cases nearly all concern high-frequency items and 'small words', such as parts of *be, have, do*, and markers of negation. It is worth noting that word-frequency distributions (see Sorell's chapter) tend to become somewhat erratic when the highly frequent words are considered. It would seem that when we get down to the most frequently occurring bits of a language, the notion of 'word' begins to dissolve.

3 Approaches to 'the word'

The above remarks notwithstanding, the reader may well be objecting that the notion of what constitutes a word is in most cases rather clear-cut. *Dog* and *cat* are words, as are *run* and *sesquipedalian*. The existence of clear-cut cases, alongside more problematic cases, points to the word as a prototype category. Prototypical words share a number of distinctive properties: orthographic, phonological, syntactic, and semantic. The problematic cases we have been considering constitute marginal words, in that they fail to exhibit the full range of characteristic properties.

Let us consider the properties in turn, taking as our reference point examples of prototypical words, with an eye on less prototypical, more marginal examples.

(a) Orthography. Orthographically, a word is separated by spaces or punctuation marks (though what counts as a punctuation mark can be open to question). Obviously, the criterion cannot be universally applicable, since some writing systems do not make use of the word-space convention (neither, of course, is it relevant to speakers who are not literate in their language). For literate English speakers, though, and for speakers of other European languages, the word-space criterion is paramount; it is also the criterion preferred by workers in computer language processing. It must be borne in mind, however, that word-space conventions are just that—conventions, which have emerged over the course of time and presumably in response to non-orthographic principles of wordhood. Even today, the conventions are not fully settled. One sometimes finds *nevertheless* and *nonetheless* written out as three words. Then there is the case of compounds. These, if conventionalized, are often written without a space or, variably, with a hyphen; otherwise, the components are separated by spaces. The conventions are different in German and Dutch; here, even nonce compounds are joined up; cf. Verkuyl's 'Word Puzzles' (this volume) example of *zuurstoftententententoonstelling* 'oxygen tent exhibition'.

(b) Phonology. A second criterion relates to pronunciation. A number of aspects are relevant. First, a phonological word must have one, and only one, main accent (or primary stress—the terminology is fluid). The number of primary stresses in an utterance is therefore a marker of the number of words (though the incidence of stresses does not indicate the location of the word boundaries.) This is why *airport* and *bus route* would be considered to constitute single words, whereas *busy port* and *direct route* would consist of two words. On this criterion, many of the 'little' words, such as articles, prepositions, and parts of *be, do*, and *have*, would not constitute words; lacking stress, they must attach to an adjacent item. Thus, *fish and chips* would consist of two phonological words: [fish and] [chips], with unstressed *and* [ŋ] attaching as a clitic to the preceding stressed word, reflecting the predominately trochaic foot structure of English. Of course,

the little words can, on occasion, be spoken with stress, for contrastive emphasis, for example.

A second criterion is that of being able to be preceded and followed by pauses. (One recalls Bloomfield's 1933: 178 definition of word as a minimal free form.) *And*, on this criterion, would count as a word. It can be utterance-initial and can be followed by a hesitation pause. Dixon and Aikhenvald (2002a: 12) propose, as a useful heuristic, the pauses that a native speaker makes when repeating an utterance 'word for word', as when giving dictation. *Fish and chips* is likely to be dictated as three words, with *and* being spoken with the full [æ] vowel.

A third phonological criterion is more subtle, and relates to the fact that some phonological generalizations (or 'rules') may be restricted to words and their internal structure, whereas others apply only across word boundaries. For example, double (or geminate) consonants are not allowed within English words. The spelling notwithstanding, *adder* is pronounced with a single 'd'; compare Italian *freddo*, which contains a lengthened 'd', spread over two syllables. Double consonants can occur in English, however, but only over word boundaries: *good dog, black cat, big girl, love Vera, his sister*, etc. The occurrence of a geminate can thus be seen as a marker of a word boundary. An interesting case is provided by examples like *non-native* and *unnatural*, which may be pronounced with a lengthened 'n', suggesting that *non-* and *un-* are (phonological) words. In contrast, there is no lengthened 'n' in *innate, innumerable*, or *innocence*, indicating that the prefix *in-* lacks the status of a phonological word.

In considering the phonological criteria for wordhood, we need to bear in mind that pronunciation—even more than writing—is variable. *Fish and chips* can be spoken, variably, with two or three main accents. *Unnatural* does not have to be spoken with a geminate 'n'. This means that the status of prefixed *un-* as a (phonological) word is also variable.

(c) Syntax. A third criterion is syntactic, or, less contentiously, distributional, having to do with the kinds of things a linguistic unit can, or must, or may not, occur next to. Here a distinction needs to be made between a word's **internal syntax** and its **external syntax**.

Internally, a word permits no variation, pauses, or insertions. Essentially, then, a word has no internal syntax. That is why compound *blackboard* (the thing to write on) is considered to be one word, whereas *black board* (referring to a board which is black) is two words. The latter can accept intrusions—*a black and white board, a blackish board*, etc.—the former cannot (at least, not if its status as a compound is to be preserved). One well-known caveat pertains to the phenomenon of expletive insertion: *abso-bloody-lutely*. The insertion is possible between two (typically trochaic) feet, each with a stressed syllable (McCarthy 1982).

Even so, there are cases which are less than clear-cut. Suppose we want to refer to a collection of blackboards and whiteboards (both single-word compounds, by most criteria). The conjunction *black- and whiteboards* (spoken with two main

accents) seems entirely plausible. Or take the case of *mother-in-law*. Is this one word or three (or two)? One relevant consideration would be what the plural form is: do we say *mother-in-laws* (suggesting that we are dealing with only one word, permitting no internal intrusions) or *mothers-in-law* (suggesting a word division between *mother* and *in-law*). According to the COCA corpus, the former is more frequent by a factor of about 5:1; a similar bias exists for other *in-law* expressions. The odds, therefore, are in favour of regarding *mother-in-law* as one word (though the very fact of variation is surely of interest, showing that the word status of *mother-in-law* is not fully fixed). Note that by the no-intrusion criterion, *in-law* would have to count as one word; indeed, it functions as a regular count noun, with a predictable plural form: cf. *my in-laws*.

Another interesting case is *mum and dad*, and its plural. We would expect *mums and dads*, and this indeed is the preferred form. However, consider the following (from the GloWbE corpus):

If the markets can't pick interest rates how can the *mum and dads* pick interest rates?

Here, *mum and dad* appears to be functioning as a single semantic unit (as a single lexeme, in fact); it does not refer to a collective consisting of a mum and a dad, but is roughly equivalent to 'typical retail investor'.

External syntax has to do with the items that a word can occur next to. There are very few restrictions on the neighbours of a (prototypical) word. Certainly, (attributive) adjectives tend to occur immediately before nouns, predicative adjectives immediately after *be, become, seem*, etc. But these are tendencies, not absolutes; for example, an adverb can easily be inserted between *be* and a predicative adjective. Compare the situation with that of a bound affix such as *-ness*. This can only occur (indeed, must occur) as an affix to an adjectival stem, with no intrusions allowed between stem and affix.

In terms of its external syntax, possessive *'s* is a word. The morpheme attaches to whatever happens to occur last in a possessor nominal. Mostly, of course, the possessive morpheme ends up attaching the possessor noun (*the man's hat*); in principle, however, the morpheme can attach to practically any kind of word (*the man I was speaking to's hat*); see Hildebrandt's chapter. The possessive morpheme would not count as a prototypical word, of course, because it is phonologically dependent and cannot occur between pauses. Its status, rather, is that of a clitic.

(d) Semantics. A prototypical word associates a stable phonological/orthographic form with a coherent semantic category, with its distribution in the language being determined by the syntax. *Dog* and *cat, airport* and *sesquipedalian*, obviously qualify for word status on this criterion—though the case of articles, parts of *be* and *do*, some prepositions such as *of*, and the possessive morpheme, is less clear. Homonymy and polysemy also muddy the picture—polysemy, in that a word form may be associated with a range of semantic values, and homonymy, in that the semantic values may be so disparate that it may be more appropriate to

speak of two or more words which happen to share the same form. Nevertheless, it is clearly the semantics which motivates the word status of *mum and dad* and *mother-in-law* (discussed above).

4 Word as prototype

As the preceding remarks will have shown, the various criteria for wordhood do not always coincide. The situation indicates a prototype approach to words; there are 'good examples' of the category (where all the criteria coincide), and more marginal examples, where only some of the criteria apply. Thus, a prototypical word will

- have a stable phonological form, intolerant of interruptions and internal variation;
- be associated with a reasonably stable semantic content (or array of related contents, in case the word is polysemous);
- be separated in writing by spaces;
- have one main stress and be pronounceable on its own, surrounded by pauses;
- be relatively free with regard to the items to which it can be adjacent.

These criteria are particularly useful when we try to differentiate words from competing categories, such as word vs. phrase, word vs. bound morpheme, word vs. clitic. The dividing line between these categories is not always clearly drawn. The definite article *the* (when unstressed) has clitic-like properties, adjective+noun combinations may waver in their status as phrases or compounds, affixes can sometimes get detached from their hosts and function as full-fledged words (*anti, pro, ism*, etc.) (Taylor 2003). This kind of fuzziness is just what one would expect on a prototype view of 'word'.

5 Are these words?

To see how a prototype-based approach might be applied, consider the various vocalizations which interlard our speech; these are represented orthographically as *oh, ah, um, er, hm, erm*, etc. Are these words? Let us go through the features:

(a) These vocalizations, if written, are separated by spaces. This makes them words. On the other hand, their 'spelling' is somewhat variable, and we would not expect to find them listed in a dictionary. This speaks against their word status.
(b) The vocalizations are phonologically autonomous in that they may bear primary stress, they do not need to lean on adjacent elements, and they may be surrounded by pauses. From this point of view, they are undoubtedly words.

(c) They are relatively free to occur at any point in an utterance and are not required to attach to items of a specified syntactic category. In this, they are like words. They differ from prototypical words in that they do not contract syntactic relations with neighbouring items; instead, they have the character of parenthetical intrusions.
(d) They are not associated with a fixed semantic content; their function is discoursal and attitudinal, signifying such things as hesitation, uncertainty, and prevarication. Yet they are not somatic noises, whether voluntary or involuntary (like coughs and yawns). They are, on the contrary, language-specific; mostly, these vocalizations are made up of phonetic segments characteristic of the language in question. English speakers do not hesitate and prevaricate in the same way as French or Russian speakers do.

In brief, the vocalizations exhibit a number of properties of typical words, yet they would by no means be considered full-fledged words; they are marginal words par excellence. Indeed, many people would not consider them to be words at all. They are typically absent from the word inventories that we find in dictionaries and the word lists that are derived from corpus analysis. And if they are taken into consideration in corpus studies, they are likely to cause all manner of problems (see Sorell's chapter).

6 Exhaustive analysis, no residues

A different approach would be to question the view that utterances can be exhaustively analysed into words.

The idea of exhaustive analysis is a common assumption in linguistic description. We divide texts up into sentences, sentences into words, words into morphemes, and all of these ultimately into phonemes. In all cases, the expectation is that once the dividing-up has been done, nothing will be left over. Now, in the case of the word–morpheme relation, 'residues' are in fact not at all uncommon. Sometimes, the residue is accorded the status of a 'cranberry morpheme', named after the *cran-* of *cranberry*. Even this ruse to save the exhaustive-analysis approach, however, quite often fails. Take the case of the names of many of the consonants; these terminate in [iː]: *B* [biː], *C* [siː], *D* [diː], etc. The strength of this association is manifest in the name of the final letter, *Z*, often pronounced [ziː] rather than [zɛd]. But if we recognize [iː] as a morpheme (with roughly the semantic value of 'name of a consonant'), what are we to say about the initial segments [b], [s], [d], etc.? To refer to these as morphemes, even as cranberry morphemes, with the semantic value of, respectively, the consonants *B, C,* and *D*, seems a bit outlandish. The proper approach, it seems to me, is to recognize that while bits of a word might have a function across several words in the lexicon, we are under no obligation to accord comparable status to all of the remaining bits.

The idea that words and utterances can be exhaustively analysed into phonemes (or, less contentiously, phones, or phonetic segments) is much more entrenched, and counterexamples are rarely entertained or discussed. Few linguists would want to quibble with Chomsky's assumption that 'each utterance of any language can be uniquely represented as a sequence of phones' (Chomsky 1964: 78). Yet there are all manner of noises that people make as they speak—coughs, laughs, giggles, grunts, smacking of lips, inhalations, sucking on teeth, and so on. These are going to be filtered out of any linguistic (phonological) analysis (though they may be of interest in a study of the communicative act in progress). On what basis they are to be ignored is, however, rarely addressed; a notable exception is Zellig Harris: see Harris (1951: 18–19) on why coughs should be overlooked in phonemic analysis.

(On a personal note: I first became aware of the problematic nature of these non-speech noises when working with some colleagues in the Information Science department on a system of phoneme—and ultimately, it was hoped, word—recognition. The problem was that the system interpreted the sound of inhalation variously as [h], [f], and [θ] and the clanking of furniture as voiceless plosives. Adjusting the sensitivity of the system did not solve the problem. This simply resulted in genuine cases of [h], [f], and [θ] being missed.)

Occasionally, the status of a noise as a linguistic or non-linguistic element is far from clear. One of the texts in Crystal and Davy (1975) consists of a lengthy deadpan monologue narrating the attempts of an accident-prone driver to reverse her car. The listener responds with an utterance transcribed by Crystal and Davy as follows:

|HM̀| –t |oh BLÌMEY|

The authors inform us (p. 46) that the 't' represents an alveolar click [!], expressive of the speaker's sympathetic appreciation. Is this sound part of the sound system of English, or is it extraneous to the system, comparable, perhaps, to a noisy inhalation of breath, or even (to take a non-acoustic example) a shake of the head? Listening to the recording which accompanies the volume suggests that the click is functioning as a consonantal onset of the word (is it a word?) *oh*. On the other hand the listener could have responded simply with the click (perhaps accompanied by a shake of the head and with raised eyebrows). At best, the click is a (very) marginal phonetic segment of English.

The relevance of this to our main topic is as follows. We can easily recognize the words in an utterance. Sometimes, though, there are bits left over which fail to achieve word status on the usual criteria. We might regard these as 'marginal' words, which fail to exhibit the full range of word-defining properties. An alternative approach, suggested by Wray (this volume), would be to acknowledge that utterances are not composed only of words; once the words have been identified, there may be bits and pieces left over which cannot easily be assimilated to the word category. As Wray puts it, words are the bits that fall off of an utterance when you shake it (p. 750). To borrow an image from Kilgarriff's chapter, what is left is like the stuff found at the bottom of a schoolboy's pocket: 'very

small pieces of a wide variety of substances, often unsavoury, all mixed together, often unidentifiable' (p. 33).

7 Overview of the volume

With quotations from over thirty writers, **David Crystal** documents the fascination of poets, novelists, and critics with words, and reflects on the various viewpoints that have been expressed in literature and linguistics about the form and function of words and their relationship to thoughts, actions, and culture. He touches on a number of topics which are dealt with in subsequent chapters, such as historical change, word innovation, and the impossibility of quantifying the size of the lexicon.

Adam Kilgarriff takes up the question of how many words there are in a given language. Dictionaries (and their users) persist in the fiction that a definitive answer is possible—if it's 'not in the dictionary', then it's not a word; conversely, if it is a word, then it has to be in the dictionary. There are many reasons why a definitive listing is not possible. Word-formation processes are productive, to a greater or lesser extent, which means that a language's vocabulary is essentially open-ended. Second, all manner of specialized interests and activities, from chemistry to cooking, have their own vocabulary, and many branches of knowledge have the means for creating their own words as needed. Then there is the question of how to handle foreign borrowings, variant spellings and pronunciations, misspellings (and mispronunciations), and dialectal forms. A more fruitful line of enquiry would concern the words that a person needs in order to function in a given context—the topic of Paul Nation's chapter.

The chapters by **Marc Alexander** and **Christian Kay** deal, respectively, with dictionaries and thesauri. A dictionary lists the words, and for each word describes its meaning. A thesaurus does the reverse—it lists the concepts, and for each concept gives the words which can express it. Both kinds of resource have a venerable history, and both are undergoing rapid developments in our digital age. Alexander addresses the sometimes conflicting concerns of scholars, users, and publishers in the design and presentation of dictionaries, while Kay raises the question of whether a universal system of concepts is possible, or whether conceptual classification should be allowed to emerge from language usage itself.

There follow two chapters dealing with quantitative aspects of the words of a language—their frequency and their length. George Zipf (1949) pioneered work on these topics, pointing to a linear relation between the logarithm of word frequencies and the logarithm of their frequency ranking, also noting that word length tends to correlate inversely with frequency. **Joseph Sorell** presents updates on the Zipfian frequency distribution, nuanced with respect to text types and genres. He also speculates that the distribution may be motivated by functional considerations, having to do, specifically, with the accessing of words stored in small world networks. **Peter Gryzbek** addresses the length of words—again building on the Zipfian thesis that the more

frequently a word is used, the shorter it tends to be. His research explores the dynamics of word length—within texts, within genres, within languages, and over time. He argues that the study of word length involves much more than just the length of words. Word length stands at the intersection of numerous language systems, and impinges on such matters as polysemy, a language's phonological inventory, and syntactic and textual organization.

The next two chapters go beyond the word, narrowly construed. **Rosamund Moon** writes about multi-word units—groups of words which have quasi-unitary status, and which need to be learned as such. Quite a lot comes under the scope of Moon's topic—from fixed idioms and proverbs to recurring phrases and preferred collocations; indeed, a significant proportion of any text will comprise formulaic material of different kinds. Of special interest are phrasal patterns which permit some degree of variation, sometimes with humorous effect.

Michael Hoey pursues the matter on the detailed analysis of a short text fragment. Speakers, he argues, subconsciously note the collocations that a word makes with other words, as well as the collocations that the combination has with other words or word combinations. Speakers also note the syntactic environments of words and word combinations (i.e. their colligations) and their association with words of particular semantic sets. Parts of words also participate in these kinds of relations. Thus, a word can be said to prime the contexts of its previous uses. These primings influence the way we interpret a word in context as well as our future uses of the word. The chapter also shows how collocation contributes to textual cohesion. The relation of a word to its neighbours thus lies at the very core of language as stored in the mind and in its use.

