

For my parents

The Oxford Guide to Etymology

Philip Durkin

OXFORD
UNIVERSITY PRESS

The Oxford Guide to Etymology

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford ox2 6dp

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 in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

© Philip Durkin 2009

The moral rights of the author have been asserted
Database right Oxford University Press (maker)

First published 2009

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organization. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above

You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data

Data available

Library of Congress Cataloguing in Publication Data

Data available

Typeset by SPI Publisher Services, Pondicherry, India
Printed in Great Britain
on acid-free paper by
CPI Antony Rowe, Chippenham, Wiltshire

ISBN 978-0-19-923651-0

1 3 5 7 9 10 8 6 4 2

Contents

viii
ix

Acknowledgements

About this book

1 Introduction

- 1.1 What is etymology? 1
- 1.2 Some basic concepts: two example etymologies 3
- 1.3 Why study etymology? 22
- 1.4 What an etymologist does 31

2 What is a word? Which words need etymologies? 34

- 2.1 What are words? 34
- 2.2 How new words arise 43
- 2.3 Lexicalization 49
- 2.4 Examples of lexicalization 51
- 2.5 Apparent reversals of the process 56
- 2.6 Cranberry morphs 56
- 2.7 Which words need etymologies? 58

3 Are words coherent entities? 61

- 3.1 Variety in form and meaning: *poke* 'bag, sack' 62
- 3.2 Do we know precisely when a word's history begins? Can we assume continuity of use? 68
- 3.3 Homonymy and polysemy 74
- 3.4 How polysemy–homonymy relations can change 76
- 3.5 Merger (or near-merger) in form and meaning 79
- 3.6 Splits in word form 83
- 3.7 A case of merger followed by a split 86
- 3.8 Homonymic clash 88

4 Word formation 94

- 4.1 Issues concerning affixation 95
- 4.2 Synonyms, nonce formations, and blocking 103

4.3 Issues concerning compounding	107
4.4 Some other important processes	111
4.5 Arbitrary and non-arbitrary linguistic signs	123
5 Lexical borrowing	132
5.1 Basic concepts and terminology	132
5.2 What constitutes a borrowing from language X into language Y?	140
5.3 Motivation for borrowing: traditional explanations	142
5.4 Examples of borrowing because of 'need'	143
5.5 Borrowing of a new word when a new product of the natural world is encountered	145
5.6 Patterns of borrowing in the history of a language	149
6 The mechanisms of borrowing	155
6.1 Perspectives from contact linguistics	155
6.2 What is basic vocabulary?	157
6.3 Language shift	161
6.4 Borrowing within and between languages	164
6.5 Borrowings from more than one language	165
6.6 Continuing semantic influence and semantic interference	167
6.7 Multiple borrowings from the same source	169
6.8 How can we tell that borrowing has occurred?	169
6.9 Lexical borrowing and code-switching	173
6.10 Some conclusions from chapters 5 and 6	177
7 Change in word form	179
7.1 Two Germanic sound changes	180
7.2 Examples of English sound changes	182
7.3 Sporadic sound changes	195
7.4 Associative change in word form	197
7.5 Metanalysis	207
7.6 How regular are regular sound changes?	208
7.7 Examples of arguments based on word form	211
8 Semantic change	222
8.1 Meaning change is a common phenomenon	223
8.2 Polysemy and meaning change	225

8.3 Semantic polygenesis	228
8.4 Meaning change in a semantically complex word: <i>quaint</i>	228
8.5 Influence from other words	230
8.6 Some basic types of change	235
8.7 Is semantic change predictable?	243
8.8 Some practical examples	254
8.9 Arguments based on form and meaning contrasted	259
8.10 Etymology and extralinguistic factors	261
9 Etymology and names	266
9.1 How and why are names different?	266
9.2 Two villages called <i>Harvington</i>	269
9.3 Change in word form shown by names	271
9.4 Which language does a name belong to?	275
9.5 Names as evidence for lexis	277
9.6 Names as evidence for word meaning	279
9.7 Names as etymons	280
9.8 Names and non-linguistic history	281
10 Conclusion	284
<i>Glossary</i>	288
<i>Suggestions for further reading</i>	297
<i>References</i>	300
<i>General index</i>	313
<i>Index of word forms</i>	326

Acknowledgements

As my 'day job' I lead the team of specialist editors researching, writing, and revising etymologies for the new edition of the *Oxford English Dictionary*. As a result my first debt of gratitude must be to all of my colleagues, past and present, for their deep expertise, stimulating questions, and very good company, over (so far) fourteen years of highly enjoyable and rewarding collaborative research.

I am very grateful to the following people and institutions for kind invitations to give lectures and papers which drew extensively on draft chapters of this book: Derek Britton and the Institute for Historical Dialectology at the University of Edinburgh; Hans Sauer and the University of Munich; Ursula Lenker and the University of Eichstätt; Hans Sauer and Ursula Lenker and the Fifteenth International Conference on English Historical Linguistics, Munich, 2008. To the audiences at each of these papers I am grateful for many helpful suggestions and observations, and not least for confirming that there is an audience for a book of this sort.