The next group of chapters address word structure and the status of words in linguistic theory. **Geert Booij** overviews the internal structure of words (morphology) and processes for word creation—inflection, derivation, compounding, blending, and univerbation (the process whereby groups of words acquire the status of single words). Complex words may acquire pronunciations and meanings which are not fully predictable from their parts, thereby obscuring a word's internal structure.

The notion of part of speech is a familiar one. However, as **Mark Smith** shows, the setting up of categories, and determining their membership, are subject to decisions which are ultimately arbitrary. Some words are 'quirky', and do not readily fit into any of the recognized categories; in a sense, they belong to categories with a membership of one. And even for the better-established categories, such as noun or verb, it is rarely the case that their members share the same range of properties.

Nikolas Gisborne argues for the lexicalist hypothesis, according to which words are 'atoms' whose combination is sanctioned by the syntax. Essentially, this boils down to the claim that the syntax does not need to 'look into' the internal structure of words. Gisborne defends the claim by addressing a number of controversial examples, including the use of the passive participle in English, pronominal clitics in French, and noun incorporation in Mohawk.

Kristine Hildebrandt discusses the word as a phonological unit, i.e. a unit around which language-specific phonological generalizations may be made. Mostly, the

phonological word coincides with the grammatical word as discussed by Gisborne, but often it does not. Moreover, different languages draw on different sets of phonological criteria. Rather than seek a universal definition, phonological words emerge on the back of phonological processes of a given language.

Andrew Hippisley turns to the question whether the word has universal status—whether, that is, all languages have items that we want to call words. Words are, he argues, the basic symbolic resources of a language, uniting a pronunciation (and spelling), a meaning, and a syntactic status. This neat association is often upset, most obviously by homonymy and synonymy, as well as by such phenomena as incorporation; at best, then, words are universals of a fairly plastic kind. Hippisley also considers the question whether words are unique to human languages—whether, that is, animal communication systems have 'words'. He argues that they do not; animal signs are tied to specific external phenomena, whereas words designate 'mind-dependent' concepts. Hippisley proposes that cross-language variation in the properties of words results from different solutions to the problem of how complex conceptual information is channelled into a one-dimensional stream of sounds.

Kate Burridge's chapter addresses word taboo. When a word denotes something unpleasant, forbidden, or emotionally sensitive, people behave as if the very sound of the word equates to what it denotes. The word itself becomes unpleasant, and should be avoided. Even words that sound similar to the taboo word may be shunned. The feelings are so intense that they may affect expressions recruited as euphemism. Taboo is therefore a potent source of lexical renewal and semantic change.

Word taboo challenges a basic tenet of Saussure's (1916) theory of the arbitrariness of the link between sound and meaning. The thesis of arbitrariness also needs to be nuanced by the widespread phenomenon of sound symbolism. The less-than-arbitrary association of sound and meaning is the topic of **Tucker Childs'** chapter. Some of these associations—such as between high front vowels and the idea of smallness—would appear to be universal, while others emerge by association within the language and can affect significant portions of a language's lexicon.

The next group of chapters deal with semantic issues. Prefacing his chapter with a warning that 'word meaning' is a theoretical notion, whose legitimacy derives from its usefulness in explaining language use, and noting that some languages do not even have a term for this supposed property of words, **Nick Riemer** discusses two major approaches to word meaning. One is based on reference, and studies the kinds of things and situations that a word may refer to. While a referential approach has the appeal of 'objectivity', it is clear that many words lack referents (*ghost* and *Martian*, presumably); the approach is also unable to capture the affective connotations of words. The other approach appeals to shared concepts; indeed, it is because of the concepts that they link to that words are able to refer to the outside world at all. A third approach—which Riemer briefly touches on—proposes that word meaning can be explicated in terms of relations amongst words themselves—a topic developed in Christiane Fellbaum's chapter.

Barbara Malt discusses the referential use of words, specially, the factors which motivate a speaker's choice in the designation of objects, events, relations, properties

of things, etc. These do not usually come associated with a name which intrinsically belongs to them and which uniquely identifies them. Even for those entities which have been assigned a proper name, a speaker is still at liberty to refer by means of a pronoun or a descriptive phrase. A speaker's choice is motivated by many factors—the options made available by the language, the ones that the speaker has learned and which are available at the moment of speaking, as well as the speaker's assessment of the knowledge base of the addressee.

Naming is also the topic of **Marie-Claude L'Homme**'s chapter on terminology. Experts in all fields of knowledge, from scientists and engineers to bureaucrats and hobby enthusiasts, have developed systems of terms for the naming and classification of concepts, with the aim of facilitating communication amongst specialists and avoiding ambiguity and misunderstandings. The chapter draws attention to the dynamic nature of terminologies and their dependence on features of the communicative situation, including the supposed degree of expertise of the addressee.

Christiane Fellbaum discusses the semantic relations amongst words in a language. She makes the distinction between relations between words and relations between concepts. Antonymy (the relation of 'opposites') is a relation between specific words (*big, little; large, small*), as are collocational preferences (*strong* collocates with *tea; powerful*, a near synonym of *strong*, does not). Taxonomic and meronymic (whole–part) relations are relations between concepts (*vehicle, car; car, wheel*). Relations (lexical and semantic) are the basis of the WordNet project, a large electronic database of about 155,000 words, incorporating properties of both a conventional dictionary and a thesaurus, organized around relations of various kinds. It is hypothesized that it is just this kind of network which underlies speakers' ready and effortless access to the contents of their lexicon.

As every language learner knows, a word in one language rarely has an exact translation equivalent in another. **Asifa Majid** examines two semantic domains—perception and the human body—in order to illustrate similarities and differences across a wide range of languages in the way in which these areas of human knowledge are structured and lexicalized. She discusses the possible sources of the variation—the environment, the ecological niche where the language is spoken, cultural practices, and historical development.

Cliff Goddard's chapter reviews different ways in which words can be carriers of culture-related meaning. Culture-laden words are untranslatable, by normal means, into other languages. The chapter reviews examples from various abstract and concrete domains, stressing that cultural themes are often conveyed by a suite of related, mutually reinforcing words. The chapter demonstrates how the Natural Semantic Metalanguage (NSM) approach is able to capture subtleties of meaning, while counteracting the danger of conceptual Anglocentrism creeping into the definitions.

Philip Durkin and **Dirk Geeraerts** discuss historical matters. Durkin takes the broader perspective of etymological research, presenting the methodology of tracing (or reconstructing) the historical past and the relationship between the words of different languages. He highlights cases where the line of descent from an earlier to a later form is blurred by mergers and blendings. Geeraerts zooms in on processes of semantic change, the emergence of new meanings, and the loss of older ones.

All languages borrow items from other languages, some extensively. This is the topic of **Anthony Grant**'s chapter. Some borrowings are for naming previously unknown concepts; others replace (or coexist with) native terms. While nouns are particularly subject to borrowing, there do not appear to be any absolute restrictions on what can be borrowed. Borrowings can be an important pointer to the history of a language, and can function as a conduit for the introduction of new phonemes and new inflectional and derivational morphemes into a language.

Margaret Winters writes on the results of borrowing on language/vocabulary structure. While English, to take one of her examples, is 'basically' a Germanic language, a large number of words (and syntactic constructions) have been borrowed from French. In some cases, the two strands have been homogenized, with only an expert being able to disentangle the influences. Often, however, the borrowed items may constitute a sub-component of the vocabulary, with its own phonological, morphological, semantic, stylistic, and even syntactic and orthographic identity, and recognized by speakers as different from the core lexicon of the language.

The next group of chapters address the mental representation and mental access of words. The topic of the chapter by **Simon De Deyne** and **Gert Storms** is research on word associations. The word association paradigm is a familiar one: given a word (such as *bread*), what is the first word that comes to mind? (Answer, for most people, *butter*.) The association paradigm provides insights into the links which exist in the mental lexicon, and is a valuable accessory to findings from usage data. The authors propose that the mental lexicon can be viewed as a large association network, whose properties facilitate lexical search and retrieval.

The accessing of words from the mental lexicon is the topic of the chapter by **Niels Schiller** and **Rinus Verdonschot**. Lexical access is a crucial component in the process of transforming thoughts into speech; a widely used paradigm requires subjects to name pictured objects, often against various kinds of distraction. The chapter reviews a number of models for lexical access. Also addressed is the storage and access of morphologically complex words, including compounds.

John Williams addresses the storage and access of words in bilingual speakers. Do bilinguals keep their two languages distinct, or do the representations overlap and interact in usage situations? Does the relation between the languages change as a function of level of proficiency and context of acquisition? Williams reviews extensive research showing that, when performing tasks in one language, bilinguals and proficient second-language learners cannot avoid activating orthographic, phonological, lexical, and semantic representations in their other language(s), suggesting that representations in a bilingual's different languages continuously compete with each other for selection. Bilinguals rely on domain-general executive control mechanisms to manage the activation levels of their different languages.

Dennis Tay writes on words and psychological disorders. He approaches the matter from two perspectives. On the one hand, the disorder manifests itself in the patient's inability to access words, or to use them appropriately. Disorders of this nature feed into models of lexical storage and access. The second perspective is to regard words,

particularly the metaphorical use of words in therapeutic discourse, as pointers to psychological disorders which are not in themselves inherently linguistic, such as psychogenic seizures and delusional thought. Tay proposes some possible directions for metaphor and corpus research in mental health discourse.

The next three chapters deal with child acquisition. Early word acquisition is the topic of **Eve Clark**'s chapter. Children normally start to talk in their second year and build their vocabulary to around 14,000 words by age 6. However, vocabulary size varies considerably with the amount of direct adult–child interaction children get to participate in before age 3. Clark discusses the factors which facilitate the process of word learning, such as joint attention with adult speakers, the presumed contrast of new words with words already known, and adults' reformulations of the child's errors.

The topic of **Katharine Graf Estes'** chapter is the strategies that infants use to identify the words of the ambient language. While the written language may demarcate its words by means of spaces, the spoken language does not (usually) demarcate words by means of pauses. However, already by one year of age, children are paying attention to cues for word boundaries, such as phonotactic constraints and patterns of lexical stress. Indeed, the ability to extract words from continuous speech may be an important driver of lexical acquisition.

Reese Heitner draws attention to what he calls the 'inherent duality' of word learning. To be sure, the learner needs to recognize that a great variety of creatures, of different shapes, sizes, colours, and temperaments, can all be called 'dog'. But the learner also has to recognize that a great variety of pronunciations can be regarded as instances of the phonological form /dɒg/. Languages differ not only in the way they categorize the environment, but also in the way they categorize speech sounds. Heitner proposes that each process is able to bootstrap the other, in a kind of virtuous circle.

Paul Nation and **Frank Boers** address vocabulary from a pedagogical point of view—Nation on the words that a learner needs, and Boers on the strategies of teaching and learning words. Given the Zipfian distribution of word frequencies, a smallish number of word types make up a largish portion of the word tokens in a text. From one point of view, the most frequent words are the most useful, in that they guarantee coverage of a large amount of a text. On the other hand, frequent words—precisely because of their frequency—are the least informative. Nation discusses the criteria for drawing up lists of words which are likely to guarantee optimal understanding of different kinds of text. With emphasis on second- and foreign-language pedagogy, Boers warns against teaching strategies which might actually impede the learning of words, while extending the discussion to the learning of multi-word phrases and idioms.

The next topic is names. Names are special kinds of words, for a number of reasons. **John Anderson** discusses their status in the linguistic system. Do proper names have a meaning? Although some philosophers have argued that they do not—names attach 'directly' to their referents, without an intervening 'concept'—the prevailing view amongst linguists is, probably, 'yes'. Concerning their syntactic status, the prevailing view is that they are a kind of noun. Anderson points out that names do not have a uniform grammar. Overall, however, their syntactic properties overlap more with those

of pronouns than with (common) nouns, as befits their use for definite reference to individuals.

Mostly, a speaker has to abide by the sound–meaning conventions of the ambient language, following the Saussurean doctrine of the arbitrariness of the sign. Especially when it comes to the naming of infants, however, people are able to establish new labels (usually from a given name pool, and respecting the prevailing cultural conventions), which they perceive to be 'appropriate' to their referent, in one way or another. Naming practices around the world are the topic of **Benjamin Blount**'s chapter. He draws attention to the equation, prevalent in many societies, of the name with the individual. Hence, the use of names may be socially restricted, and after the death of the individual the name may become taboo (its mention conjuring up the deceased), thus necessitating the invention of alternative descriptive names. Even words phonetically similar to the deceased's name may be affected—a major factor in lexical renewal, in some societies.

Carole Hough introduces the field of onomasiology—the study of names, with special reference to place names, man-made structures, and features of the natural environment. In view of their conservative nature, these kinds of names are of special interest for the light they shed on settlement patterns. Hough concludes with some suggestive remarks on the sociolinguistic dimension of naming, for example, in the construction of community identity.

Word creation is also the topic of **Robert Kennedy**'s chapter on nicknames, their form, content, and function. Nicknames range from forms internally derived from formal names to items coined via more creative processes. In function, they can be used for reference or for address. Nicknames for males and females tend to have different patterns of phonemic structure, coinage, and semantic content. Like Hough, Kennedy draws attention to the sociolinguistic dimension of nicknaming, in that, for example, nickname coinage may reflect the relative power of coiners over recipients.

Cynthia Whissell addresses the factors which may influence the choice of a name—usually for a child, but also for pets, and even a novelist's choice of names for his or her protagonists. Her research shows, once again, that names are not arbitrary strings of sounds but can be felt to be appropriate to their subjects. She explores, amongst other things, the emotive associations of the sounds in a name and, related to this, the phonological differentiation of male and female names and changes in naming fashions over time.

Victor Raskin writes on verbal humour. He notes that words as such are not funny—the only exception, perhaps, being names (here again, the special status of names vis-à-vis other words is worth noting). But it is the words, in their appropriate combination, which make up verbal humour. Defending his theory that humour arises through conflicting scripts, he notes that jokes depend on the possibility of words and expressions being compatible with more than one script, a punchline making the ambiguity evident.

People's fascination with words finds its expression in all manner of word games, from Scrabble to palindromes and anagrams, and, of course, crosswords, especially the 'cryptic' crosswords so popular in Anglophone (and Dutch) cultures. **Henk Verkuyl**

offers a linguistic analysis, both erudite and entertaining, of word puzzling in English and Dutch.

Alison Wray sums up with remarks on the paradox of words. We think we know what they are, we believe that they exist, yet find it extraordinarily difficult to define them precisely and efficiently. It is not just that there are conflicting understandings of 'word' (word form, lexeme, lemma) and different (and not always consistent) definitions of word (phonological, orthographic, semantic, grammatical). The situation suggests a prototype account. Wray offers an alternative. Words, she says, in a striking image, are the bits that fall off when you shake an utterance. What this means is that we can certainly pick out words in an utterance—these are mostly the high-content nouns, verbs, adjectives, and adverbs. But then we are left with a residue of bits and pieces which can be assigned word status only with difficulty or by relaxing our notion of what a word is.

PART I

WORDS: GENERAL ASPECTS

CHAPTER 1

THE LURE OF WORDS

DAVID CRYSTAL

I have never met anyone who has not at some time been lured by words. The word is one of those concepts that seem to accompany us from the cradle to the grave. Parents are excited by (and never forget) the emergence of their child's 'first word'. At the opposite end of life, we pay special attention to 'last words'—and if their owners are famous, collect them into books. In between, we find 'words' entering idiomatically into virtually every kind of daily activity. We 'have words' when we argue. We 'give people our word' when we promise. We can eat words, bandy them, mark them, weigh them, hang upon them, and not mince them. People can take words out of one mouth, and put words into another.

Words operate within parameters of linguistic extremes. One such parameter is length. At one end, we see words as single strings of sounds separated by pauses, or of letters separated by spaces. They are the entities we identify when we do crosswords or play word games. At the other end, we make words equivalent to entire sentences or discourses. We talk about news travelling 'by word of mouth', and when we say 'a word in your ear', or we 'put in a good word' for someone, the utterances might be any length.

Another parameter is meaning. At one end we pay scrupulous attention to the meaning words convey, and many books have been written attempting to explicate what is involved when we say a word 'has meaning'. At the other end, there are contexts where the meaning is totally irrelevant. In a game such as Scrabble, the critical thing is to find a word that fits into the grid and is allowed by the official dictionary, rather than to know what it means. Most people have little clue about the meaning of some of the two-letter words they look up in the word lists, such as *en*, *qi*, and *ka*. The important thing is that they help the player to score well.

A third parameter is scope: 'words' can be equivalent to 'language', and then they evoke another contrast of responses, ranging from positive to negative. The proverbs of the world express both attitudes. On the one hand, we have the Arabic maxim 'Words draw the nails from the heart', the Bulgarian 'A gentle word opens an iron gate', and the Chinese 'A kind word warms for three winters'. On the other hand, we hear that 'Fair words butter no parsnips' (or 'cabbage', as it is in parts of south-east Europe), that 'Words don't season soup' in Brazil, and that in Germany 'Words are good, but hens lay eggs'.

The contrast here is variously expressed: between words and things, words and deeds, words and thoughts, words and ideas. Writers throughout history have pondered the relationship between these pairings. Two broad trends are apparent. One is to see words as inadequate representations of thoughts, poor replacements for actions, or a dangerous distraction from experiential realities. The other is to see them as indispensable for the expression of thoughts, a valuable alternative to actions, or a means of finding order in inchoate realities.

We see the first position at work when words are described as 'the small change of thought' (by French novelist Jules Renard in his *Journal*, 1988) or 'merely stepping stones for thought' (by Arthur Koestler in *The Act of Creation*, 1964) or 'the great foes of reality' (by Joseph Conrad in *Under Western Eyes*, 1911). Francis Bacon is in no doubt: 'Here therefore is the first distemper [abuse] of learning, when men study words and not matter' (1605, *The Advancement of Learning*).

On the other hand, for British poet and novelist Osbert Sitwell, 'A word is the carving and colouring of a thought, and gives it permanence' (*Laughter in the Next Room*, 1949); for American longshoreman philosopher Eric Hoffer, 'Action can give us the feeling of being useful, but only words can give us a sense of weight and purpose' (*The Passionate State of Mind*, 1954); and for science-fiction author Philip K. Dick, 'The basic tool for the manipulation of reality is the manipulation of words' (*I Hope I Shall Arrive Soon*, 1986). The writer of the Book of Proverbs is in no doubt: 'Deep waters, such are the words of man: a swelling torrent, a fountain of life' (18:4, *Jerusalem Bible* translation).

Several writers search for a middle way, stressing the interdependence of words and thoughts. This is German philologist Max Müller's view: 'Words without thought are dead sounds; thoughts without words are nothing. To think is to speak low; to speak is to think aloud. The word is the thought incarnate' (*Lectures on the Science of Language*, 1861). English poet Samuel Butler gives the relationship poetic form: 'Words are but pictures, true or false, design'd / To draw the lines and features of the mind' (*Satire upon the Imperfection and Abuse of Human Learning*, 1670s). And Bronislaw Malinowski provides an anthropological perspective, observing the way different languages express different visions of the world: 'The mastery over reality, both technical and social, grows side by side with the knowledge of how to use words' (*Coral Gardens and Their Magic*, 1935).

The metaphors increase and multiply, as writers struggle to find ways of expressing the relationship between words, on the one hand, and thoughts, deeds, and things, on the other. American historian Henry Adams: 'No one means all he says, and yet very few say all they mean, for words are slippery and thought is viscous' (*The Education of Henry Adams*, 1907). British novelist Aldous Huxley: 'Words form the thread on which we string our experiences' (*The Olive Tree*, 1937). An Indian proverb, much loved by Samuel Johnson: 'Words are the daughters of Earth, and things are the sons of Heaven'.

Some writers focus on what words actually do. Malinowski emphasizes their dynamic and pragmatic force: 'Words are part of action and they are equivalents to actions' (*ibid.*), and makes his point with some convincing examples: 'In all communities, certain words are accepted as potentially creative of acts. You utter a vow or you forge a signature and

you may find yourself bound for life to a monastery, a woman or a prison'. German novelist Thomas Mann adopts a social perspective, thinking of individuals: 'The word, even the most contradictious word, preserves contact—it is silence which isolates' (*The Magic Mountain*, 1924). British management educator Charles Handy also thinks socially, but on a grander scale: 'Words are the bugles of social change' (*The Age of Unreason*, 1991). Lord Byron gives words a mind-changing power: 'But words are things, and a small drop of ink, / Falling like dew upon a thought, produces / That which makes thousands, perhaps millions, think' (*Don Juan*, 1819–24). American columnist Peggy Noonan captures their emotional force: 'words, like children, have the power to make dance the dullest beanbag of a heart' (*What I Saw at the Revolution*, 1990).

It is the tension between the two perspectives that some writers see as critical, for it generates a creative impulse. American novelist Julien Green puts it like this: 'Thought flies and words go on foot. Therein lies all the drama of a writer' (*Journal*, 1943). For Ralph Waldo Emerson, 'Every word was once a poem. Every new relation is a new word' (*Essays*, 1844). T. S. Eliot describes the tension as an 'intolerable wrestle / With words and meanings' ('East Coker', in *Four Quartets*, 1944). It's the challenge that provides the lure, evidently, especially for the poets. For Thomas Hood, 'A moment's thinking, is an hour in words' (*Hero and Leander*, 1827). For American poet laureate Richard Wilbur, writing is 'waiting for the word that may not be there until next Tuesday' (in *Los Angeles Times*, 1987). And Lord Tennyson expresses the quandary thus: 'I sometimes hold it half a sin / To put in words the grief I feel; / For words, like Nature, half reveal / And half conceal the Soul within' (*In Memoriam A.H.H.*, 1850).

The whole situation is made more fascinating by language variation and change. Words and their meanings do not stand still, and perpetually offer new possibilities to the creative user. 'For last year's words belong to last year's language / And next year's words await another voice' (T. S. Eliot, 1944, 'Little Gidding', in *Four Quartets*). 'A word is dead / When it is said, / Some say. / I say it just / Begins to live / That day' (Emily Dickinson, *Complete Poems, c.*1862–86). And creativity extends to going beyond the existing wordstock. One of the most popular competitions I ever ran in my BBC radio series *English Now*, back in the 1980s, was the challenge to invent a word that the language needs. I received thousands of entries. The winner was the word we need when we are waiting by an airport carousel for our luggage, and everyone else's bags appear except yours. We are *bagonizing* (see Crystal 2006).

Word competitions are held every day, in some newspapers. How many words can you form from a string of letters? Which is the most beautiful word in the language? What is the longest word? What is the longest isogram (a word in which every letter appears the same number of times)? Can you make a humorous anagram out of the letters in the name of the prime minister? Can you write a poem in which every word contains the same vowel (a *univocalic*)? Can you write a text that doesn't make use of a particular letter of the alphabet (a *lipogram*)? Some people spend huge amounts of time on such tasks. Ernest Wright's novel *Gadsby* (1939), which uses no letter *e*, has 50 000 words. There seems to be a very fine dividing line between allurement and addiction (see Crystal 1998).

Exploring the history of words provides a further dimension. 'The etymologist finds the deadest word to have been once a brilliant picture', says Ralph Waldo Emerson (*Essays*, 1844), concluding that 'Language is fossil poetry'. The etymological lure is undoubtedly one of the strongest. I never cease to be amazed at the way word-books attract interest. Mark Forsyth's *The Etymologicon* topped the best-seller Christmas list in 2011. I have had more online reaction to my own *The Story of English in 100 Words* than to any other of my books: making a personal selection of words seems to encourage others to talk about their own favourites. Any listing of obsolescent words generates a nostalgia which can turn into a call for resurrection. A word can be given a new lease of life through online social networking—or a good PR campaign.

When in 2008 Collins decided to prune a couple of dozen old words from its dictionary—such as *agrestic*, *apodeictic*, *compossible*, *embrangle*, *niddering*, *skirr*, and *fubsy*—a cleverly managed campaign generated huge publicity for the next edition. Collins agreed to monitor public reaction, and to retain words that obtained real support. *The Times* took up the campaign (Adams, 2008). Celebrities agreed to sponsor the words: British poet laureate Andrew Motion, for example, adopted *skirr* (the sound made by a bird's wings in flight); British television personality Stephen Fry adopted *fubsy* (short and stout) and used it on his BBC panel/quiz show QI (i.e. Quite Interesting). A 'savefubsy' petition was launched online. An art exhibition featuring the words ran at the German Gallery in London. The result: both *fubsy* and *skirr* were reprieved, along with a few others, and all of the endangered words were retained in the online version of the dictionary.

Why do words get this kind of response? Henry Thoreau provides one answer (*Walden*, 1854):

> A written word is the choicest of relics. It is something at once more intimate with us and more universal than any other work of art. It is the work of art nearest to life itself. It may be translated into every language, and not only be read but actually breathed from all human lips;—not to be represented on canvas or in marble only, but be carved out of the breath of life itself. The symbol of an ancient man's thought becomes a modern man's speech.

Oscar Wilde provides another (*Intentions*, 1891):

> Words have not merely music as sweet as that of viol and lute, colour as rich and vivid as any that makes lovely for us the canvas of the Venetian or the Spaniard, and plastic form no less sure and certain than that which reveals itself in marble or in bronze but thought and passion and spirituality are theirs also, are theirs indeed alone.

I take these responses from the literary canon, and that is how it should be, for, as Ezra Pound affirms, talking about the writing of *Ulysses*, 'We are governed by words, the laws are graven in words, and literature is the sole means of keeping these words living and accurate' (quoted by George Steiner in *Language and Silence*, 1967). But the lure of words extends well beyond literature in its canonical form.

Perhaps it is the sheer number of words that provides the attraction. The size of a language's vocabulary is such that there are always new lexical words to explore. When learning a language, the task of mastering the pronunciation, orthography, and grammar is a finite task. There are only so many sounds and symbols, and only so many ways of constructing a sentence. But there is no limit to the words. I have elsewhere called vocabulary 'the Everest of language learning', to capture the challenge learners face; but even that metaphor is misleading, for vocabulary has no summit or end-point. To count the words of a language is an impossible task, and estimates of the number of words in, say, English, are always wide of the mark. Great publicity surrounded the claim made by an American agency, Gobal Language Monitor, in 2009 that the millionth word had entered the English language (Payack 2008). All they had done, of course, was devise an algorithm which was able to count up to a million. The English language has long had more than a million words.

The reason that the task is impossible is partly empirical, partly methodological, and will be discussed in detail later in this book. It is empirical because the English language is now used worldwide, and thousands of fresh words—and fresh meanings of words—are being introduced by the 'new Englishes' that have evolved. Dictionaries and word lists of Jamaican, South African, Indian, Singaporean, and over fifty other global varieties of English show the extent to which the emerging identities of recently independent countries is reflected in lexical innovation (see Crystal 2003). There are 15 000 words listed in a dictionary of Jamaican English, for example—that is, words used in Jamaica that aren't known globally. Many of them are colloquial or slang expressions, unlikely to appear in print, but that does not rob them of their status as words. Many of these words come and go like the tides. It is impossible to keep track of all of them.

The word-counting task is also complicated by methodological considerations. For what counts as a word? Are *cat* and *cats* one word or two? How many words are there in *flower pot* or *flower-pot* or *flowerpot*? Does an abbreviation count as a word? Do proper names count as words? Normally, we exclude names (such as *David* and *London*) from a word-count, assigning them to an encyclopedia rather than a dictionary; but we include them when they take on an extended meaning (as in 'The White House has spoken'), and there are many cases where we need to take a view ('That's a Renoir'). We need to be alert to these issues, to avoid making false claims. How many 'different words' does Shakespeare use? If we count *go*, *goes, going*, *goeth*, *gone*, etc. as separate words, the total is around 30 000 (it can never be a precise figure because of uncertainties over editions and what counts as part of the canon); if we count them as variants of a single 'word', *GO*, then the figure falls to less than 20 000. It is the need to clarify which motivated linguists to introduce a new term into the literature: *lexical item*, or *lexeme*. *Go, goes*, etc. are said to be variant forms of the lexeme *GO*.

The other counting task is more feasible: how many words do you, the reader of this book, know? If you have the time, all you have to do is go through a medium-sized dictionary and make a note of them. (Most people don't have the time, so they base their estimate on a sampling of a small percentage of the pages.) This would be only a first approximation, because not all the words you know will be in that dictionary—especially

if you are a scientist and have a large specialized vocabulary—but it will not be too far away from the truth. An English desk dictionary of 1500 pages is likely to contain around 75 000 boldface headwords. Most people find they have a passive vocabulary (i.e. the words they know) of around 50 000; their active vocabulary total (i.e. the words they use) is significantly less. Authors and word-buffs might have a vocabulary that is double this figure (Crystal 1987). One can nonetheless do a great deal with a relatively small active vocabulary, as the Shakespeare total illustrates—or the 8000 or so different words (excluding proper names) that are in the King James Bible.

Using this perspective, we now can quantify the lure of words. For if there are over a million English words waiting in the wings, and the best of us knows perhaps a tenth of these, there is an unimaginable lexical world waiting to be explored—unimaginable also because the vast majority of these words has more than one meaning. And they are all waiting in dictionaries to be used in new contexts. British novelist Anthony Burgess found a vehicular metaphor apposite: 'A word in a dictionary is very much like a car in a mammoth motorshow—full of potential but temporarily inactive' (*A Mouthful of Air*, 1992). American physician and essayist Oliver Wendell Holmes, Sr used a gustatory one: 'Every word fresh from the dictionary brings with it a certain succulence' (*The Autocrat of the Breakfast Table*, 1858).

Once again, looking to the poets helps us identify what it is that makes people talk about the 'magic' of words. Dylan Thomas, in his *Poetic Manifesto* (1961), picks up on the theme of quantity when he describes his first experience of reading:

> I could never have dreamt that there were such goings-on in the world between the covers of books, such sand-storms and ice-blasts of words, such slashing of humbug, and humbug too, and staggering peace, such enormous laughter, such and so many blinding bright lights breaking across the just-awaking wits and splashing all over the pages in a million bits and pieces all of which were words, words, words, and each of which was alive forever in its own delight and glory and oddity and light.

Sylvia Plath (in *Ariel*, 1965) describes the consequences of word choice. For her, words are 'Axes / After whose stroke the wood rings, / And the echoes! / Echoes travelling / Off from the centre like horses'.

So who should have the last word on *lurement* (first recorded usage, 1592, and marked 'rare' in the *Oxford English Dictionary*)? Or is it *luresomeness* (no attestation, yet, though there is a single record of *luresome* in 1889)? Perhaps we need a reality check from Samuel Johnson (in Boswell's *Life*, 1791): 'This is one of the disadvantages of wine, it makes a man mistake words for thoughts'. Or from Thomas Kyd (in *The Spanish Tragedy*, *c*.1589): 'Where words prevail not, violence prevails; / But gold doth more than either of them both'. Given the range of enthusiasms evident in the following pages, I opt for Evelyn Waugh, in a *New York Times* article in 1950: 'Words should be an intense pleasure, just as leather should be to a shoemaker'. Clearly, in this book, they are.

CHAPTER 2

HOW MANY WORDS ARE THERE?

ADAM KILGARRIFF

2.1 Introduction

Words are like songs. The ditty a mother makes up to help her baby sleep, the number the would-be Rolling Stones belt out in their garage, the fragment in a strange dialect recalled by the octogenarian, these are all songs. The more you look, the more you find.

The dictionary, as an institution, is misleading. The big fat book has an aura of authority to it, carefully cultivated by its publishers. On the back covers of the dictionaries on my shelf we have: 'Full and completely up-to-date coverage of the general, scientific, literary, and technical vocabulary', 'No other single-volume dictionary provides such authoritative and comprehensive coverage of today's English', 'The new authority on the world's language', 'The most comprehensive and up-to-date picture of today's English'. This is sales talk. They want to give their potential purchasers the impression that they have all the words in them (and more than their competitors). They also have numbers—always a bone of contention between the editorial department and the marketing department:

MARKETING:	How many words are there, for the press release?
EDITOR:	Well, there are 57,000 full entries.
MARKETING:	That's no good, Chambers and Websters both have far more.
EDITOR:	Well, we could count run-on items, the embedded compounds, phrasal verbs and phrases, that gets us up to 76,000.
MARKETING:	Still not enough, I'm sure you can do better, what about these bolded bits in examples?
EDITOR:	But they're just common expressions, they are not even defined.
MARKETING:	Are you forgetting who pays your salary? We need to sell!

There is even something strange about the syntax. We don't say 'Is it in a dictionary?', always 'Is it in the dictionary'. This is a triumph of marketing. Another word that works like that is *bible*. In the case of *bible*, it is reasonable to say that, at source, there is just one, and that all editions, in all languages, are just versions of that. The use of *the* for *dictionary* suggests some Platonic ideal that any published item is a more or less true version of.

Dictionaries have a variety of uses. Consider Scrabble. The simple role of the dictionary in Scrabble is to say if a string of letters is a word. It can only do that by having all the words in it. Alongside word games, there is resolving family arguments. A dictionary that does not allow a protagonist to say 'I told you so, it's not in the dictionary' is not worth the paper it is written on.

The impulse to document a language has much to do with comprehensiveness. 'Today's lesson is about glaciation. Let's start with gelifluction and move on to polynas,' says a character in a cartoon in the 'Horrible Geography' series (Ganeri 2002: 7). There are, indeed, a lot of words. All sorts of nooks and crannies of human activity have their own terms, not known to the general public but nonetheless, straightforwardly and unequivocally, words of English. *Gelifluction* does not have an entry in the largest dictionary I had available to check, the *Oxford English Dictionary*, although it does occur (apparently misspelt *gelifiuction*) in an example sentence for the related word *solifluction*. *Polynya* (note the difference of spelling) does have an entry. *Gelifluction* occurs just four times in a database of 12 billion words of text crawled from the web, *polynya/s* occurs 328 times, mostly with the second 'y', sometimes without it.

All this makes it hard to give a number. The primary reason is the sheer number of nooks and crannies of human activity that there are: how might we cover all of them? There are other reasons:

(a) Rules for making new words up. This is the province of derivational morphology and word formation rules (see Booij, this volume). Some specialisms even have their own rules for generating an unlimited number of specialist words (see l'Homme, this volume). The *Nomenclature of Inorganic Chemistry: IUPAC Recommendations* (Connelly 2005) is a collection of rules for naming inorganic compounds. If the rules are followed, then different chemists working independently will give the same name to a new compound according to its chemical composition, thereby reducing ambiguity and confusion. The rules sometimes give rise to terms with spaces in, sometimes to terms containing hyphens, brackets, numbers (Arabic and Roman), Greek letters, the + and – signs, and sometimes to long strings with none of the above. Examples (from Wikipedia) include *ethanidohydridoberyllium, bis(η^5-cyclopentadienido)magnesium, pentaamminechloridocobalt(2+) chloride, di-μ-chlorido-tetrachlorido-1κ^2Cl,2κ^2Cl-dialuminium*, and *Decacarbonyldihydridotriosmium*.

(b) Homonymy. Where there are two different meanings, when do we want to say we have two different words? Some cases are clear, e.g. *file* 'type of tool' and 'collection of documents', others less so (see Durkin, this volume).

(c) Multi-words. Do we allow in words written with spaces, like *all right*? When does a sequence of words turn into a single word, and *vice versa*? (See Wray, this volume, and Moon, this volume.)
(d) Imports. There can be uncertainty about the language that a word belongs to; when do words borrowed from other languages start to count? (See Grant, this volume, and Sorell, this volume.)
(e) Variation: when do two different spellings, or pronunciations, start to count as two different words?

First, we present a little data, and then we say some more about imports and variation.

2.2 A LITTLE DATA

The question 'How many words are there?' may be asked of any language. All the aspects discussed here relate to any language, though sometimes in different ways. Here, we mainly discuss English, with occasional reference to how different considerations play out differently in other languages.

enTenTen12 (Jakubíček et al. 2013) is a database of 12 billion words of English gathered from the web in 2012. The 12 billion is the number of tokens, not types: that means that the 547 million occurrences of *the* count as 547 million, not as just one, as they would if I was counting types. To put it another way, how many words are there in *dog eats dog*? There are two possible answers: three, if I am counting tokens, but two, if I am counting types. The question 'How many words are there?' clearly relates to types, not tokens.