I am enormously grateful to all of those who have commented on parts of this book in draft: Kathryn Allan, Joan Beal, Alan Durant, Anthea Fraser Gupta, Christian Kay, Peter McClure, Inge Milfull, Nicholas Sims-Williams; and to the following who were generous and brave enough to read drafts of the whole book: Anthony Esposito, Meg Laing, Roger Lass, Anna Morpurgo Davies, John Simpson, and Edmund Weiner. The book as it stands today has benefitted enormously from the keen eyes and still keener intellects of all of these people. Needless to say, all errors and omissions are to be laid, with complete justice, entirely and exclusively at my own door.

I am hugely grateful also to John Davey at OUP for his help, advice, and encouragement at every stage in the process of shaping, writing, and producing this book. Elmandi du Toit, Malcolm Todd, and Lesley Rhodes provided expert input on the production of the book.

I would also like to thank Kathryn Allan a second time, for putting up with me on the many days when even brimming pots of coffee were not enough to focus the mind and drive away irritation. And finally I would like to thank the dedicates of this book for tolerating a child's at times rather obsessive interest in very old documents and even older words.

About this book

Etymologies appeal to people with a very wide variety of interests and intellectual backgrounds. A very few people, such as myself, spend most of their time researching etymologies. A slightly larger number do so very occasionally. Many, many more people look at etymologies, but have never researched any themselves. Some people will never even have thought of etymologies as things which need to be researched. Particularly when etymologies are encountered in the compressed form found in many dictionaries, they can seem to be a given, rather than the (often very tentative) results of extensive research.

This book is intended for anyone who has taken the important first step of realizing that etymologies are the result of research, and would like to discover something about the nature of that research, and the principles and methodologies which underlie it.

I have attempted to frame this book so that it is addressed most centrally to someone who has an interest in historical linguistics, the study of how languages change and develop over time. Etymology is a part of this wider field, and anyone's understanding of etymology will be greatly enriched by at least some acquaintance with the broader concerns of the discipline as a whole. Readers who are entirely new to this field may find that they get much more out of this book if they read it in conjunction with one of the many excellent general textbook introductions to historical linguistics, such as Schendl (2001) or, in slightly greater depth, Millar (2007, which is a revised edition of Trask 1996) or Campbell (2004); for an excellent introduction to a wide variety of linguistic topics focussing on the vocabulary of English see Katamba (2005).

When deciding what to cover in this book and in how much detail, I have tried to pay particular attention to those areas which are important for etymology but which receive relatively little attention in most introductory books on historical linguistics. Nonetheless, I have also endeavoured to ensure that the book provides a balanced account of all aspects of etymology, especially for readers who are prepared to follow up references to fuller discussions of any topics which may be new or unfamiliar.

Most of my examples will be drawn from English, since this is the one language that any reader of this book will necessarily have some knowledge of. However, my aim has been to assume no particular knowledge about the history of the English language, beyond the explanations and further references given in the text. Drawing examples from the history of English also brings the advantage that I have in many cases been able to make use of very recent research for the new edition of the *Oxford English Dictionary* with which I have been involved personally.

There are no exercises, but at various points in the text I have listed further examples of the phenomena discussed, which readers can pursue if they wish in etymological dictionaries. Access to a good etymological dictionary of English would be of great benefit to anyone reading this book. In particular, access to the full *Oxford English Dictionary*, especially in its online version (www.oed.com), would be of especial benefit, so that many examples given here in summary form can be pursued in greater detail. (The dictionary can be accessed online via most institutional libraries and many public libraries.)

1

Introduction

1.1 What is etymology? 1	1.3 Why study etymology? 22
1.2 Some basic concepts: two example etymologies 3	1.4 What an etymologist does 31

1.1 What is etymology?

As we will see in this chapter, etymology can tell us that English *fríar* was borrowed from Old French *frere* ‘brother’, which in turn developed from Latin *frāter* ‘brother’. It can also tell us, perhaps rather more surprisingly, that Latin *frāter* is ultimately related to English *brother*, and that English *foot* is related to Latin *pēs* ‘foot’ and Armenian *otn* ‘foot’. Just as surprisingly, it can tell us that, in spite of the resemblance in form, English *care* and Latin *cūra* ‘care’ are definitely not related to one another, nor are Latin *deus* ‘god’ and Greek *theós* ‘god’. Etymology can also trace dramatic changes in meaning: for instance, English *treacle* originally had the meaning ‘medicine’, and comes ultimately from a Greek word which originally meant ‘antidote against a venomous bite’; *sad* originally had the meaning ‘satisfied’. How we trace such developments, and what they tell us about linguistic history, will be the topic of this book.

Etymology is the investigation of word histories. It has traditionally been concerned most especially with those word histories in which the facts are not certain, and where a hypothesis has to be constructed to account either for a word’s origin or for a stage in its history. That might be a stage in its meaning history, or in its formal history, or in the history of its spread from

one language to another or from one group of speakers to another. The term is also used more broadly to describe the whole endeavour of attempting to provide a coherent account of a word's history (or pre-history). As we will see in the course of this book, many of the basic methodological assumptions made in etymological research are the same regardless of whether we are looking at well-documented periods of linguistic history or at periods earlier than our earliest documentary records. Indeed, even someone who is primarily concerned only with attempting to solve hitherto unresolved difficulties of word history can only do so by building on the knowledge of many other word histories which have been much more securely established. For this reason, very many of the illustrative examples in this book will come from word histories which are very secure and not in any doubt, since they often provide the surest foundation for further investigation. Nonetheless, we will also look at some rather more difficult cases along the way.