Another ambiguity to draw attention to is between inflected forms of words and lemmas. Do *invade, invading, invades, invaded* count as forms of the same word, or as different words? If we say 'forms of the same word', we are talking about lemmas, or dictionary headwords. If we say 'four different words' we are talking about word forms. For English, the difference between the two is not so great, since very few lemmas are associated with more than four forms (the standard number for verbs, like *invade*), with nouns having just two (singular and plural). For many languages, the numbers are higher, sometimes running into hundreds. In this section all discussions are of word forms, largely because they are easier to count.

There are 6.8 million different types in enTenTen12 (including only items comprising exclusively lower-case letters, separated by spaces and punctuation). Their distribution is Zipfian: the commonest items occur far, far more often than most, and very many occur only once (see Sorell, this volume). Here there are 1,096 words that occur over 1 million times, and 3,745,668 words occurring just once. The distribution is broken down in Table 2.1.

At the 1,000,000 point (capturing words which occur more than 1,000,000 times) we have mainstream, core vocabulary words.

Table 2.1 Selected words from 9 frequency bands in the 12-billion-word corpus enTenTen12

Frequency band	No. of words	Random sample from lower edge of frequency band
1,000,000+	1096	active expensive floor homes prior proper responsible round shown title
1,000–999,999	60,789	ankh attunements diatom dithered limoncello mobilisations sassafras seemeth softgel uremic
100–999	109,362	alledge dwellin faceing finacee frackers neurogenetic sacralized shl symbole vigesimal
10–99	511,714	abbut arquebusses bundas carcer devilries feace hotu petronel taphophiles theaw
5–9	611,146	athambia dowter hazardscape humanracenow kernelled noatble producest stancher sullens trattles
4	307,309	boarwalk intercousre layertennis locutory meritest nonhumanistic pitiyankees scapularies starbeams uitrekenen
3	483,720	rokas faraa cuftucson cremosas topboard brahmanam samuebo messenblokken regenica
2	941,181	androgynized bolibourgeoisie lascomadres lowspot neoliberialism nonmorbid oapmaking projectst salesm whatsoevery
1	3,745,668	cirumscriptions digatel dramturgy figurability frelks inactivazed mixtore shunjusha teires wrider

At the 1,000 point we have:

(a) words from specialist domains, found in large dictionaries:
- An *ankh* is an Egyptian symbol usually meaning 'life' or 'soul'.
- A *diatom* is a single-cell alga.
- *Sassafras* is a species of tree with aromatic leaves and bark, and the extract drawn from it.
- *Limoncello* is an Italian alcoholic drink made from lemons. Also note that *limoncello* is on the margins of being a name, and in addition to 1,000 lower-case occurrences, there are 729 capitalized. On the borderline between regular words and names, see Anderson (this volume); also see restaurants section below.
- A *softgel* is an oral dosage form for medicine similar to capsules.

(b) inflected forms for familiar, if not specially common, words: *attunements, dithered, mobilisations*; also *seemeth*, an archaic inflected form of a common word; and *uremic* (relating to the disease *uremia*).

At the 100 point we have

- *vigesimal*, a number system based on twenty, present in the larger dictionaries.
- One simple spelling error, *faceing* (the target form was *facing* in all cases that I checked).
- *finacee*, target form: *fiancé, fiance, fiancée, fiancee*, depending on gender and the tricky business of how accented characters in imported words relate to English spelling. One thing is clear: the *a* should be before the *n*.
- Spelling errors mixed with old or other non-standard forms: *alledge, dwellin*. A mixture is a case where some of the instances are of one kind, e.g. spelling errors:

  > If you hear or read anyone in the United States assert or *alledge* that we have a democracy, a representative democracy or anything short of a kleptocracy

  while others are of another kind, e.g. an old form:

  > That the Debts either by Purchase, Sale, Revenues, or by what other name they may be call'd, if they have been violently extorted by one of the Partys in War, and if the Debtors *alledge* and offer to prove there has been a real Payment, they shall be no more prosecuted, before these Exceptions be first adjusted.

- A spelling error mixed with a foreign word: *symbole*.
- Inflected forms of derived forms of words: *frackers* is plural of *fracker*, 'someone who fracks', where fracking is a process of extracting gas from underground reserves, currently a politically and environmentally contentious topic; *sacralized*, past tense of 'made sacred' or 'treated as sacred'.
- A prefixed form: *neurogenetic* (where *neuro* is a mid- to low-frequency prefix).
- *shl*: a mixture of programming language command, url-parts, shortened *shall*, abbreviations.

At the 10 point, *abbut, arquebusses, devilries*, and *taphophiles* are recognizably words of English, albeit obscure and/or misspelt and/or inflected/derived forms, while the remaining six are not even that, and so it is as we carry on down to the items occurring just once. These are like the residue at the bottom of a schoolboy's pocket: very small pieces of a wide variety of substances, often unsavoury, all mixed together, often unidentifiable. One would rather not have to look into them too closely.

In sum, at the top of the list—at least the top 1,000—we have core vocabulary. By the time we have reached 60,000 we have obscure vocabulary and marginal forms. Another 100,000 items, and dictionary words are thin on the ground, though we still often have their inflected and derived forms, and their misspellings. After a further half million, half the items no longer even look like English words, but are compounded from obscure forms, typos, words glued together, and other junk, and so on down to *bolibourgeoisie, whatsoevery*, and *frelks*.

2.3 Imports

2.3.1 Restaurant English

As explained by Douglas Adams in *The Hitchhiker's Guide to the Galaxy*, a distinct form of mathematics takes over in restaurants at that moment when it comes to working out each person's contribution to the bill. Likewise, a distinct form of English. Let us make a linguistic visit to the grandest of our local vegetarian restaurants, Terre a Terre. A sample of their menu:

> Red onion, mustard seed, cumin crumpets with coconut curry leaf and lime sabayon, ginger root chilli jam and a fresh coriander, mint salsa sas. Served with thakkali rasam of tamarind and tomato, nimbu bhat cardamom brown onion lemon saffron baked basmati rice with our confit brinjal pickle.

The peculiar thing about this form of English is that, while the language is English, most of the nouns don't seem to be. They form a subtext to the history of the population itself, with:

- indigenous: *onion, mustard, seed, crumpet, leaf, root, jam, mint, pickle*
- fully naturalized: *cumin, coconut, curry, lime, ginger, coriander, tamarind, tomato, cardamom, saffron, rice*
- recent (within my lifetime): *salsa, bhat, basmati, confit, brinjal*
- novel: *sabayon, sas, thakali, rasam, nimbu*

A restaurant like Terre a Terre is at the leading edge of both culinary and linguistic multiculturalism. All sorts of other areas have their borrowings too: wherever we share artefacts or ideas or practices with another culture, we import associated vocabulary, for example in music (*bhangra, didgeridoo*), clothes (*pashmina, lederhosen)*, or religion *(stupa, muezzin)*. The question 'But is this word English?' feels narrow-minded and unhelpful. To give a number to the words of English, we would need to be narrow-minded and unhelpful.

2.3.2 Naturalization

A side-effect of importing words is: how much naturalization do we do?

There are assorted reasons—some good, some bad, most contentious—for having immigration policies and controlling which people are allowed into a country, and those policies are then strenuously policed. For words, some countries (famously France, with its Académie française) have, or have had, policies, and we may argue about the reasoning behind those policies being good or bad. They are also hard to police. English does

not have such a tradition. We welcome all sorts of words—but are often not sure how to say them or how to write them. The Nepalese staple lentil soup, in enTenTen12, is found as *dal baht, dal bat, dahl bat, dahl baht, dahl baat.* If the source language does not use the Latin alphabet, the imports will suffer vagaries depending on the source-language writing system and transliteration schemes. The *dal baht* case suggests a problematic mapping for the /a:/ sound between Nepali (usually written in Devanagari script) and English (written in Latin). Arabic usually does not write vowels, which is a main reason why there are so many options for how *Mohammed* is spelt in English. *Mohammed, Mohammad, Mohamed, Mahmoud, Muhammed, Mehmet, Mahmud, Mahmood, Mohamad, Mahomet*, and *Mehmood* all occur more than 1,000 times in the enTenTen12 corpus.

English can be seen as an imperialist language, currently the world's pre-eminent imperialist language, with its words marching into other cultures and taking over. English speakers, at least so far as their language is concerned, have no anxieties about being taken over and fading out. But the situation looks quite different from the other side. All over the world, languages are threatened and are dying, usually where, more and more often, bilingual speakers choose the alternative over their indigenous language (Crystal 2000). One part of this process is at the level of vocabulary, with speakers, even when speaking the indigenous language, using imports more and more often, either in preference to a local term or because there is no well-established local term. Many languages have government-supported terminology committees, charged with identifying, or creating, local language terms where as yet there is nothing well-established in the local language. Most often the non-local term is an English one.

The question 'Do we include this word in the count for *our* language?' is an interesting one for English—but for many languages it is also a political one, closely related to the very survival of the language.

2.3.3 Variants

Most English words have a single standard spelling; if the word is spelt in any other way, it is a spelling error. We all learnt that at school. We are troubled by the few exceptions: does *judg(e)ment* have an *e* in the middle? Answer: it can. There are also the transatlantic variants, including the *or/our* group (*colo(u)r, favo(u)r, hono(u)r*, etc.) and the *ise/ize* group. In our count of the words of the language, do we treat variants as different words?

Many languages have far less stabilized spelling than English, in particular languages which do not have a long written tradition.

There is interplay between standardization, pronunciation, and dialects. How far can a word stray and still be the same word? When I first came across *eejit*, when working with Glaswegians, I was puzzled as I felt I did not know the word. It was some months before I discovered it was a variant of *idiot.*

2.4 Conclusion

'How many words are there?' begs a set of further questions about what a word is: across time, across languages, across variation in meaning and spelling and spaces-between-words, across morphological structure. There is also the question of whether we are talking about the core of the language, or about the whole language including all the specialist corners where some small groups of people have developed their own terms and usages.

Dictionaries are no help. They have pragmatic solutions to the question that they face, namely, 'How many words shall we include?', and the answer varies from dictionary to dictionary. Whatever they say on the back cover is to be treated with the greatest scepticism.

'How many words are there?' is not a good question. A better question is 'How many words do various different speakers of a language (of various levels of education, etc.) typically know?' or, moving on from a purely academic perspective to one where the answer has practical implications, 'How many words do you need?' For that, we pass you on to the chapter by Paul Nation (this volume).

CHAPTER 3

WORDS AND DICTIONARIES

MARC ALEXANDER

3.1 Words and dictionaries

Dictionaries seek to be a core reference guide to words, presenting knowledge about a word as separate facts, such as its meaning, its pronunciation, or its history. They are some of the oldest forms of reference ever produced, with modern dictionaries tracing their bilingual roots to Sumerian-Akkadian word lists assembled a few millennia BCE (Snell-Hornby 1986: 208), with the first monolingual dictionary, the *Erya* or *Ready Guide*, being a Chinese collection of word glosses around 300 BCE (Yong and Peng 2008: 3). Their evolution has mirrored changes in world culture and technology across the ages, from being individualistic—and idiosyncratic—manuscripts to some of the first printed materials produced, then to centrally planned nationally authoritative tomes, to research outputs based on empirical scientific methodologies, and now to interactive and dynamic electronic databases. At each stage in this development, each change inherits, modifies, and develops the techniques of the past—taking always as their core the necessity for delimiting, cataloguing, and explaining words to a reader, but shifting often in approach, execution, and evidence.

Modern dictionaries, as the inheritors of this tradition, are born at the confluence of three contradictions: they are both scholarly and commercial, both judicious and impartial, and both artificial and yet grounded in natural language. These three contradictions explain much of what we need to know about both dictionaries themselves and their relationship with the words they contain: they are produced by expert scholars, but most often at the behest of publishers concerned with sales; they often strive to be descriptive and neutral, but users often require them to act as arbiters; they are unnatural things, but they are rooted in the natural use of words. These intersections explain why dictionaries are the unusual contraptions that they are, and each contradiction is examined in turn below, following an overview of dictionary history and structure.

3.2 A BRIEF HISTORY OF DICTIONARIES

The study of dictionaries, formally called *lexicography*,[1] requires, for much of dictionary history, the study of their compilers; because for many years the vast undertaking of dictionary compilation was undertaken by a single individual, so the judgements, biases, enthusiasms, and—in some cases—the personality of the compiler were irrevocably bound together with their work. Perhaps the most famous example of this is the 1755 *Dictionary of the English Language* by Samuel Johnson (now often referred to simply as *Johnson's Dictionary*, or just as *Johnson*). Although sensible and thorough in the main, Johnson's *Dictionary* contains a range of idiosyncratic choices, omissions, and definitions—he defines, for example, a *lexicographer* as 'a writer of dictionaries; a harmless drudge, that busies himself in tracing the original, and detailing the signification of words'. Johnson is somewhat exceptional in so forcefully displaying his personality and humour in his dictionary, although the history of lexicography is filled with biases, unfortunate choices, and unusual omissions.

Early dictionaries are very far from the modern conception of what a dictionary should be. Primarily taking the form of glosses and word lists, providing bilingual translations or explanations of difficult words, their sole shared characteristic is a focus on the word as a unit and in giving information about that word. Many consist of collections of manuscript glosses, where a reader or scribe would annotate the margins or line-spaces of an existing manuscript with word meanings or translations (Sauer 2009). This phenomenon represents a key aspect of the formation of dictionaries; the conceptual jump from annotating a word in context with a meaning to formally decontexualizing this word and treating it as an atomistic unit which can be defined separately is one which should not be underestimated from a modern point of view. Similarly, what we now call an early dictionary could also be easily categorized as a thesaurus (see Kay, this volume); although dictionaries are thought of nowadays as being characterized by their alphabetical order, many pre-modern works do not even countenance this sort of structure, instead preferring to list words thematically. Johannes Balbus, in his Latin dictionary of 1286, was sufficiently concerned with what he thought of as the innovation of alphabetical ordering to explain it in great detail, ending: 'I beg of you, therefore, good reader, do not scorn this great labor of mine and this order as something worthless' (Daly 1967: 73). Such ordering is counterproductive in many ways—it places unrelated items next to each other, refuses a user the opportunity to make connections between words close in meaning, and blocks easy browsing on one particular theme or topic—but it has one enormous advantage, that of a previously unprecedented ease of lookup. These two

[1] While *lexicography* generally refers to the creation and analysis of dictionaries, its companion term *lexicology* is somewhat broader, referring to the scholarly and linguistic study of words, such as that found in this Handbook. The distinction is somewhat blurred; lexicology often uses dictionaries as a source of evidence for the investigation of the lexicon of a language, and the lexicographical process of analysing words in order to create a dictionary can be easily seen as a process of lexicology.

innovations—abstraction from context and alphabetical order—primarily describe all dictionaries as we know them today; containers of facts about words, assembled according to alphabetical order.

Dictionaries began to homogenize in the Early Modern period, starting with the 16th century growth in bilingual dictionaries between European languages. In Paris, Robert Estienne produced his *Dictionarium, seu Latinæ Linguæ Thesaurus* in 1543, which was intended to be a monolingual Latin dictionary but whose first edition contained many definitions in French. The *Dictionarium* remains a significant work of Latin lexicography, and was reprinted and revised by later hands, often without the author's consent (Greswell 1833: 199–200). The classical languages dominated the beginning of this bilingual period; Sir Thomas Elyot's 1538 Latin–English dictionary, for example, was considered a major work of scholarship at the time, and most bilingual dictionaries linked a European language with either Latin or Greek.

Later English monolingual lexicography grew out of a 'hard words' tradition, beginning with what is usually recognized to be the first English dictionary, Robert Cawdrey's 1604 *A Table Alphabeticall*, which contained a list of 'difficult' or rare words each with a very brief English gloss. This sparked a long chain of plagiarism of English dictionaries over the next few hundred years, where authors took Cawdrey's work (itself based on earlier sources), added and changed some material, and then published their own dictionary, which in turn would be plagiarized by later authors (this rather endearing story of repeated and brazen theft is told briefly in Landau 2001: 48ff, and in more detail in Green 1996 and Considine 2008). By the time of Nathan Bailey's 1721 *Universal Etymological English Dictionary*, English had a dictionary containing not only hard words but also some of the core vocabulary of the language, alongside some etymologies and occasional invented examples of usage. There was some dissatisfaction with the state of these dictionaries, however, as they were highly uneven in quality and accuracy; following the publication of a large and authoritative French dictionary, an awareness grew in Britain that a better dictionary was required for the English language, which Samuel Johnson supplied in 1755.

The rapid evolution of English dictionaries can be seen by comparing some of their entries: *abash*, Cawdrey's second entry in 1604, was defined by him simply as 'blush', while Bailey in 1721 gives an etymology plus the definition 'to make ashamed or confound', and Johnson gives 'To put into confusion; to make ashamed. It generally implies a sudden impression of shame', alongside its part of speech, a cross-reference, a note on its use as a phrasal verb (*abashed at, abashed of*), and five illustrative quotations (in all, the entry totals 201 words, and is fairly short by Johnson's standards). The use of illustrations particularly marks out Johnson's work, with entries furnished with a range of quotations chosen from authors in high regard (usually from the 16th century, and frequently mistranscribed); this marks lexicography's move towards a respect for natural language, rather than invented examples and editorial guesswork.

The 17th and later centuries therefore saw the formation of major, rigorous national dictionaries, either officially sanctioned or considered generally authoritative by users: as well as Samuel Johnson's 1755 *Dictionary*, this period saw the publication of

the 1612 Italian *Vocabolario Degli Accademici della Crusca*, the 1694 *Dictionnaire de l'Académie française*, the 1780 Spanish *Diccionario de la lengua española*, and the 1880 German *Vollständiges orthographisches Wörterbuch der deutschen Sprache*, generally known as the *Duden* after its author. The New World was also represented here, with Noah Webster's 1828 *An American Dictionary of the English Language* following his 1783 *The American Spelling Book* (which deliberately introduced many spellings now considered characteristic of American English, such as *color, traveled, honor*, and *center*).

More ambitious multigenerational scholarly projects began later in the 19th century, their completion taking decades or even over a century, including the Grimm brothers' 1838–1961 *Deutsches Wörterbuch,* the Dutch 1863–1998 *Woordenboek der Nederlandsche Taal*, the (still in progress) Latin 1894–2050 *Thesaurus Linguae Latinae*, and the English 1884–1928/1933 *Oxford English Dictionary* (*OED*), originally titled the *New English Dictionary on Historical Principles*. None of these, as is common to dictionary projects since the age of Samuel Johnson, was ever predicted to take as long to complete as it eventually did (the *OED* was originally intended to take only a decade). To continue the example above, the *OED* uses 628 words to define and exemplify *abash*, including four sub-senses, 94 words of etymology, 14 spelling variants, and 394 words of quotations spanning five centuries.

Few new dictionaries of the same scope have been founded to match these projects, with later work instead focused on the updating and revision of existing multi-volume works; the *OED*, for example, currently has a third edition in preparation, its first full revision since its original publication, and the Académie française is working on a ninth edition of their *Dictionnaire*. More recent dictionary innovations rely on technological advances to reformulate their structure, style, or evidence base, although apart from the use of computers for editing and assembling sample citations, the same process is followed now as in the 19th century: assemble uses of a word, analyse and classify them, and then describe them. This process is explored further in the following section.

3.3 Types of dictionary and word facts

The process of analysing words depends, primarily, on what style of dictionary the analysis is necessitated by. Dictionary types vary along six main dimensions: their languages, variety, audience, timespan, format, and specialization.

3.3.1 Languages

A dictionary can be *monolingual, bilingual*, or the semi-intermediate category of a *learner's dictionary*—one which is monolingual but aimed at non-native speakers. While the normal market for non-native learners of a language is a bilingual dictionary (where each word meaning in the first language is given an appropriately chosen alternative

in the second language), a learner's dictionary is monolingual, but aims to define and give facts about a word using a restricted and simplified vocabulary alongside a detailed explanatory style. They are therefore appropriate for students of a language who are ready for a monolingual dictionary, but require more explicit assistance with words, and an avoidance of overly technical or difficult terminology in the definition. These dictionaries are often the source of many present-day innovations in dictionary-making, such as full-sentence definitions and more assistance with usage (for example, the *Collins COBUILD English Language Dictionary* defines *abashed* as 'If you are abashed, you feel embarrassed and ashamed', and notes that it is mostly found only in written texts).

3.3.2 Variety

Monolingual dictionaries can vary depending on which variety of the language they choose to represent; most choose the standardized form of the language, but some represent regional uses (such as dictionaries of Australian, New Zealand, Indian, or Canadian English, or the comprehensive *English Dialect Dictionary* and the *Dictionary of American Regional English*). On occasion, one dictionary can be adapted for an alternative variety, such as the British *New Oxford Dictionary of English*, later adapted to become the *New Oxford American Dictionary*.

3.3.3 Audience

Dictionaries can be aimed either at a scholarly or a commercial audience; if they are scholarly, then they are generally given more freedom to change their form to fit the data they describe, whereas commercial dictionaries tend to impose external restrictions on the size, scope, and other features of the dictionary (see section 3.4).

3.3.4 Timespan

A dictionary can be either *synchronic* (covering only one point in time) or *diachronic* (covering a span of time). The majority of commercial dictionaries sold are synchronic dictionaries focusing on the present day, while diachronic dictionaries tend to be aimed at scholars. These include the large multigenerational projects described in section 3.2, as well as such examples as the period dictionaries of Scots and English: the (in progress) *Dictionary of Old English* (600–1150); the *Dictionary of Middle English* (1100–1500), completed in 2001; the *Dictionary of Early Modern English* (1475–1700), begun and then abandoned, with its data folded into the revisions for the *OED*; and the complete *Dictionary of the Scots Language*, made up of the *Dictionary of the Older Scots Tongue* (early Middle Ages to 1700) and the *Scottish National Dictionary* (1700 to the present).

3.3.5 Format

A modern dictionary can be *print, print and electronic*, or *electronic-only*. Fewer dictionaries are now sold in hardcopy and more are sold as computer programs, online services, and smartphone apps. It is unlikely that a new major dictionary will now appear and not be available online; it is fairly likely that the majority of dictionaries will begin to move to electronic-only distribution.

3.3.6 Specialization

Not all dictionaries cover the general language; some specialize in their subject coverage. These include dictionaries of names, of medical or legal language, of abbreviations, or of other specialist areas, often aimed at expert practitioners needing a reference guide to obscure terminology in their professional area (such as *Black's Law Dictionary* or *Brewer's Dictionary of Phrase and Fable*). Many standard dictionaries contain frequently occurring specialist terms such as these, but very few will claim to be comprehensive in these areas. Some dictionaries take normal features of a standard dictionary and expand on them, such as the *Oxford Dictionary of English Etymology*, giving the derivations of words but no other information, or specialist pronunciation dictionaries, which are often aimed at learners or, in one notable case (the *Oxford BBC Guide to Pronunciation*), newsreaders.

These six features serve to mark each dictionary as separate from the others, and dictate its form, size, budget, and selection principles. Variation within these categories is usually comparatively minor, and consists of the ways in which various publishers and dictionaries distinguish themselves in the marketplace. It is therefore on the grounds of these features that the dictionary will allocate its resources and so decide the likely size and budget (and then determine how much of the dictionary can be revised from earlier editions, how much can be original, and how much has to be imported without alteration).

These final decisions then aid in the selection principles of the dictionary, those guidelines which determine which words should be included and which should not. These are not trivial matters; even Samuel Johnson, in his 1747 'Plan' for his dictionary, was anguished at how it was 'not easy to determine by what rule of distinction the words of this dictionary were to be chosen', concluding that such a rigid and inflexible rule was unworkable, because 'in lexicography, as in other arts, naked science is too delicate for the purposes of life' (2008: 20). As a result, selection principles tend to be guidelines rather than regulations, and the base principle followed by all modern dictionaries is that a candidate word for inclusion should both have definitively entered the language in question and have some currency amongst its speakers. Oxford University Press's modern English dictionaries, for example, have amongst their selection principles that a word must be found in a variety of different sources by different writers, must not be limited in its usage to one group of users, and should both have a fairly long history of use and be likely to be used in future. This last principle is notoriously difficult to predict but is nonetheless essential; dictionaries do not

record every minor neologism which is not likely to enter common currency, as otherwise they would be packed with useless 'nonce words'. This term was coined by James Murray, first editor of the *OED*, to describe words which are invented purely on the spur of the moment and used as one-offs, such as the verb *forficulate*, meaning to feel a creeping sensation, 'as if a forficula or earwig were crawling over one's skin' (*OED*), recorded only in use once but nonetheless included in the dictionary by Murray and marked as a nonce word. This is not the only principle that can be often broken; Oxford's criterion of not including a word limited to a single group of users is almost always violated in the case of technical vocabulary or slang which is characteristically used by one particular group—such as chemical engineers or teenagers—but with which the general public may well come into contact. Drawing a line in the sand to indicate where a word is said to be definitely 'in' the language is therefore somewhat superfluous, with some general rules discussed in a dictionary's front matter to give a broad sense of principle, but with editors normally given broad leeway to include whatever words they see fit.

Beyond deciding what new words to include, one of the major issues for a new dictionary is to what extent it will cover the core of the language—not all dictionaries give equal coverage both to basic, frequent words and to rare, difficult words. The core of English, for example, includes such very frequent words as *the, and, into, was*, and *of*, and these words are rarely looked up by native speakers as they are too basic for a user who is already competent in the language. These users will instead prefer a dictionary in the 'hard words' tradition, which will define and elucidate the periphery of the language in order to assist with composition, spelling, and comprehension on the one hand and, on the other, leisure activities such as crossword puzzles and word games (such as Scrabble) (see Verkuyl, this volume). Learners of English, by contrast, frequently find the very common and highly polysemous words of a language challenging, and therefore their dictionaries require significant resources to be devoted to explaining these words, with a corresponding lack of attention on rarer and more specialized words. Both types of user may benefit from the inclusion of encyclopedic content (such as that often found in US English dictionaries), but again a line must be drawn. Most British dictionaries reject encyclopedic content and only define generic words, such as *king* or *country*, and so assign the task of providing information about particular countries or kings to an encyclopedia; the *OED* has no entry for *Bhutan*, for example, while the *New Oxford American Dictionary* gives its location, population, capital, and official languages. The *OED* does, however, have an entry for *0898 number*, a British term for a premium-rate phone line, generally sexual, while neither dictionary has an entry for the *M25*, London's orbital motorway which is used frequently as a shorthand for the boundaries of the city.[2]

[2] The M25 is a good example of an entity which might not normally be considered a natural candidate for inclusion in a dictionary, and yet it is used frequently in the UK Parliament's *Communications Act 2003* (which stipulates that in the UK 'an appropriate range and proportion of programmes [must be] made outside the M25 area' in order to assist economic growth outside of London). There is therefore a strong case for considering this term enough a part of the language that it should be included in a dictionary. (I owe this example to John R. Taylor.)

Once coverage has been decided, the creators of a dictionary must turn to what goes into each individual word entry. Such entries contain, at a minimum, a word and a corresponding definition. Many contain more, including separate sub-definitions for different meanings of the same word, alongside pronunciations, parts of speech, lists of variant spellings, example sentences, usage labels, and word origins and etymologies. With regards to these elements, dictionaries inherit all the issues addressed in this Handbook, and from the outset, lexicographers have to engage with each problem that languages throw at them. One issue is a base form of a word, what dictionaries call a *headword* or *citation form*, and linguists call a *lexeme* (Lyons 1977: 18ff.). This headword is intended to collapse multiple forms of a word, such as *play, playing, plays, played*, into a single lookup item for ease of reference (in this English example, the unmarked *play*).

This collapse may be straightforward in a language like English, with its relatively straightforward prefix and suffix structure, but is rather more difficult in a non-alphabetical language or one where words have a complex internal structure. For example, monolingual Chinese dictionaries, such as the 1993 *Hanyu Da Cidian*, arrange their entries by *radicals* or *bùshǒu*, the meaning-bearing components of a character, while those marketed to foreigners are arranged alphabetically based on a transliteration of each component into the Roman alphabet. A Chinese arrangement by radical (or the provision of an index of radicals) involves ordering the radical list based on how many strokes the character takes to write—for example, two strokes for 人 'man/person', seventeen for 龠 'flute'. Each further character (made up of a radical with other elements) is then listed according to the remaining number of strokes, so that 人 has listed underneath it 介 'to lie between', requiring two additional strokes, and 企 'to plan a project' has the radical plus four strokes. Other arrangements, such as phonetic order, are also possible. Similarly, Arabic and other Semitic dictionaries use as their headword the *consonantal root*, the basic sequence of consonants which combine with other features to form a word. Languages with highly complex word structure, such as Turkish, Eskimo, and some Mesoamerican languages, generally structure their dictionaries around *morphemes*, or meaning-carrying units smaller than a word (for example, the English morphemes inside *novelization* are *novel, -ize, -ate*, and *-ion*, each with its own meaning; see Booij, this volume). Further elements are arranged under these roots in an intricate grammatical ordering. Beyond this, the system in Romance/Germanic languages of using the least-modified infinitive form of a verb, such as *play* or *achever*, does not apply in other languages, where the least-marked simplest form can be in another form—in Greek, for example, this is the present indicative 1st person singular. Some semantic units are not words at all, but rather opaque idiomatic phrases (such as *raining cats and dogs*), which must be given their own entry as their meaning cannot be reduced to that of their component words (see Moon, this volume).

Similar issues arise when choosing a pronunciation of a word (in countries with an accepted 'standard' or 'reference' pronunciation, there is often an issue as to how well this fits actual widespread usage); when dealing with exceptional word forms (so that, for example, the adjectival form *imperial* may not be best placed under its parent noun

empire, even when this sort of placement is normal practice for a dictionary); when choosing a part of speech for an entry (which requires the choice of a particular grammatical model); arranging the order of an entry (historical dictionaries tend to order senses by date, while modern dictionaries order by frequency or other criteria); and more. Even the style of phonetic transcription is a major consideration, with the majority of dictionaries using the standard academic IPA system (in which *late* is /leɪt/ and *far* is /fɑː/),[3] but many others, particularly US monolingual dictionaries, preferring instead a simplified 'respelling' system using standard alphabetical characters with various diacritics (so in *The New Oxford American Dictionary* of 2001, *late* is given as /lāt/ and *far* is /fär/, with footnotes on each page indicating that /ä/ is the sound in *car* and /ā/ in *rate*; this is simpler for native speakers but all but useless for early-stage learners who may not know the pronunciation of the reference words). In all these cases, the lexicographers working on the dictionary must aim to carefully resolve each of the issues which arise, based on the best of evidence, and in accordance with established dictionary policies, publisher policies, and the large number of considerations listed above.

With all this in mind, it is an amazing feat for any dictionary to be produced at all. But these are not the only issues which face dictionaries and their makers.

3.4 Scholarly but commercial

Almost all lexicographers must work with the tension between scholarly and commercial aims at the front of their mind.

Dictionaries are ultimately commercial because they are produced at the behest of publishers in order to sell to a public. Not every dictionary is entirely commercial; in the past, some were produced under the patronage of rich individuals (although patronage only goes so far—Johnson's *Dictionary* of 1755 had a patron but was mainly funded by a consortium of booksellers, as the patron contributed very little to the finances of the work; Johnson's feelings on the matter were clear both from his letters and his definition of *patron* as, in part, 'a wretch who supports with insolence, and is paid with flattery'). Many wholly scholarly dictionaries of modern times are funded by national academies, university presses, charities, or public research funders, who pay for the work on the grounds that the end result will be of significant public and scholarly benefit. But such funding is comparatively rare; large, multi-generational projects which take up significant resources over long periods are normally not attractive projects to funders, whose resources are often constrained and who have an uncertain ability to commit to future spending for decades to come. Scholarly dictionary projects instead often rely on patchy

[3] IPA stands for the International Phonetic Alphabet, overseen by the International Phonetic Association (also IPA); their homepage, with details of the Alphabet, can be found at http://www.langsci.ucl.ac.uk/ipa/.

portfolios of funding, seeking small grants for limited times from a range of funding bodies and charities, always with the uncertainty that future resources may not be forthcoming. The majority of new dictionaries produced are therefore commercial projects, funded by publishers on the expectation of future sales, usually with a firm budget and a tight scope. Such dictionaries are often commercially very successful, particularly in the lucrative non-native learners market.

However, the instincts of a well-schooled lexicographer are scholarly; they are experts who puzzle over large amounts of data regarding a word, and who take pride in setting themselves the task of teasing apart the fine variations of meaning that the word can realize, and in so doing create a well-turned explanation of the nature, structure, meaning, history, and usage of the word which they are examining. This expertise is hard-won, and the complexity of language does not lend itself easily to quick investigation (academics can often write whole articles or even books concerned with the meaning and usage of a single word, particularly if it is a culturally significant, contested, or semantically complex term). This does not always sit well with the commercial need for rapid turnover and efficient progress towards a publication date, and the tension between the scholarly and commercial needs of lexicography is often revealed in the occasionally fractious interactions between dictionary editors and their publishers or sponsors.

It is also notable that a publisher's budget will dictate to what extent a dictionary is wholly new and to what extent it is a reprint or reuse of existing material alongside additions and updates. It is very rare for dictionaries to be written from scratch; following the constant plagiarism of early dictionaries, in modern years almost all publishers with an existing dictionary will update their data rather than begin again wholesale. Those rare exceptions are where entirely new projects are undertaken in order to address shortcomings in existing dictionaries, such as the *OED*, or the 1987 *COBUILD* dictionary (see chapter 5 of Béjoint 2010 for more on this). Tight budgets often result in an increased tolerance for a lack of revision between dictionary editions, meaning that most new editions of a dictionary for the general market are now marketed on the basis of a much-trumpeted short list of very recent famous words, often those words which are seen to be part of the general cultural zeitgeist of recent years. This is the means by which an existing dictionary can highlight its new revision in order to establish itself as a sufficiently up-to-date reference source, and, in some cases, a justifiable purchase for owners of a previous edition. However, inclusion of such words can also easily date a dictionary, and often those entries which are inserted for the sake of a marketing push are themselves removed from later editions as they no longer meet the notability requirements many dictionaries place on their entries. The American Dialect Society's 2006 Word of the Year, the verb *to Pluto* (meaning to demote something, as was done to the former planet Pluto), was added to some dictionaries in that year and included on their marketing material—but eight years later is not to be found in recent editions of any major modern dictionary. This habit of featuring zeitgeist terms which violate a dictionary's standard inclusion policy is one notable example of commercial needs overriding a lexicographer's guidelines.

Scholarly and commercial interests also sit uncomfortably alongside each other when it comes to the inclusion of taboo words. A neutral linguist, examining the language, will say that all words are worthy of interest, and that swear words are generally of more interest than others (see Burridge, this volume); however, a publisher may be wary of including examinations of such words in a dictionary intended for general readership. This was a key concern in the mid-20th century: the 1961 *Webster's Third New International Dictionary* was intended to include an entry on *fuck*, but the publisher vetoed this and a reviewer later criticized its 'residual prudishness'; the 1966 *Random House Dictionary of the English Language* did not include the word, and so the *New York Times* review of the book discussed its 'stupid prudery […] in a dictionary of this scope and ambition the omission seems dumb and irresponsible' (Sheidlower 2009: xxx–xxxi). By contrast, the scholarly *Dictionary of Middle English* included *cunte* in 1961, although not all scholarly dictionaries followed the urge for maximal inclusion; the *OED* editor C. T. Onions supported the omission of *fuck* and other strong taboo terms from the first edition of the *OED* in the early 20th century, but it appeared along with a range of other taboo terms in the later 1972 *Supplement* (see also Burchfield 1972). Such omission was not universal, however. Onions argued for the 1933 *OED Supplement* to include *lesbianism*, which he described in a letter as 'a very disagreeable thing, but the word is in regular use and no serious Supplement to our work should omit it' (Brewer 2007: 49).