Etymology forms part of the wider field of historical linguistic research, that is to say of attempts to explain how and why languages have changed and developed in the ways that they have. However, it does not concern itself exclusively with a particular linguistic level, as does for instance historical phonology (the study of speech sounds and of their deployment in ways which convey distinct meaning), historical morphology (the study of word forms as used to convey grammatical relationships), historical semantics (the study of the meaning of words), or historical syntax (the study of the meaning relations between words within a sentence). This is not to suggest for a moment that historical phonologists, morphologists, semanticists, or syntacticians never pay any attention to anything other than phonology, morphology, semantics, or syntax respectively. However, etymology is rather different, in that an individual word history will almost never be explicable in terms of only one linguistic level. Typically, some arguments or at least tacit assumptions about word form, probably involving issues of both historical phonology and morphology, will be combined with some arguments or assumptions about word meaning. In fact, etymology can be defined as the application, at the level of an individual word, of methods and insights drawn from many different areas of historical linguistics, in order to produce a coherent account of that word's history. One of the most exciting aspects of etymology is that this sort of detailed work on individual word histories sometimes throws up interesting results which can have a much broader significance in tracing the history of a language (whether that be with regard to phonology, morphology, etc.), especially when we can find

parallels across a group of different word histories. Additionally, it is often crucial that questions of (non-linguistic) cultural and intellectual history are considered in tandem with questions of linguistic history.¹

As well as using the word *etymology* as an abstract noun, we can also talk about *an etymology*, that is to say an account of a word's history. In the next section, we will look at two representative etymologies in some detail, as a practical way of introducing some basic concepts and at the same time some questions and issues which will concern us in much more detail later. The first example involves some very well-documented periods of linguistic history, while the second (which is rather more complex) will offer a first foray into historical reconstruction at a very considerable time depth. Concepts that we will explore include:

- tracing the linear history of a word
- change in word form
- change in word meaning
- borrowing
- genetic relationships between languages
- cognates
- comparative reconstruction
- sound change

1.2 Some basic concepts: two example etymologies

1.2.1 Example one: *friar*

The etymology of the English word *friar* can be sketched very crudely as follows:

Latin *frāter* 'brother'

develops into

Old French *frere* (modern French *frère*) 'brother', also 'member of a religious order of "brothers"'

which is borrowed as

Middle English *frere* 'friar'

which develops into

modern English *friar*

¹ For a short survey of previous definitions of the term 'etymology', accompanied by an adventurous attempt to formulate a fully adequate formal definition, see Alinei (1995).

The symbol 'greater than' is frequently used to stand for both 'develops into' and 'is borrowed as', and so we can represent the same development in a more 'shorthand' way as:

Latin *frāter* brother > Old French *frere* brother, also member of a religious order of 'brothers' > Middle English *frere* friar > modern English *friar*

Or we can reverse the arrows, and trace backwards from the modern English word. In fact, this is the style most frequently encountered in dictionaries and in most other scholarship:

modern English *friar* < Middle English *frere* friar < Old French *frere* brother, also member of a religious order of 'brothers' < Latin *frāter* brother²

The etymology of the Latin word could also be traced back a lot further than this, and can be linked ultimately with English *brother*, but this requires an acquaintance with some topics which we will investigate in section 1.2.4.

Obviously, this is a summary of a series of events in linguistic history. We will now examine each of those events in turn, and to do so we will require a little background at each stage. The Latin language is the direct antecedent of French. That is to say, French, like the other Romance languages (Portuguese, Spanish, Italian, Romanian, etc.), developed from Latin, albeit probably from a form of the language rather different from that reflected by the majority of our literary records. French also shows many borrowings and some structural influences from other languages, especially the Germanic language spoken by the Franks, but its basic line of descent is indisputably from Latin. In the vulgar Latin and proto-Romance varieties which eventually developed into French, the Latin word for 'brother', *frāter* (or more accurately its oblique case forms, such as the accusative singular *frātrem*) underwent a number of (perfectly regular) changes in word form, resulting in Old French *frere*. Old French is the term used to denote the earliest recorded stage of the French language, up to the early fourteenth century.³ Thus we have our first step:

Latin *frāter* > Old French *frere*

² Some scholars use the symbols '<' and '>' only to link forms related by direct phonetic descent, and use different symbols for processes such as borrowing or derivation, but in this book I will use them to link any two consecutive stages in an etymology.

³ Unusually, in this particular case, an intermediate step in the formal development of the Old French word is recorded in the very early Old French form *fradre* preserved in the *Strasbourg Oaths*, a unique (and very short) document from the year 842 which records (partly in Latin, partly in French, and partly in German) the oaths taken by Louis the German, Charles the Bald, and their followers during a time of conflict.

frere remained the basic word in French for 'brother', but it also acquired a secondary meaning denoting the (metaphorical) 'brothers' who belonged to various religious orders. This usage in French followed similar use of *frāter* in medieval Latin.⁴ The word was then borrowed into English from French. This happened in the Middle English period, the stage of the English language from roughly 1150 to 1500. More accurately, the word was borrowed from the Anglo-French variety of Old French which was used in England in the centuries after the Norman Conquest.⁵ The usual form in Middle English, *frere*, matches the French form exactly, and the pronunciation is likely to have been almost identical in Anglo-French and in Middle English. However, in Middle English the meaning is much narrower, showing only the religious sense and occasionally one or two other metaphorical uses. Thus we have our second step:

Old French *frere* brother, also member of a religious order of 'brothers' > Middle English *frere* friar

It is very common for a borrowed word to show only a very restricted and possibly rather peripheral portion of its meaning when it is borrowed into another language. In this particular instance, it is easy to see why (Anglo-)French *frere* was not borrowed into English with the much more basic meaning 'brother': the word *brother* (inherited from the Old English period, and from the Germanic antecedent of English before that) already had that meaning and was in common use, and even in the Middle English period, when very many words were borrowed from French into English, it is relatively uncommon for words with quite such basic meanings as this to be borrowed in place of native words. We will look at this issue in more detail in chapters 5 and 6. In fact English *brother* also had the meaning

⁴ The macrons which indicate vowel length in forms like classical Latin *frāter* are not normally given when citing Latin forms from later than the classical period, although this does not necessarily indicate any change in the vowel length in particular words.

⁵ In this book I use the term 'Anglo-French' to denote French as used in England (and elsewhere in Britain) in the centuries following the Norman Conquest. Scholarly practice is divided in this area: 'Anglo-Norman' is often used to denote this variety (as in the title of the *Anglo-Norman Dictionary*), but increasingly the broader term 'Anglo-French' is used instead, in order to reflect better the varied inputs from different varieties of Continental French which occurred both immediately after the Norman Conquest and in the subsequent centuries: for a useful discussion and further references see Rothwell (2005). For convenience, where a form or meaning belonged to both Insular and Continental French I use the style (Anglo-)French.

'(fellow) member of a religious order' in the Old English period on the model of use in Latin, and this meaning continued in the Middle English period (as it does today), reinforced by the similar use in both Latin and French. When *frere* is first found in Middle English it duplicates this meaning, as well as showing the more specialized meaning 'member of one of the mendicant orders (chiefly the Franciscans, Augustinians, Dominicans, and Carmelites, as opposed to the non-mendicant Benedictines, etc.)'. By the end of the Middle English period a process of semantic specialization took place, with *brother* used in the general sense 'member of a religious order' and *friar* in the narrower sense 'member of one of the mendicant orders'. Thus we might say that the borrowing filled a lexical gap in the vocabulary of English, providing a word specifically for 'a member of one of the mendicant orders', although we should perhaps be slightly cautious about such assumptions, since the same gap remained unfilled by any single word in French, even though the two languages were being used in very similar societies. Indeed, Anglo-French and Middle English were being used in precisely the same society. (See section 5.6 for discussion of the different functions of each language.) As we will see later, we can often run into problems of this sort when we attempt to explain word histories in functional terms, although this does not necessarily mean that the attempt is not worthwhile.

In its development from Middle English to modern English the word did not show any further change in meaning, but it did show an unusual change in form. The usually expected modern (British standard) pronunciation of a word which had the Middle English form *frere* would be /fri:ə/ (compare *here*, *deer*) but instead we find /fri:ə/. The same development is found in a small number of other words such as *briar* and *choir*. It probably shows a sporadic phenomenon of vowel raising before a following /r/.

Summary so far We can trace the history of a word's sound and form. In doing so we are looking for regularity, i.e. developments which are the same as those which happened to the same sounds or combinations of sounds in other words. Where something unexpected or irregular has happened, as with the development of /fri:ə/ rather than /fri:ə/, we will want to find parallels, such as *briar*, etc. Ideally we will want to find an explanation for this as well.

The meaning of the word can also be traced historically. We can see how the meaning broadened in Latin and French, but how the English

borrowing showed only a very narrow component of the donor word's meaning. We can also see how this borrowing fitted into a set of meaning relations with existing words in English (specifically *brother*). The meaning history of this word also shows the importance of factors from non-linguistic history: if we did not know something about the history of the religious orders in medieval Europe we would have considerable difficulty in explaining the historical development in the meaning of this word.

1.2.2 Example two: *sad* from modern English to proto-Germanic

For our next example we will start with the present day and work backwards. Modern English and Middle English *sad* show the reflex or linear historical development of Old English *sæd*. The symbol æ which occurs in the written form of this word and of many other Old English words (and some early Middle English ones) represents a front vowel phoneme /a/ (perhaps in fact [æ] rather than [a]) which in Old English was distinct from the back vowel /ɑ/, represented by a. (Its italic form æ is unfortunately very similar to that of the ligature œ, which can sometimes lead to confusion for the unwary.) We could represent this word history as Old English *sæd* > Middle English *sad* > modern English *sad*, but this would be rather artificial, since what we in fact have is a continuous history across all periods in the history of the language.

If we turn to the word's semantic history, a basic dictionary definition of the word *sad* as typically used in modern English is:

Of a person, or his or her feelings, disposition, etc.: feeling sorrow; sorrowful, mournful.

This meaning is first recorded *a*1300 (which stands for 'ante 1300', that is '1300 or a little earlier').⁶ A similar basic dictionary definition for the word's earlier meanings would be:

⁶ Some scholars use 'ante' in the more literal sense 'before', but most, including most dictionaries, use it in the generally more useful sense 'this date or a little earlier'. In this book the dates given for English words, forms, and senses are normally those provided by the *OED*. For words from other languages the data I give is generally drawn from the standard historical or etymological dictionaries of each language. Glosses and definitions of English words are normally based on those in either the *OED* or *The Oxford Dictionary of English* except where otherwise noted, although I have frequently shortened or otherwise adjusted them.