One final area for commercial interests to be prioritized by a dictionary is in the protection of a modern work's copyright. It is fairly easy to plagiarize dictionaries, given that they assemble facts about words in common use and are generally available in electronic form. It is therefore somewhat common in reference sources to insert a fake item, sometimes called a *copyright trap* or a *Mountweazel*, after a well-known trap entry in a 1975 encyclopedia; should this fake entry be found in another reference work, it could only have been taken directly from the inventing source rather than be the result of independent work. The most famous word of this sort is *esquivalience*, an entry inserted in 2001 in the *New Oxford American Dictionary*; the non-word was later found without attribution on the online *dictionary.com* resource, and then taken down from that site (Alford 2005). There are now, interestingly, some natural uses online of *esquivalience* (meaning the wilful avoidance of one's official responsibilities)—a consequence, explored in the following section, of the popular view of a dictionary as an absolute authority.

3.5 Impartial but judicious

The second contradiction is that dictionaries are both judicious and impartial. Expert lexicographers generally wish to act as other exploratory scientists do—like biologists, catching butterflies and examining their provenance, or cartographers, exploring new vistas and charting their extent. Lexicographers of this type are not concerned with value judgements, just as biologists do not speculate on whether a butterfly is morally dubious and cartographers do not solely chart those areas of which only the well-bred

approve. However, one of the primary uses of a dictionary is to assist users in finding words which are appropriate for a purpose, and in some cases to avoid words which are considered by other speakers to be offensive or otherwise inappropriate. This tension is frequently problematic, as a core tenet of modern descriptive language study is that it is concerned with describing all aspects of a language within its natural habitat, free of value judgements.

Nonetheless, many dictionary users wish a dictionary to be judicious and to act as an authority, describing what is and is not considered part of 'the language'. This perhaps arises from the privileged position of a dictionary in the educational system; dictionaries are often purchased by or for students in order to assist them with language learning and as an aid in composition (the sales categories for dictionaries, such as *Collegiate, Student*, or *School*, reinforce this). As Kilgarriff points out in this volume, dictionaries are usually referred to with the definite article; people say that things should be checked in *the* dictionary, not *a* dictionary. This 'implicit belief that the dictionary is one's linguistic bible' (Quirk 1982: 87) has become so entrenched that judges often use dictionaries in court cases to establish 'natural' meaning (Solan 1993). This desire is, on occasion, explicitly fulfilled; the modern prescriptive dictionary par excellence is that of the Académie française, which acts as the canonical authority on the use and vocabulary of French—but these works are rare.

The desire for prescription relies on the assumption that a dictionary can be more of an authority than is actually possible; dictionaries are necessarily imperfect because they are created through many compromises, by people working on often-imperfect data within a tight timescale, and often have problems with regards to the constant struggle to be up-to-date. Nonetheless, dictionaries are given an unusual privilege amongst reference materials. Should a person look up their location on a map and find features to be missing, they would believe the map to be incorrect; if they were to hear of a country and search fruitlessly for it in an encyclopedia, they will consider the encyclopedia out of date; however, if they were to look to a dictionary for a word they have just heard and find that the word is not in the dictionary, there is a significant chance that the user may conclude that the word is itself improper, to be avoided, or even that it isn't a 'real' word at all, rather than the more natural conclusion that the dictionary is flawed or out of date.

When dictionaries break away from this tradition, they may find themselves following the desires of lexicographers and of linguists, but acting against the wishes of users. This is particularly evident in the United States, where the 1961 publication of *Webster's Third New International Dictionary* was the source of much controversy when it adopted an impartial, descriptive approach to the language, including many slang words in common use but not generally approved of by prescriptive authors (the most famous being the contraction *ain't*). This resulted in virulent displeasure from many writers who desired a more judicious dictionary; the *New Yorker* published an unrestrained and full-throated attack on descriptivism in response, arguing against this 'trend toward permissiveness, in the name of democracy, that is debasing our language' and concluding, quoting *Troilus and Cressida*, that *Webster's Third* had 'untuned the string, made a sop of the solid structure of English, and encouraged the language to eat up himself'

(Macdonald 1962: 166ff.). *Life* magazine's review further complained that the dictionary had 'abandoned any effort to distinguish between good and bad usage—between the King's English, say, and the fishwife's' (quoted in Skinner 2012: 246–7, who with Morton 1995 presents a detailed account of this controversy). The wholly descriptive account was not, it became clear, what many users desired from their dictionary.

One frequent strategy to appease both the descriptive urge and the prescriptive desire is to describe word usages neutrally in the dictionary text itself, and then to include separate usage notes (often in a box or in a separate font in order to highlight their different status). *The American Heritage Dictionary of the English Language* (currently in its 5th edition, 2011) is well known for its usage notes, with a large panel of eminent writers forming its 'usage panel' (including, in its first edition, a number of the most vehement critics of *Webster's Third*). For words such as *epicenter*, the note states what the term 'properly' means and then supplies a figure for what percentage of the usage panel approves of various types of figurative extensions (for example, only 50% per cent of the panel 'accept' figurative uses of this word outside of its earthquake or explosive meaning—such as 'New York City is the epicenter of European immigration'). This is now the standard means of giving prescriptive judgements in a dictionary, but doing so at arm's length; lexicographers can avoid having to say, from their position as impartial experts, what is 'correct' and 'incorrect', or 'good' or 'bad', but rather can give evidence that, should words be used in certain ways in certain situations, that use may be judged inauspicious by some readers.

3.6 Artificial but natural

The final contradiction at the heart of dictionaries is that they are highly artificial but must be entirely rooted in natural language.

They are artificial because a dictionary by its nature unnaturally separates a word from its context. Words derive their meaning from their context and from their usage; a dictionary, however, must remove words from their linguistic situation and present them alphabetically, separated even from words in the same semantic field. This artificiality occurs throughout every part of the dictionary entry: words change their pronunciation depending on where they are in a sentence, yet a reference form must be given; they alter their meaning depending on the context of their use, yet dictionaries must strip away contextual influence while, in some cases, pointing out its influence. It is perhaps a sign of the naturalization of dictionaries, and of their significance within modern culture, that they are not more often viewed as the extreme oddities which they are.

This artifice must then sit alongside the need for dictionaries to reflect natural language, as far as the prescriptive/descriptive balance can allow. Importantly, a dictionary's evidence source should always be rooted in samples of natural language, as the dictionary seeks to represent either the word-stock of a language or of a well-defined subset of this (such as modern German words of a certain level of currency, or all

Middle Scots words in written use). This inventory is impossible to gather without linguistic evidence. Originally, this was accomplished by means of a reading programme, where employees or volunteers read through books, journals, newspapers, and other texts and transcribed examples of words in context onto paper slips. Enough of these slips, once assembled and collected by word, would then be given to a lexicographer, whose job it would then be to identify and define all the meanings found in the slips; their task would be complete when they had accounted for every meaning and every sense represented by the collection of evidence for each word. A reading programme is a very expensive and time-consuming undertaking, however, and can be unconsciously biased. Henry Bradley, an editor of the *OED*, once discovered that there were no citations for the word *glass*, used as the name of articles made of that substance, presumably because it was not exotic enough to be noted by readers (Mugglestone 2005: 41). Similarly, Jürgen Schäfer has noted the reliance of the *OED* citations on 'great literature, a fact which has proved a boon for the literary scholar [...] however, this policy leads to distortion' (1980: 13).

Preferable is the modern method of using an electronic corpus (usually shortened to just *corpus*), 'a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research' (Sinclair 2005: 16). These resources can contain huge amounts of language data, and are used by all modern dictionaries in order to provide the evidence from which their lexicographers draw conclusions about words. The choice of electronic texts to go into a corpus has now become a key concern of lexicography. While a corpus's creators must balance it to best represent what is being analysed, there will always be questions of both quality and inclusion: should internet sources be treated as valid standard language use, for example, or should *Jerry Springer* transcripts, soap opera scripts, and articles from the sensationalist US *National Enquirer* be included, to the detriment of literary novels and financial newspapers?

It is unfortunately not sufficient simply to take as much data as possible and trust that issues of quality and representativeness even themselves out. Corpora have grown enormously across the years; one of the first corpora of English was the Brown Corpus, a 1960/70s collection of 1 million words of American English, followed by the 20-million-word Birmingham Collection of English Text in the 1980s (later renamed the Bank of English, and used for the *COBUILD* dictionary), the 100-million-word British National Corpus in the 1990s, the billion-word Oxford English Corpus in the 2000s (used by Oxford University Press for their dictionaries), and the 155-billion-word Google Books corpus in the 2010s. As corpus size increases, the issue for lexicographers is no longer how to get enough raw examples of word use to create dictionaries, but rather how best to deal with the overwhelming number of examples which can be found. A search in the Brown Corpus for *inside* has 178 example results, a reasonable number to analyse, whereas the British National Corpus gives 13,449 results, and the Google Books corpus 10,057,203. Modern lexicographers must then take random samples from these corpora instead of aiming, as was once the practice, to account for all meanings found in a search; software is also used to automate some of the process of working out what is

happening with a word's context, its usage, and its common collocates. Corpus selection and design are therefore still of high importance.

The contradiction here lies in the unusual situation of seeking to remove words from their natural context while gathering as many samples of natural context as possible. The way in which these two, in practice, act in harmony with each other is perhaps the greatest art of the lexicographer; many recent advances in the quality and usability of dictionaries' treatment of words, particularly for non-native language learners, have arisen from digital techniques applied to corpora to assemble as much word data as possible.

3.7 The future

Where next? The likely future directions of dictionaries are predictably expected to rely on electronic formats; developments in this area have led to new lexicographic projects challenging every orthodoxy established over the past few centuries, with varying degrees of success. Perhaps most unusually, it has been hypothesized that corpora can replace dictionaries for most users—a dictionary entry may not need a carefully written definition, but can instead consist simply of a set of well-chosen sample sentences which themselves do the job of providing that definition. This has already been done on occasion by many older dictionaries, including large scholarly works such as the *OED*, where an editor has chosen to let an example stand as the entry's definition (see e.g. the *OED2* entries for *calicle* or *reel-bird*); the corresponding modern extension is that, provided with a sufficiently large corpus and the ability to automatically search for text patterns which characterize pre-existing definitions in texts, details of the meaning of any given word could theoretically be retrieved from the corpus automatically and dynamically (Hearst 1992 describes some of these definitional patterns). There is an attraction to this idea—after all, words are frequently defined in text, and when it comes to rare and unusual recent coinages, it is usually possible to find their original definition in an electronic copy of the text in which they were coined. The online resource *Wordnik* (http://www.wordnik.com) does this; a search for *decimate*, for example, finds example sentences with patterns like 'the word decimate' or *decimate* in inverted commas, which are more likely to give definitions than randomly chosen samples.

But lexicographers are not so easily replaced. As large amounts of raw data become more and more widely available, there is a corresponding increase in the desire of users to be guided appropriately through the overwhelming amount of variable information at hand. The same desire that causes users to ask for prescription and guidance through the thicket of words also requires expert intervention in the face of a wall of sample text—a well-tended and comprehensive set of clear, accurate definitions is preferable to something as easily confusing as natural language. A lexicographer, in this view, is not a stern and prescriptive schoolmaster but rather a curator of data; much as a museum curator is valued for their expert ability to present their exhibits for best comprehension, the value of a modern dictionary is in the expert summations which a lexicographer can provide

for the language user. While the surface form of dictionaries will inevitably change, usually to the advantage of users, the underlying function of the dictionary—that of having an expert examine language use and provide an authoritative encapsulation of facts about any given word—will probably not be lost.

Dictionaries also have a new lease of life in their use in digital and online applications. There is an increased desire to have computers *understand* texts as best they can, rather than simply search for words as they do now; it is in the comprehension of the meaning of texts that computers can extract useful information from vast data sources, including the web. Words are used by corpus linguists, computer programmers, and others as *search proxies*, simple means of hunting through texts for information about culture, history, language, literature, and the human experience. It is, however, rare that the word is of interest in its own right, rather than the meaning behind it; in this way a word is used as a proxy for its meaning, and in enhancing this proxy with dictionary data computers can begin to deal with information about meaning rather than just word forms. As computing grows beyond its infancy in this field, and as computers become able to tag and extract word meanings from texts, new prospects open up with regards to the aggregation and investigation of culture using semantic techniques, and it is through dictionaries and thesauri that this becomes possible. Here, the dictionary and the thesaurus have finally returned to their roots as not simply august reference volumes for the aid of the curious, but as comprehensive inventories of meanings, of words, and of the relationship between the two.

CHAPTER 4

WORDS AND THESAURI

CHRISTIAN KAY

4.1 What is a thesaurus?

The English word *thesaurus* derives from a Greek word meaning a store or treasure, and also refers metonymically to the place in which these things were kept, a storehouse or treasury. According to the *Oxford English Dictionary* (*OED*), it appeared in English in the Latin titles of reference books from 1565, and in English titles from 1736 (*OED* sense 2a).[1] Its first recorded use specifically to describe a thematically organized collection of words is in the title of probably the most famous thesaurus of all, Roget's *Thesaurus of English Words and Phrases* (*Roget*), first published in 1852. In modern use, the word can also be applied to an alphabetically organized dictionary of synonyms and antonyms, and to a classified list of terms used in various technical applications, such as indexing and information retrieval (*OED* senses 2b and 2c).The primary focus in this chapter will be on the thematically organized type of thesaurus exemplified by *Roget* and on the issues raised by attempting to present the diverse and expanding lexicon of English within a structure based not on the straightforward progression of the alphabet but on the much more unwieldy and controversial concept of a classification based on meaning.

4.2 Historical overview

In discussing the development of thesauri, scholars are hugely indebted to the work of the late Werner Hüllen, and especially to his monumental *English Dictionaries 800–1700. The Topical Tradition* (1999). So used are we in western societies to the dominance of the

[1] The online *OED* is under revision; dates or other information may have changed by the time this chapter is read.

alphabetical tradition, that it may come as a surprise to learn that the topical or thematic tradition is much older. Indeed, Robert Cawdrey, author of what is generally agreed to be the first English-English alphabetical dictionary, *A Table Alphabeticall, Conteyning and Teaching the True Writing, and Understanding of Hard Usuall English Words* (1604), felt obliged to explain this novel system in laborious detail to his readers, writing in his preface:

> If thou be desirous (gentle Reader) rightly and readily to vnderstand, and to profit by this Table, and such like, then thou must learne the Alphabet, to wit, the order of the Letters as they stand, perfecty [sic] without booke, and where euery Letter standeth: as *b* neere the beginning, *n* about the middest, and *t* toward the end. Nowe if the word, which thou art desirous to finde, begin with *a* then looke in the beginning of this Table, but if with *v* looke towards the end. Againe, if thy word beginne with *ca* looke in the beginning of the letter *c* but if with *cu* then looke toward the end of that letter. And so of all the rest. &c.

Modern dictionary users, being entirely attuned to the alphabet, might find such a level of instruction more appropriate when attempting to understand the mysteries of a semantic classification such as Roget's.

Hüllen (1999: 30–1) traces the existence of topical word lists as far back as the ancient civilizations of Egypt and China. These early compilations focused on the natural world and on the conditions and objects of everyday life, including plants and animals, buildings, kinship terms, and so on. Their purpose was to record knowledge and to pass it on to others. The chosen topics remind us of the close relationship between language and ethnography: much can be revealed about a culture by examining the concepts it chooses to lexicalize and record in writing. Hüllen comments that the Egyptian lists were 'lists of entities rather than lists of words', thus raising a fundamental issue in semantics—the relationship between objects in the material world and the words by which we name or classify them. Whether words should be regarded as names for things, or as signs representing the more abstract notion of mental concepts, has been much discussed in philosophy and linguistics over the years (see Riemer, this volume).

In Europe, the practice of glossing difficult words in texts, such as foreign or dialectal words or archaisms, can be traced back to ancient Greece (Hüllen 1999: 44). As time went on, such glosses were gathered together into wordlists or glossaries for ease of reference, at first related to particular texts, then as more generally useful independent lists divorced from the texts. In England, Latin texts with marginal or interlinear glosses in Old English (OE) are found from the 8th century onwards. Their primary purpose, as in other parts of Europe, was the study and teaching of Latin. Frequently compiled lists included such useful topics as the body and its parts, precious stones, medicinal herbs, and natural kinds such as animals, birds, fish, and plants. The practice of compiling glossaries continued during the Middle English period (1150–1500), when we also find glosses in Anglo-Norman and Old French, the languages of vernacular literacy at the time. Perhaps because of the disruption and development of society during this

period, increasing attention was paid to civil domains such as the church, society, arts and crafts, and the home.

In the 15th century, social changes such as the introduction of printing, and increased literacy and mobility among the population, created a demand for materials for learning vernacular European languages. These materials often consisted of multilingual thematic lists, with words from up to eight languages appearing in parallel. English, however, was a low-prestige language at the time, and was rarely included (Hüllen 1999: 105). The Renaissance period also saw the appearance of many new or translated works on technical subjects such as warfare, navigation, and horticulture, some of which were accompanied by thematic glossaries. As Hüllen notes (1999: 54), 'The beginnings of English lexicography, and indeed of the lexicography of other European languages, lie in glosses.'

From the 17th century onwards, the lexicographical focus in the English-speaking world has been very much on the ever-increasing size and sophistication of alphabetical dictionaries, with some notable exceptions, including John Wilkins' *An Essay towards a Real Character, and a Philosophical Language* (1668), and Peter Mark Roget's *Thesaurus of English Words and Phrases* (1852), which acknowledges a considerable debt to the earlier work. More recent thesauri, such as the *Historical Thesaurus of the Oxford English Dictionary* (*HTOED*, 2009), have followed somewhat different pathways in devising their classifications.[2] Various aspects of these works will be dealt with below.

4.3 What are we classifying, and why?

Words and their meanings are notoriously hard to pin down. Some word forms are polysemous, i.e. they have more than one meaning—sometimes considerably more; the verb *set*, for example, is divided into 126 main categories of meaning in the *OED*, not counting subcategories and phrases. In the opposite case, we may have several words, synonyms, which appear to refer to the same concept, but which on closer examination are subtly different, depending on where and when they are used. Issues like these can often be resolved by examining the words in context, but in the case of thesauri, words are usually removed from their contexts and treated as independent entities. We may thus find *beautiful, handsome*, and *pretty* grouped together as words referring to pleasing appearance, although a finer-grained analysis would show that we rarely, if ever, talk about *handsome babies* or *pretty sausages*. It is, of course, on the basis of collocations which *do* occur, such as *pretty baby* or *handsome man*, that dictionary definitions are formulated.