Having had one's fill; satisfied, sated; weary or tired (of something).

If we consider the likely historical development of these meanings, we can hypothesize that the meaning 'weary or tired (of something)' developed from 'satisfied, having had one's fill (of something)', hence showing a metaphorical, narrowed, negative meaning; compare the modern English idioms *to have had enough of something* or *to be fed up with something* for similar developments. Subsequently the sense 'weary or tired (of something)' broadened again (but still with an exclusively negative sense) to 'sorrowful, mournful' in general. Hence we can hypothesize that a meaning development occurred with two main steps:

satisfied, having had one's fill (of something)

[metaphorized and narrowed] > weary or tired (of something)

[broadened] > sorrowful, mournful

We get some further support for the last stage in this hypothesized development when we look at the meanings of the closest relatives of the Old English word, its cognates in the other Germanic languages. The next step back in the history of *sad* can be expressed as follows:

Old English *sad* is cognate with Old Dutch *sat*, Old Saxon *sad*, Old High German *sat*, Old Icelandic *saðr*, Gothic *saps*, all of which have meanings broadly corresponding to the Old English one, 'having had one's fill; satisfied, sated; weary or tired (of something)'.

However, the concept expressed by 'cognate with' needs some unpacking, and we will now look at this in more detail.

1.2.3 Cognates and language families

What does it mean to say that Old English *sad* (English *sad*) is 'cognate with' the words from Old Dutch, Old Saxon, etc. listed at the end of the previous section? Just as the Romance languages all developed from (some form of) Latin (see section 1.2.2), so English and a number of other languages, which linguists call the Germanic languages, developed from a common antecedent called proto-Germanic. Unlike Latin, we have no historical records for proto-Germanic, but we can reconstruct a good deal of information about it from the evidence of the languages that developed

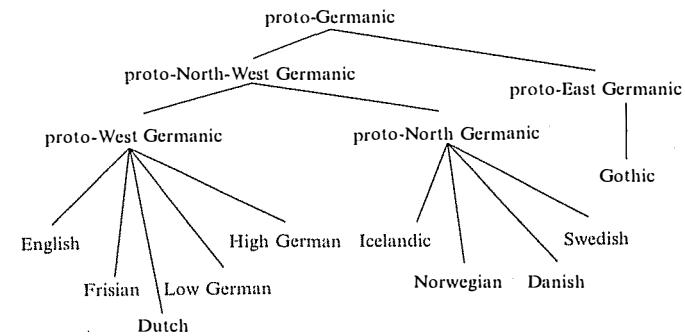


Fig 1.1 The major Germanic languages

from it. The other Germanic languages include Dutch (and hence Afrikaans), German (and hence Yiddish), Danish, Norwegian, Swedish, and Icelandic, as well as others such as Frisian (the closest relative of English, but with very few speakers today) and the extinct language Gothic (which is the Germanic language for which we have the earliest extensive documentary records, in the form of a bible translation dating from the fourth century AD). The cognates of an English word are the words in these other Germanic languages which can be explained as having developed from the same (unrecorded) antecedent word in proto-Germanic.

In fact, we can also identify subdivisions within the larger group of Germanic languages, on the basis of shared innovations that allow us to group the Scandinavian languages together as descendants of a common North Germanic sub-branch and likewise (albeit with rather more rough edges) English, Frisian, Dutch, Saxon/Low German, and High German as descendants of a West Germanic sub-branch. In turn, many scholars would now group together West Germanic and North Germanic as being descended from a shared North-West Germanic sub-branch with shared differences from East Germanic.⁷ Thus the relationships between the major Germanic languages can be represented schematically as in figure 1.1. We can reconstruct a similar tree structure for the major Romance languages, with the difference that in this instance the common ancestor, Latin, is of course attested (figure 1.2).

⁷ See for example Ringe (2006) 213. For a useful introduction to the early Germanic languages, see Robinson (1992).

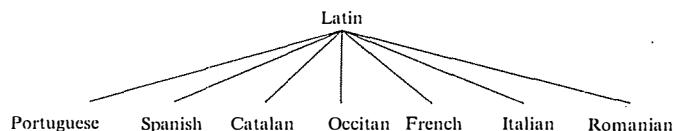


Fig 1.2 The major Romance languages

It is as well to pause for a moment and consider in a little more detail what this concept of a reconstructed antecedent language implies, because it will be crucial to many arguments later in this book. From present-day English to Old English (back as far as the eighth century, or even earlier in runic inscriptions) we have a chain of documents which enable us to trace the history of the English language in reasonable detail. In fact, these documents reflect many different local varieties of the language, showing many divergent developments. Some of these are reflected in different varieties of English today, such as the English of Chicago, or London, or Cape Town. We may analyse these as forming part of larger varieties, such as US English (or perhaps North American English), British English, or South African English. Alternatively we may subdivide them further, by looking for instance at different geographical or administrative areas of London, or at the language of different social classes within the city, or of different age groups, etc. Such variation must have been present throughout the history of English, although in earlier periods the nature and amount of the surviving evidence mean that we can only reconstruct a very limited picture. Modern US English and British English have developed as distinct varieties in different geographical locations from roughly the same antecedent, English as spoken in Britain in the early modern period (usually defined as approximately 1500–1750), but the historical record, as well as the evidence of modern US and British English, shows us that this common antecedent showed considerable internal variation. Similarly English and all of the other Germanic languages developed from a common antecedent (as did French, Spanish, etc. from Latin), but there is no reason to doubt, and every reason to suspect, that Germanic already showed internal variation. (Even though our surviving records for classical Latin are mostly literary and reflect a highly homogeneous literary language, there is indeed some variation in our surviving Latin evidence, and the later evidence of the Romance languages suggests the existence of a good deal of further variation in Latin