[2] Further details about *HTOED* and the procedures used in its classification can be found in its *Introduction* (2009: xiii–xx), and in Kay (2010; 2012).

Such problems were of concern to scholars in the 17th century for a variety of reasons. As a result of the explosion of knowledge during the Renaissance, and increased contact with modern languages, the vocabulary of English had expanded rapidly, leaving many people feeling insecure about their ability to use it. Learned words derived from Latin presented particular problems to the less well educated, and it was their needs that early dictionaries were designed to address. In the title page of his *Table Alphabeticall* of 1604 (see section 4.2), Cawdrey announced that his aim was to tackle these difficult words 'with the interpretation thereof by plaine English words, gathered for the benefit & helpe of Ladies, Gentlewomen, or any other unskilfull persons'. However, it was not only 'unskilfull persons' who were facing linguistic problems. At the same time as these Latinate words were entering English, Latin itself was losing the place it had held throughout the medieval period as the lingua franca of European scholarship. Academic works were increasingly written in vernacular languages, including English, and scholars struggling to develop a suitable style for these works had to confront problems such as the ambiguity and vagueness of words in natural language.[3] Scholars in the rapidly developing field of science, especially the biological sciences, were particularly exercised by such issues. They also engaged in discussion about the relationship between words and the things they designate in the external world, and about what goes on in our heads when we connect the two.

4.3.1 John Wilkins and universal languages

The most draconian solution to the inadequacy of natural language is simply to replace it with an artificial language, rather in the manner in which words for basic relationships are replaced by signs in symbolic logic. Theoretically at least, the inventor of such a language can control the situation by ensuring that each sign designates one, and only one, meaning; polysemy and synonymy are outlawed. Discussion and implementation of artificial languages with such an aim occupied a good deal of scholarly time in the 17th century.

Although the first person to attempt an artificial language in England was Francis Bacon, the most influential scholar in the field, and the one who took his plans furthest, was John Wilkins, a botanist whose primary interest lay in the classification of plants.[4] His work culminated in *An Essay towards a Real Character, and a Philosophical Language*, published in 1668. In his title, *character* means a set of written symbols, while *real* is used in *OED* sense 4b, 'Of written characters: representing things instead

[3] For a discussion of how such issues, and others raised in this chapter, have affected the development of specialist terminologies in more recent times, see L'Homme in this volume.

[4] A full account of Wilkins' work in the context of his time is given in Slaughter (1982). For a detailed account of his work on classification, see Hüllen (1999: 244–301, the sample tables in pp. 459–67, and 2004).

of sounds; ideographic'. Several of the *OED* citations refer to the Chinese writing system, knowledge of which had recently reached England. The language projectors, as they are sometimes called (because they had a project), realized the potential of such a system: just as it enabled speakers of mutually unintelligible languages like Mandarin and Cantonese to communicate in writing, so it could be used in a much grander way to enable communication among speakers of all languages, thus diluting the effects of the tower of Babel. The projectors, however, wanted to go beyond a universal language where signs represented words and to create one based on the 'things and notions' which occurred in nature; that is, a 'philosophical language'. As Slaughter writes, they aspired to a system that

> … directly referred not to words but to things or to notions of things, just as 2 refers to the quantity two and not the word *two, deux, duo*, etc. To say that one will invent a common language of ideographic signs presupposes, however, that there is a common set of notions which the ideographic signs are to represent. While these common notions are obvious in the indisputable case of numbers they become more problematical once we go beyond numbers. The language projectors, since many were also scientists, soon made it their business to set out or to discover precisely what those universal notions were.
>
> (Slaughter 1982: 1–2)

The second part of this quotation identifies one of the problems that beset any attempt at creating a universal language. Notions which seem to be so basic that they must be common to all humankind often appear to be less than universal on closer examination. Much 20th-century linguistic research on the world's languages, in areas such as colour or kinship terminology, proves just this point: what is meant by such apparently straightforward concepts as 'red' or 'sister' can vary widely across languages and cultures.

It is impossible in a few paragraphs to do justice to Wilkins or to the many contemporary scholars and fellow members of the Royal Society who shared his interests, such as Christopher Wren, John Ray, Isaac Newton, and Gottfried Leibnitz. Like his colleagues, Wilkins was concerned with finding order and structure in the apparent chaos of the external world. As a foundation for his philosophical language, therefore, he set about experimenting with a classification of plants, which involved identifying their distinguishing characteristics and organizing them into classes on that basis. From these basic units he built up a taxonomy, a hierarchical classification in which each level was related by a distinctive feature to those above and below. From these relationships, definitions of the classified things and notions could be constructed. To modern readers, used to elaborate scientific taxonomies such as the Linnaean classification of plants, such an arrangement may not seem particularly novel, but it was revolutionary in its day. Until the 17th century, such taxonomies as existed were essentially folk classifications, based on popular knowledge, such as whether plants were edible, or poisonous, or useful in medicine.

By the time his book was published, Wilkins had gone far beyond the domain of plants and devised a taxonomy that encompassed, in varying degrees of detail, all phenomena capable of observation. Robins observes:

> The Essay, which runs to 454 pages … sets out what purports to be a complete schematization of human knowledge, including abstract relations, actions, processes, and logical concepts, natural genera and species of things animate and inanimate, and the physical and institutionalized relations between human beings in the family and in society.
>
> (Robins 1967: 114)

Wilkins himself acknowledged that information was deficient in many areas, but defended his decision to publish on the grounds that he wanted to show the potential of taxonomically organized data.

Once the taxonomic tables had been established, the next stage was to assign symbols to them, rather in the manner of headings in a thesaurus such as *Roget*, but avoiding the ambiguity of natural language. Thus the letter <g> indicates the superordinate category 'Plants', and is followed by subordinate categories such as <ga> 'plants classed by flowers' and <ge> 'plants classed by seed vessels'. Three letters indicate a further subdivision, as in <gab> 'plant, herb/flower, stamineous', and so on for a further fifteen levels. These elements constitute the primitives or root words of the system, unambiguous units denoting a single idea out of which more complex units might be constructed (Slaughter 1982: 168–70). The final stage in the process is to assign to each primitive and composite form a symbol that will fix its meaning for all times and all natural languages. For composite forms, these utilize repeated signs, such as a semi-circle above the middle of a character indicating 'male' or a short vertical line on the left indicating metaphorical use (Robins 1967: 115).

4.3.2 After Wilkins

Interest in philosophical languages had virtually disappeared by the end of the 17th century, although universal languages such as Esperanto, usually based on existing languages, are a recurrent phenomenon, as is the search for components of meaning which can be combined to form the words of natural languages. Much of this kind of work has been done by anthropological linguists, who find it useful when analysing areas like kinship terminology to employ components such as plus or minus [MALE], [SAME GENERATION AS EGO], and [MATRILINEAL]. Componential analysis, as this process is usually called, assumes structured lexical fields, where words can be defined in terms of one another. It can be applied to specific languages or to a renewed search for deep structure features common to all languages. The former objective informs the thesaurus of New Testament Greek compiled by Eugene Nida and described in Nida (1975). Here a total of some 5,000 words yields around 15,000 different

meanings, classified under approximately 275 semantic domains tailored to the cultural context of the time. Another thesaurus-related use aimed at a particular language was the setting up of the twenty-six major categories of *HTOED*, based on the extraction of components from *OED* definitions of key words (Kay and Samuels 1975). A form of componential analysis also played a part in the development of generative semantics, as in Katz and Fodor's influential paper of 1967, 'Recent issues in semantic theory'. In this project, the aim was to formalize the components that distinguish the various meanings of a word such as *bachelor*, with a view to supplying a dictionary as part of the base component of a generative grammar. An ongoing universalist approach is represented by the work of Anna Wierzbicka, initially published in *Semantic Primitives* (1972), the first step in the development of her Natural Semantic Metalanguage, which offers a set of indefinable meaning units by means of which more complex terms in any language can be defined (Goddard, this volume).

Although Wilkins' work never passed into popular use, its very existence contributed to future endeavours. Roget, in the 1852 introduction to his own thesaurus, praised 'the immense labour and ingenuity' expended in the construction of Wilkins' tables, but considered the work 'far too abstruse and recondite for practical application' (2002: xxx, note 2). Ironically, this comment foreshadowed some of the reaction to his own work when it first appeared. A reviewer of the first edition of *Roget* in *The Critic* wrote:

> This is at least a curious book, novel in its design, most laboriously wrought, but, we fear, not likely to be so practically useful as the care, and toil, and thought bestowed upon it might have deserved.
>
> (cited in Emblen 1970: 272)

In fact, *Roget* did not really become a best-seller until a craze for crossword puzzles swept North America and Britain in the 1920s. During this period, his publisher, Longmans, was reprinting a run of up to 10,000 copies at least once a year (Emblen 1970: 278–81).

Roget nevertheless shared many of Wilkins' objectives and insights, noting that 'classification of ideas is the true basis on which words, which are their symbols, should be classified', and pointing out the advantages for improved communication among speakers of different languages of constructing a 'polyglot lexicon' using the *Roget* framework (2002: xxx). As Hüllen says, Wilkins did great service not only to science and theories of classification, but also to his native language by introducing 'the idea of a comprehensive and monoglot onomasiological dictionary of the language into lexicography (possibly without being aware of it)' (1999: 271–2).

Nor have scientists lost their interest in the classification of ideas through words. One of the most original volumes of recent years is Henry Burger's *Wordtree* (1984), described on its title page as

> A Transitive Cladistic for Solving Physical and Social problems. The dictionary that analyzes a quarter-million word-listings by their processes, branches them binarily

> to pinpoint the concepts, thus sequentially tracing causes to their effects, to produce a handbook of physical and social engineering.

Having found existing dictionaries 'overly humanistic', the editor turned to the language of technology, 'an increasingly important part of the mapping of any culture seeking to control its environment'. Over a period of twenty-seven years, he collected transitive verbs, analysed them into binary semantic primitives, and combined them to form a multiply cross-referenced hierarchy of lexical items, where each word is defined by a word from the level above, plus a differentiating component (*Wordtree*: 13–14). This is a book like no other, yet Burger's comment 'Each scientific revolution produces a somewhat different grammar and world-view' reminds us both of the relationship between thesauri and culture, and of the role that scholars from different disciplines can play in their development.

4.4 Synonymy

A somewhat less draconian approach to the problem of synonymy is the development of the synonym dictionary, a hybrid form, which combines the convenience of alphabetical order with the opportunity to consider a range of possibilities for expression in particular contexts. Despite their lack of theoretical interest for aficionados of thesauri, such works continue to be popular nowadays and appear on many publishers' lists, often with the word *thesaurus* in their titles. Full-scale thematic thesauri, on the other hand, are often daunting to users unfamiliar with their structures—an observation confirmed by the fact that thesauri from Wilkins onwards have included an alphabetical index for ease of reference.

The development of synonym dictionaries filled the gap between Wilkins and Roget, although the primary focus in the period was the production of ever larger and more comprehensive monolingual dictionaries, culminating in the beginning of publication of the *OED* in 1884. The purpose of these synonym dictionaries is not to unravel the intricacies of nature, but to offer opportunities for stylistic, and possibly social, improvement. In this they have something in common with the early monolingual dictionaries, as indicated by Robert Cawdrey's reference on his title page to 'ladies and unskilfull persons' (see section 4.3). Hester Lynch Piozzi, one of the early compilers of a synonym dictionary, and one of the first women to appear in the annals of lexicography, perhaps had a similarly delimited audience in mind when she compiled *British Synonymy; or, An Attempt at Regulating the Choice of Words in Familiar Conversation* (1794). As Hüllen points out:

> In the preface, Piozzi conforms to the limited role that the eighteenth century allowed a woman, even one of her panache. It is a woman's work, she writes, to direct the choice of phrases in familiar talk, whereas it is a man's work to prescribe grammar

> and logic. Synonymy, her topic, has more to do with the former than the latter, with the elegance of parlour conversation rather than with truth.
>
> (2004: 224)

Having married an Italian and lived in Italy, Piozzi also had a concern with the problems of foreign learners faced with stylistic choice. Rather than simply offer lists of words, she put them into articles establishing their collocations, as in the group containing *abandon, forsake, relinquish*, etc., where she writes:

> … a man *forsakes* his mistress, *abandons* all hope of regaining her lost esteem, *relinquishes* his pretensions in favour of another; *gives up* a place of trust he held under the government, *deserts* his party, *leaves* his parents in affliction, and *quits* the kingdom for ever.
>
> (cited in Hüllen 2004: 226–7)

One can imagine the parlour gossip which might have given rise to that scenario.

A similar model is followed in a later work, George Crabb's *English Synonymes Explained, in alphabetical order with copious illustrations and examples drawn from the best writers* (1816). This was a popular and influential book, running through many editions until the final one in 1953, and aspired to cover all the synonyms in the language in alphabetical order (Hüllen 2004: 254). As with other synonym dictionaries, such ambition entailed a good deal of cross-reference. The inclusion of quotations from the 'best writers', such as Addison, Dryden, Johnson, Milton, and Pope, was a practice that had been developing steadily in alphabetical dictionaries since the 17th century, and is, of course, a cornerstone of the *OED*. An example of Crabb's style is his article for *Dispel, disperse*:

> *Dispel*, from the Latin *pellere*, to drive, signifies to drive away. *Disperse* comes from Latin *dis*, apart, and *spargere*, to scatter, and means to scatter in all directions.
>
> *Dispel* is a more forcible action than to *disperse*: we destroy the existence of a thing by *dispelling* it; we merely destroy the junction or cohesion of a body by *dispersing* it; the sun *dispels* the clouds and darkness; the wind *disperses* the clouds or a surgeon *disperses* a tumour.
>
> (1916: 276)

Entries of this kind go some way towards dispelling (or dispersing) the criticism often levied against thesauri, that they give the user no means of discriminating amongst the words on offer. Such criticism was anticipated by Roget, whose thesaurus simply listed words under headings and, sometimes, subheadings. Addressing those who are 'painfully groping their way and struggling with the difficulties of composition', he assured them that their 'instinctive tact will rarely fail to lead [them] to a proper choice'

(2002: xx). As anyone will testify who has read a student essay that relied heavily on *Roget* for elegant variation of style, such an assumption may be over-optimistic. On the other hand, Roget's category 621 *Relinquishment* in the most recent edition (2002: 338) contains over a hundred verbs, comparing favourably in quantity at least with the much smaller number offered by Piozzi or Crabb.

Discrimination among lists of synonyms is a particular problem for foreign learners of a language. Roget's work is clearly aimed at native English speakers, even if their discriminatory powers are not as finely tuned as he suggests. Modern thesauri intended for learners, such as Tom McArthur's pioneering *Longman Lexicon of Contemporary English* (1981), give examples of usage as well as supplementary information, for instance about grammatical patterns. The *Oxford Learner's Thesaurus* (2008), despite its title, is a dictionary of 17,000 synonyms and antonyms, and includes short definitions as well as collocations and usage notes. Thus the entry for *beautiful* is divided into 1. a beautiful woman, and 2. a beautiful place. Synonyms for the former, based on frequency of occurrence in a corpus, are *beautiful, pretty, handsome, attractive, lovely, cute, good-looking, gorgeous, stunning, striking*. A usage note under *pretty* tells us that it 'is used most often to talk about girls. When it is used to talk about a woman, it usually suggests that she is like a girl, with small, delicate features' (53). *Plain* is given as the antonym for *pretty*, whereas that for *beautiful* is *ugly*. Overall, the learner is offered a good deal of assistance in making appropriate choices. Interestingly, right at the back of the book, there is a topic index, grouping the dictionary entries under thirty headings such as *the arts, conflict*, and *health* (903-12). The difference between a thesaurus and a dictionary is to some extent relative.

4.4.1 What is a synonym?

Dictionaries like those discussed above draw attention to the problems involved in discriminating amongst synonyms. Underlying such problems is the superficially straightforward question, 'What is a synonym?', to which the superficially straightforward answer would be that it is a word which 'means the same' as another word or words. This is a particularly pertinent question for English, which has accumulated large numbers of synonyms or near-synonyms as a result of its long history of absorbing words from other languages, and coining new ones from its own resources. Two key periods here are Middle English, where French was the predominant source language for borrowings, and the Renaissance, which saw an influx of words from Latin.

During much of the twentieth century, when structural linguistics was the dominant paradigm in many parts of Europe and America, the question of what constituted synonymy was widely debated. John Lyons, for example, proposed a distinction between a strict interpretation of the term, where synonyms enter into the same set of sense relationships, and a loose one, which covers the sorts of lists found in *Roget*, where some, but by no means all, meaning is shared (1968: 446–53). In fact, language has a natural

tendency to differentiate words that may originally have been synonymous, often by narrowing the meaning of one of them; there is no particular advantage, other than elegant variation of style, in having two or more words that mean exactly the same thing. In theory at least, it is possible to say everything in Wilkins' philosophical language, with its exclusion of synonyms, that one can say in a much larger and more unwieldy language like modern English.

Within cognitive semantics, defining synonymy is no longer such an issue. On the one hand, building on the notion of a fuzzy set, prototype theory allows for categories that contain both good and less good examples, as in a *Roget* list. On the other hand, individual meanings of words are defined within a framework of knowledge, rather than solely in terms of their relationships to one another. As Geeraerts writes:

> Cognitive semantics . . . takes a maximalist perspective on meaning, one in which differences between semantic and encyclopedic knowledge, or more generally, between semantics and pragmatics, are not taken as a point of departure. Giving up the distinction is relevant for the description of separate lexical items: it implies that it is no longer necessary to draw the borderline between strictly definitional and merely descriptive features.
>
> (2010: 222)

Exclusion of real-world knowledge is in fact an impossible position for anyone attempting to construct a thematic thesaurus (see further section 4.5). Information deduced from the words often has to be supplemented by information about the things or notions they designate. It would, for example, be difficult, if not impossible, to attempt a classification of Old English words for agricultural implements and their parts without any knowledge of what a plough or a mattock was like.

In passing, it may be noted that polysemy is not an issue for thesauri. If a word has ten different meanings, then they will appear in ten thesaurus categories as independent entities, related only through the alphabetical index.

4.5 The bigger picture

Some of the points in section 4.4.1 can be illustrated from *Roget* category 252 *Rotundity*:

> **N.** *rotundity*, rondure, roundness, orbicularity 250 *circularity*; sphericity, sphericality, spheroidicity; globularity, globosity, cylindricity, cylindricality, gibbosity, gibbousness 253 *convexity.*
>
> *sphere*, globe, spheroid, prolate s., oblate s., ellipsoid, globoid, geoid; hollow sphere, bladder; balloon 276 *airship*; soap bubble 355 *bubble*; ball, football, pelota, wood (bowls), billiard ball, marble, ally, taw; crystal ball; cannon ball, bullet, shot, pellet; bead, pearl, pill, pea, boll, oakapple, puffball, spherule, globule; drop,

> droplet, dewdrop, inkdrop, blot; vesicle, bulb, onion, knob, pommel 253 *swelling*; boulder, rolling stone; hemisphere, hump, mushroom 253 *dome*; round head, bullet h., turnip h.[5]
>
> (2002: 132)

Pairs like *sphericity*/*sphericality* or *rotundity*/*roundness* might be accepted as synonyms, but in the second pair there is a difference of register, with *rotundity* more likely to be used in formal or humorous contexts. *Gibbosity* and *gibbousness* are now mostly in technical use, and for many speakers would occur only in collocations such as *gibbous moon*. In the second set, *sphere* and *globe* might be considered synonyms, but *spheroid, prolate s[pheroid], oblate s*., *ellipsoid, globoid*, and *geoid* are not synonyms but hyponyms of *sphere* in that they refer to a particular sub-type. Thereafter, the classification wanders off into a miscellaneous collection of words referring to objects that just happen to be round—in Aristotelian terms, 'roundness' is an accidental property of such objects rather than an essential one. If people are asked where they would expect to find such words, they are likely to allocate *pea, onion*, and *mushroom* to a category of *Vegetables, football* to *Sport*, and *bullet* to *Armaments;* that is to lexical fields representing domains of use in the external world. In current *Roget*s, many words appear in both types of categories.

To some extent, *Roget* categories of the kind represented by 252 *Rotundity* are a historical accident. Early editions of his work focus on abstract vocabulary areas, which are more likely to attract synonyms than words with material referents. (*Pea*, for example, according to *HTOED*, has had only one synonym in its long history, and that was *roly-poly*, recorded twice from a single colloquial source in 1784.) Subsequent editors have added in large amounts of vocabulary for material objects, generally using the existing categories rather than, as might have been preferable, setting up new ones. Thus, because of both its original plan and its subsequent history, many people find Roget's classification less easy to use than he claimed in the passage from his *Introduction* below:

> In constructing the following system of classification of the ideas which are expressible by language, my chief aim has been to obtain the greatest amount of practical utility. I have accordingly adopted such principles of arrangement as appeared to me to be the simplest and most natural, and which would not require, either for their comprehension or application, any disciplined acumen, or depth of metaphysical or antiquarian lore (2002: xxii).

This statement draws attention to one of the many issues confronting anyone rash enough to embark on compiling a thesaurus from scratch: do you start with an *a priori* scheme of classification into which words can be slotted, or do you build up the

[5] In *Roget*'s layout, semi-colons separate minor divisions of meaning. Numbers with italicized headings, as in 250 *Circularity*, are cross-references to related categories. In cognitive semantic terms, these indicate the peripheral members of the category, as opposed to the clear members at its core.

classification from an examination of the lexical materials?[6] The first of these approaches was favoured by Wilkins and Roget, both of whom, as we have seen, had a desire to establish order in the world before embarking on lexical analysis. Indeed, Roget's most recent biographer suggests that his desire for order was an attempt to compensate for an insecure childhood, spent wandering around England with his over-anxious widowed mother (Kendall 2008). The *a priori* approach was also favoured by the major alternative to their systems proposed in the 20th century, Rudolph Hallig and Walther von Wartburg's *Begriffssystem als Grundlage für die Lexikographie: Versuch eines Ordnungsschemas* (1952). Although the scheme of classification and some sample entries are given in French, this is not a thesaurus of a particular language but a taxonomy of concepts into which, they claim, the lexicon of any language could be inserted, thereby enabling comparative lexical and cultural studies. Ullmann reports that the work caused considerable interest when it was revealed at the Seventh International Congress of Linguists in 1952 (1957: 314–15; see also Hüllen 1999: 18–21), but there is no record of its being used in its totality or of it having much effect on practical lexicography.

Schemes such as these perpetuate the idea that a universal system of concepts is discoverable (see section 4.3.2). In fact, such an objective is unlikely to be accomplished. As Lyons writes of Hallig and von Wartburg:

> It is difficult to justify, for English at least, even the highest-level tripartite division of the vocabulary into lexemes relating to the universe, to man, and to man and the universe; as it is difficult to justify, in terms of hyponymy and quasi-hyponymy, Roget's six main classes of lexemes, (i) abstract relations; (ii) space; (iii) matter; (iv) intellect; (v) volition; (vi) sentient and moral powers.
>
> (1977: 300–1).

In the case of Hallig and von Wartburg, the decision to place 'man' at the centre of the universe is sociopolitical, rather than linguistic. Humans and other living creatures have a great deal in common when it comes to the processes of physical existence: living, breathing, eating, sleeping, and so on. A thesaurus of English based on their separation would lead to huge amounts of duplication in listing the vocabulary for these processes.

It is often a sorrow for thesaurus-makers, and an argument against the universality of human conceptual structures, that one person's self-evident system of classification will be largely mysterious to others. Robert Chapman, for example, editor of the fifth edition of the American *Roget's International Thesaurus* (1992), made some radical changes to Roget's framework, re-organizing the highest level of the taxonomy into fifteen categories, starting with *The Body and the Senses*. He claimed that Roget's scheme:

> … does not coincide with the way most people now apprehend the universe. Casting about for a more fitting arrangement, I chose what I call a 'developmental-existential' scheme … The notion has been to make the arrangement analogous

[6] For further discussion of the structure of thesauri, see Kay and Alexander (forthcoming).

> with the development of the human individual and the human race … This seems to me 'the simplest and most natural' array in the mind of our own time.
>
> (quoted in Fischer 2004: 43; see also Hüllen 2009: 44).

Thesaurus-makers are nothing if not ambitious, but one is tempted to ask who 'most people' are, and on what grounds the editor speaks for them.

A more modest approach is taken in Buck's *A Dictionary of Selected Synonyms in the Principal Indo-European Languages* (1949). After criticizing *Roget* for its large groupings and consequent 'lack of coherence', he says that his own classification is by 'semantically congeneric groups', and continues:

> The particular order and classification adopted is not copied from others, but no remarkable merit is claimed for it … There will be much that is frankly arbitrary, both in the classification and in the selection of synonyms to be included.
>
> (1988: xiii)

His 'congeneric groups' are somewhat similar to lexical fields, starting with The Physical World in its Larger Aspects, with subcategories at three levels of delicacy, and concluding with Chapter 22, Religion and Superstition. Despite the fact that there is 'some recourse to Miscellaneous' as a category (1988: xiii), the classification is relatively easy to navigate.

More recent thesauri, such as *A Shakespeare Thesaurus* (1993), *A Thesaurus of Old English* (*TOE*) (1995), and the *Historical Thesaurus of the Oxford English Dictionary* (2009), have taken the second approach noted above and started from the lexical data. In so doing, they aim to construct the world-view from the lexicon, and to postpone the identification of things and, especially, notions until the words have been analysed. As the editors of *HTOED* explain:

> It was acknowledged from the start that each section should be allowed to develop its own structure. Within the general taxonomic framework [the 26 major categories], classifiers were given a free hand, being told simply to 'sort, sort, and sort again' until an acceptable structure emerged; in other words, the classification was 'bottom up' from the data rather than imposed 'top down'.
>
> (2009: xviii)

A similar procedure had already been followed for *TOE*, which served as a pilot study for the larger work. The editors go on to remark that the most successful classifiers were often people who combined linguistic and philological skills with some knowledge of the subject being classified, especially in the sections dealing with the external and social worlds. In this they show some affinity with the tenets of the *Wörter und Sachen* (words and objects) movement in early 20th-century Germany, of which Geeraerts writes: 'The principal idea is that the study of words, whether etymological, historical, or purely variational, needs to incorporate the study of the objects denoted by these words'

(2010: 24). Given the ethnographic nature of their interests, it is not surprising that this group concentrated on the vocabulary of the material universe and its culture. One of the challenges for the *HTOED* team was to extend this system to abstract concepts, where boundaries of both meanings and categories are less clear-cut, and the taxonomy tends to be much flatter (Kay and Wotherspoon 2005).

It is perhaps no coincidence that all three of the works mentioned above are historical thesauri, covering particular periods in English in the case of the first two, and the complete recorded history of English in the case of *HTOED*. World-views change with the passage of time, as well as with synchronic variational factors, making it impossible in both theory and practice to devise a classification where one size fits all.

Abbreviations

HTOED	*Historical Thesaurus of the Oxford English Dictionary*
OE	*Old English*
OED	*Oxford English Dictionary*
Roget	*Roget's Thesaurus of English Words and Phrases*
TOE	*A Thesaurus of Old English*

CHAPTER 5

WORD FREQUENCIES

JOSEPH SORELL

5.1 Introduction

Science fiction authors and filmmakers have long faced an inconvenient linguistic problem. How does one write realistic dialogue for extraterrestrials? C. S. Lewis in his *Space Trilogy* took into account that not all *hnau* (the term for 'sentient species' in Lewis's interplanetary lingua franca, Old Solar) necessarily share the same vocal apparatus as homo sapiens. The makers of *Star Trek* commissioned a linguist, Marc Okrand, to create deliberately non-human-like languages, rather than ask viewers to suspend disbelief as Klingons converse with each other in American English. This raises the question of how we would even recognize a transmission from E.T. as meaningful. The answer lies, surprisingly, not in understanding the meaning of the message, but in analysing the word frequencies.

The story behind this solution begins over a century ago with an astute observation by a French stenographer. Jean-Baptiste Estoup (1916) noticed that some words were not just more frequent than others; they towered far above the average by a factor of thousands. Harvard professor George Kingsley Zipf (1949) made the same observation and wrote extensively on this pattern of word frequencies. Today, this frequency distribution is known generally as Zipf's law.

The statistical tools for gauging the central tendency in a normal distribution turn out to be of little use when dealing with word frequencies in naturally produced texts. The vast majority of word types (the words in the vocabulary of the text) are less frequent than the mean (or average). In H. G. Wells's *The First Men in the Moon*, the novel that inspired Lewis's trilogy, the mean frequency is 9.7. Of the 7,179 word types in the novel, 89 per cent are less frequent than the mean. Rather than being near to the mean, the median frequency (the frequency of the type at the mid-point of the distribution, i.e. type number 3,589 or 3,590) is only 2.0. The mode (the most common frequency) for naturally produced texts is always 1.0. In *The First Men in the Moon*, 54 per cent of the word types are hapax legomena (types that occur only once in the text).

What makes this distribution of words interesting, though, is not just that there are a few oddly frequent words. Estoup and Zipf found that this highly skewed pattern of frequencies followed a markedly consistent pattern. The second most frequent word in a text occurred roughly half as many times as the most frequent word. The third most frequent word occurred around one-third as often as the most frequent word, and so on. To put it another way, if one lists the words in a text in descending order of frequency and then multiplies a word's frequency (f) by its rank (r), the product remains approximately constant (C).

$$r * f \approx C$$

As can be seen in Table 5.1, the first two types do not fit the pattern very well. This is actually typical and a potentially significant phenomenon that will be discussed later in

Table 5.1 Sample of word types and frequencies from H. G. Wells's *The First Men in the Moon*

Word type	Frequency (f)	Rank (r)	Product (C)
the	3759	1	3759
and	2680	2	5360
of	2396	3	7188
I	2193	4	8772
a	1796	5	8980
was	928	10	9280
on	407	20	8140
our	304	30	9120
an	232	40	9280
or	195	50	9750
through	96	100	9600
life	41	200	8200
anything	25	300	7500
feeling	19	400	7600
held	15	500	7500
master	7	1000	7000
clumsy	3	2000	6000
halfway	2	3000	6000
bedroom	1	4000	4000
zuzzoing	1	7179	7179

the chapter. At the bottom of the table, also notice that each type has been assigned an individual rank. One could argue that all types with the same frequency should have the same rank. One could assign each of the types at a given frequency the highest rank among that group or the lowest or an average. This would be logical, since *bedroom* and *zuzzoing* were assigned their particular rank in the table only because the list was also alphabetized after sorting for frequency.

In Table 5.1, ranks have been shown in even increments for clarity, but choosing the highest rank among each group of types with the same frequency would have given even more consistent results. In Fig. 5.2, words with the same frequency show up as a horizontal line of overlapping circles. The point furthest to the right, i.e. the highest rank in that frequency group, is typically a better fit to the expected distribution.

A chart of word frequencies is usually shown on a double-logarithmic graph, i.e. the first increment on both the *x* and the *y* axes is 1, the second 10, the third 100, etc. A non-logarithmic graph (Fig. 5.1) looks like a large letter L with a few very high-frequency words lined up tight against the *y*-axis and the rest of the words spread out along the *x*-axis. If the logarithm of each type's frequency is plotted instead, the points line up at a roughly 45° decline (Fig. 5.2).

This type of distribution is referred to as a power law in mathematics, and Zipf's law is the best-known example. This mathematical insight allowed Estoup to improve French stenography by choosing simpler and shorter symbols for the most common items. Zipf, a linguist, also found this frequency pattern in the words, and even the morphology, of Chinese, Dakota, Gothic, High German, Plains Cree, Yiddish, and many

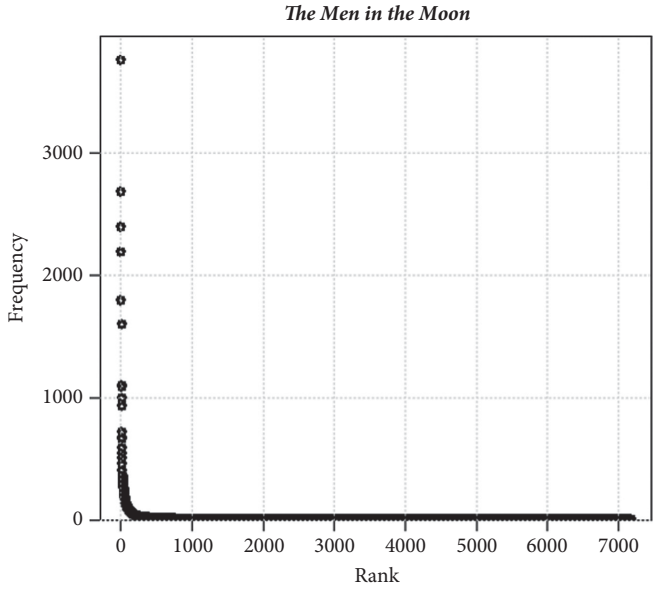

FIG. 5.1 Non-logarithmic Zipf graph of H. G. Wells's *The First Men in the Moon.*

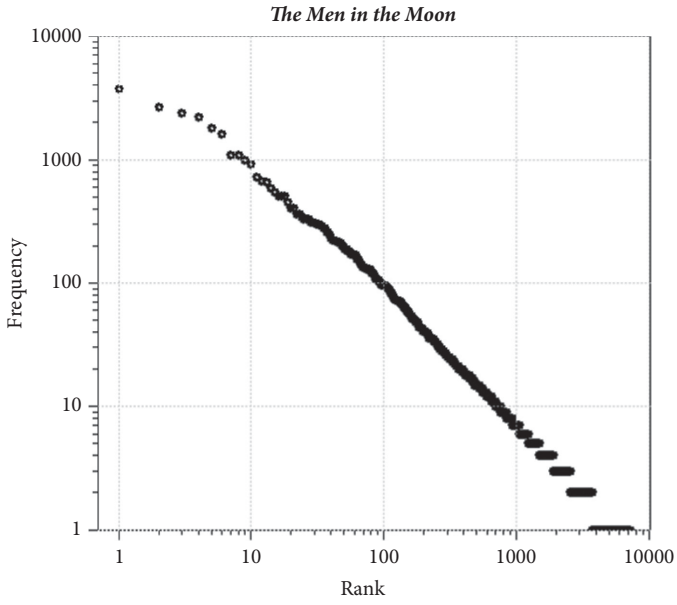

FIG. 5.2 Double-logarithmic Zipf graph of H. G. Wells' *The First Men in the Moon.*

other languages (Zipf 1949). He argued that this curious and apparently universal pattern arose as a compromise between speakers and hearers (1949: 20–21), each of whom wished to expend the least effort possible.

After Zipf published his *Human Behavior and the Principle of Least Effort* (1949), a young mathematician, Benoît Mandelbrot, was looking for something to read on the Paris Métro. He retrieved a review of the book from his uncle's wastebasket (Mandelbrot 1982: 346). His uncle, Szolem Mandelbrojt, was professor of mathematics at the Collège de France. The young Mandelbrot was inspired, and wrote his first academic paper on the connection between Zipf's law and thermodynamics (1982: 345). He envisioned himself becoming the Isaac Newton of linguistics, but he later concluded that Zipf's law 'is linguistically very shallow' (1982: 346; see also Ferrer-i-Cancho and Elevåg 2010; Li 1992; Miller 1957). This fateful encounter eventually led to Mandelbrot's famous work in fractal geometry, and though he described his work on Zipf's law as a 'self-terminating enterprise', some of his insights into Zipf's law have yet to be fully explored.

Mandelbrot (1953, 1982: 347) realized there was an important connection between Zipf's law, thermodynamics, and the recent work of Claude Shannon on information theory (1948). Understanding the efficient communication of information is a good first step towards understanding the mechanics behind Zipf's law. Efficient communication is, in fact, the common denominator that links better systems of stenography, the vocabularies of the world's languages, and spotting a potentially meaningful transmission from another world.