which is not reflected in the surviving documentary evidence).⁸ Over the course of time, groups of Germanic-speaking peoples developed distinct communities in different geographical locations (to some of which, like England, they had spread as part of the considerable movements of peoples which occurred in the later stages of the history of the Roman Empire and in the following centuries). As they did so, linguistic differences would have become more pronounced, as different variants from among the existing variation in Germanic came to predominate in different speech communities, and as new variation arose in each speech community.

At the time of our earliest substantial records for English, from several centuries after the Anglo-Saxons arrived in England, there are already important differences between English and its continental relatives, but these clearly took time to develop. We can also trace significant differences between different regional varieties of English in this early period, although the surviving documents leave very many questions unanswered.⁹ The demarcation of the various national languages of modern Europe owes a great deal to geography and, especially, politics. In the sixteenth and seventeenth centuries Scots was well on the way to developing a standard, ‘official’ form, distinct from the English of England, but subsequent political developments led to the adoption in official functions of a highly anglicized variety now usually referred to as Scottish English (although in recent decades as a result of the political process of devolution there have been some interesting developments in the use of Scots once again as an officially recognized variety in some functions). Today Dutch and German are well-defined national languages, sufficiently different from one another that monolingual speakers of either standard language have only an extremely limited degree of mutual intelligibility, but the situation is different among speakers of traditional dialects on or near the geographical boundaries between the two countries: such speakers can with a little effort understand the speech of their neighbours on the other side of the national border, even though one person is speaking something that is classified as a dialect of Dutch and the other something that is classified as a dialect of German. We can say that there is a dialect continuum which crosses the Dutch–German border. Another crosses the French–Italian border, and

⁸ On the degree of regional variation shown by surviving Latin documents from antiquity see Adams (2008).

⁹ For an introduction to the various issues involved see Hogg (2006).

similar cases can be found in many other parts of the world, essentially wherever languages have developed from a common source in adjacent territories.¹⁰

Such dialect continua lead us fairly directly to some limitations in the tree diagrams for the Romance and Germanic languages which I offered above. Diagrams of this type are a good way of representing where the most important shared innovations are found among various dialects in a group, but they have the disadvantage of making linguistic history appear artificially simple and neat. When two speech communities diverge, as represented by the branching on a tree, each takes with it a particular selection of features from the parent language. When further divergences occur subsequently, we may find that a particular feature is retained, quite by chance, in two languages or dialects which the weight of evidence places on completely different sides of the tree. In other cases the same innovation may occur independently in two different places, giving a false indication of inherited similarity. Additionally, where languages or dialects remain in contact, especially when they are spoken in geographically contiguous or overlapping territories, we can find that some features spread by diffusion (i.e. contact) from one variety to another, hence muddling the apparently clean branching shown by a tree. A better metaphor for such diffusion of features through language contact may be the spreading of a wave from a point of origin, rather than the branching of a tree.¹¹

1.2.4 Example two revisited: *sad* from proto-Germanic to proto-Indo-European

If we return to our example of *sad*, we can push this particular word history back further than just to proto-Germanic. The Germanic languages themselves form one branch of a much larger language family which historical linguists call Indo-European, which has numerous other branches, sub-branches, and isolate languages including for example:¹²

¹⁰ For an introductory account of these issues see Chambers and Trudgill (1998) 3–12. On the concept of a traditional dialect see especially Wells (1982) 4–8.

¹¹ For discussions of this issue with reference to the Germanic languages see Trask (1996) 181–7 (also Millar (2007) 225–31) and, at a rather more advanced level, Lass (1997) 139–59. On more general issues to do with language trees see McMahon and McMahon (2005).

¹² For an overview of the Indo-European languages see Fortson (2004).

- the Celtic languages: Welsh, Irish, etc.
- the Italic languages: Latin (and hence the Romance languages), Oscan, Umbrian, etc.
- Greek
- the Balto-Slavonic languages, comprising the Slavonic languages (Russian, Polish, etc.) and the Baltic languages (Lithuanian, Latvian, etc.)
- Albanian
- Armenian
- the Indo-Iranian languages, comprising the Iranian languages (Persian, etc.) and the Indic languages (Sanskrit and hence modern Hindi, etc.)

All of these languages can be shown to have developed from a single parent, proto-Indo-European, although of course all of them show the effects of contact with other languages during their histories. The identification of a shared ancestor for all of these languages rests upon the evidence of regular correspondences of sounds between the various languages, which we will look at in more detail below, and also upon systematic grammatical similarities, which are largely outside the scope of this book.

Many people have attempted to link Indo-European with other language families, but all such attempts remain extremely controversial, and the general view is that no genetic relationship has been reliably established between Indo-European and any other language family.

Precisely when and where proto-Indo-European existed as a spoken language is the subject of a very great deal of debate. This is complicated by the fact that the earliest recorded Indo-European language, Hittite, the oldest documentation for which dates back approximately 4,000 years, belongs to a branch, Anatolian, which probably split from the rest of Indo-European very early. However, what is reasonably certain is that proto-Indo-European began to split into its various daughter languages very much earlier than the date of our earliest documentary records for those languages. It is therefore unsurprising that many of the cognate forms bear little if any superficial resemblance to one another, since we are working at such a great time depth, and centuries of linguistic change lie between proto-Indo-European and even our earliest documentary evidence.

In this section we will trace the history of the word *sad* from proto-Germanic back to proto-Indo-European, and we will examine some of the procedures by which etymologies can be established at this time depth.

In doing so, we will encounter some principles and procedures which are equally applicable to much more recent linguistic history, and which we will investigate mostly from the standpoint of rather more recent linguistic evidence in the remainder of this book. However, reconstruction of linguistic data at a very considerable time depth is one of the big attractions of etymological research for many people, and it is also true that many of the most important aspects of modern etymological research came to fruition in the context of research into proto-Indo-European in the second half of the nineteenth century. We will therefore begin our investigation of the relationship between sound change and etymology by taking a look at how the sound changes known as Grimm's Law and Verner's Law help explain the etymology of *sad*.

By comparing the forms found in the Germanic languages with one another and also with forms in other Indo-European languages, we can reconstruct the proto-Germanic ancestor of *sad* as **sada*.¹³ An asterisk conventionally marks reconstructed forms, i.e. forms which are not actually recorded. **sada*- ends with a hyphen because it is a reconstructed word stem, i.e. the morphological stem to which inflectional endings were then added. In this book I will usually give reconstructions using IPA symbols, but without using square brackets [] implying that they are hypothetical phonetic transcriptions, nor // slashes implying that they necessarily have phonemic status. This is a traditional philological practice, which is useful for three main reasons: (i) we cannot always be certain about the precise phonetic quality of reconstructed sounds; (ii) any past historical sound system almost certainly showed considerable variation in the realization of sounds, which we cannot recover in detail from our historical evidence; (iii) we cannot always be sure whether certain distributions of sounds were phonemic or allophonic in a given historical period.¹⁴ We will look at issues

¹³ The exact phonetic quality and phonemic status of the consonant I have represented here as **ð* is in fact very uncertain. Many scholars choose to use **d* in reconstructions of proto-Germanic forms to represent any sound which may have been either a voiced plosive /d/ or a voiced fricative /ð/. In many modern etymological dictionaries the proto-Germanic form of this particular word is hence represented as **sada*. However, since the sound in this instance was almost certainly a voiced fricative at an early stage in proto-Germanic, I have used the reconstruction **sada*-, which has the advantage of making the changes from proto-Indo-European to proto-Germanic easier to follow.

¹⁴ For a recent detailed argument for this position see Lass and Laing (2007) §§2.4.2, 8.3.2.

to do with variation and change in any linguistic system in more detail in chapters 3 and 7.

The reconstruction **sada*- depends upon the evidence of the various Germanic languages, and also upon the evidence of forms in other Indo-European languages which can plausibly be referred to the same root form. Most crucially, it depends upon:

- (a) regular sound correspondences between the various languages
- (b) sound changes which can be posited to explain apparent irregularities

To get from proto-Germanic **sada*- to the recorded words Old English *sæd*, Old Dutch *sat*, Old Saxon *sad*, Old High German *sat*, Old Icelandic *saðr*, Gothic *sahs* requires just a couple of small steps:

- In West Germanic, proto-Germanic **ð* regularly became the voiced plosive /d/, as in our Old English form *sæd* /sad/ or Old Saxon *sad*. Old Dutch *sat* and Old High German *sat* show subsequent devoicing of this plosive (compare section 2.1.1.3).
- Old English *sæd* additionally shows Old English (and Old Frisian) fronting of West Germanic **a* to /a/.

These are regular, predictable sound changes in a word of this phonological shape in these languages.

This reconstructed proto-Germanic form **sada*- itself shows the reflex of an earlier Indo-European form **səto*- (The symbol **ə* in this reconstruction represents a sound which was realized as a vowel when it occurred in this position, hence giving rise to vowels in the daughter languages, but which is now generally believed to have resulted from the vocalic realization of one of a series of so-called laryngeal sounds which are hypothesized for proto-Indo-European. They are called laryngeals for historical reasons, although no one in fact knows exactly what their phonetic quality was. This particular laryngeal is sometimes represented as *ə* or as *h* or as *H*, depending on which transcription conventions are being followed. We will return to this topic in sections 1.3.1 and 4.4.1.)

Related words in other Indo-European languages include:

- classical Latin *sat*, *satis* 'enough', *satur* 'satisfied, full'
- Lithuanian *sotus* 'filling, full, satisfied, substantial'
- ancient Greek *áatos* 'insatiate' (showing a negative prefix)

We can see that the meanings of these words help support our hypothesis about the meaning development in the Germanic languages from 'satisfied, having had one's fill (of something)' to 'weary or tired (of something)'. It is difficult to be certain about the precise relationships between these words. They probably reflect two different variants, **sə-* and **sā-*, of a single Indo-European root for which the approximate meanings 'fill up, (make) replete' can be reconstructed. In our surviving cognates various different suffixes, **-to-*, **-ti-*, and **-tu-*, have been added to this root. The cognates thus do not represent the reflexes of a single word form, but rather the survivors of an extended word family, derived in various different ways from a common root.¹⁵ The Germanic words probably show what was originally a suffix which formed verbal adjectives, proto-Indo-European **-to-*. The same suffix is probably found in *old* (< proto-Germanic **al-da-*) and *cold* (< proto-Germanic **kal-da-*; compare Latin *gelidus*), and in many Latin words ending in *-tus*. (On roots and their meanings see further sections 4.4.1 and 8.7.3.)

The assumption made in the last paragraph that proto-Germanic **saða-* is likely to have developed from proto-Indo-European **səto-* may seem rather startling to anyone who does not have a prior acquaintance with Indo-European linguistics. On the face of it only the initial consonant **s* is common to both forms. However, the development of the vowels is easily dealt with, by the principle of regular sound correspondences. Proto-Indo-European **ə* (with the caveats given above) and (short) **o* both regularly develop to **a* in proto-Germanic, thus **səto-* > **saða-*. A sound change of this sort is called a merger: the phonetic development of **ə*, **o*, and **a* in proto-Germanic led to loss of the distinction between the three separate proto-Indo-European phonemes and merger as the single phoneme **a* in proto-Germanic. Compare Latin *hostis* 'stranger, enemy' with its cognate Gothic *gasts* 'guest', or Latin *hortus* 'garden' with its cognate Gothic *gards* 'garden'. (Latin *h* and Gothic *g* in these words show the regular development in Latin and in proto-Germanic of proto-Indo-European **g^h*; we will look further at the Germanic side of this in the next paragraph. The modern English cognates of these words are respectively *guest* and *yard*, showing the result of a number of sound changes during the history of English.)

¹⁵ For a specialist readership, the best recent detailed account of the Germanic component of this etymology is provided (in German) by Heidermanns (1993) 458–9; on the Indo-European component see especially Szemerényi (1979).

Probably, on the basis of the evidence of other Indo-European languages, in proto-Germanic the reflexes of proto-Indo-European **ə* and **a* merged first as **a*, with which **o* then also merged. Conversely, the proto-Indo-European long vowels **ō* and **ā* merge as **ō* in proto-Germanic.

The explanation for the medial consonant in proto-Germanic **saða-* is a little more complicated, and involves two reconstructed sound changes. Comparison among the Indo-European languages excluding Germanic leads to the reconstruction of three sets of stop consonants: voiceless stops (**p*, **t*, **k*, **kʷ*), voiced stops (**b*, **d*, **g*, **gʷ*), and breathy-voiced stops (**b^h*, **d^h*, **g^h*, **gʷʰ*). Comparison with the forms in the Germanic languages leads to the conclusion that a series of sound shifts occurred in proto-Germanic:

- *p > *f
- *t > θ (represented in traditional philological notation as *þ)
- *k > *h
- *kʷ > *hw
- *b > *p
- *d > *t
- *g > *k
- *gʷ > *kw
- *b^h > *β (in some environments > *b)
- *d^h > *ð (in some environments > *d)
- *g^h > *γ (in some environments > *g)
- *gʷʰ > *γw (in some environments > *gw)

Thus the voiceless stops became voiceless fricatives, the voiced stops became voiceless stops, and the breathy-voiced stops lost their breathy-voice and probably became fricatives before becoming voiced stops in many environments. Experts in fact differ on many details of this process, especially as regards the proto-Indo-European breathy-voiced stops and also the proto-Indo-European voiced stop **b* (which was very rare, and some argue did not exist at all), but this is not of importance for our present purposes.¹⁶ This sound change (or series of changes) is known as Grimm's Law, after the German philologist Jakob Grimm (1785–1863), who compiled with his brother Wilhelm both the celebrated fairy tale collection and the early

¹⁶ The literature on Grimm's Law, and Verner's Law,¹⁷ is vast. For a recent detailed account of the changes see Ringe (2006) 93–116; for particularly useful analyses see also Bynon (1977) 83–6, Collinge (1985) 63–76. See also the discussion in section 7.1 below.

fascicles of the major historical dictionary of the German language. Grimm produced an important early formulation of this sound change, although it had in fact been described earlier by other scholars. An alternative name for this sound change is the Germanic Consonant Shift.

We can illustrate the changes in the proto-Indo-European voiceless stops with the following examples:

$*p > *f$

I-E root **ped-* 'foot': ancient Greek *poús* (stem *pod-*), Latin *pēs* (stem *ped-*); Gothic *fōtus*, English *foot*

$*t > *θ$

I-E **tū* 'you (singular)': Latin *tū*, Old Irish *tū*; Gothic *þū*, English *thou*

$*k > *h$

I-E root **kerd-* 'heart': ancient Greek *kardía*, Latin *cor* (stem *cord-*); Gothic *hairtō*, English *heart*

$*kʷ > *h₂w$

I-E **kʷós* 'who': Sanskrit *kás* 'who', Lithuanian *kàs* 'who, what'; Gothic *hwas* 'who', English *who*

In the first example here, 'foot', Grimm's Law explains not only the shift of the initial consonant from **p* to **f* but also the shift of the final consonant of the stem from **d* to **t*. However, it will be obvious at a glance that there are other differences between the cognates apart from those explained by Grimm's Law, even though I have attempted to select forms which have an unusually close mutual resemblance (another of the cognates of English *foot* is in fact Armenian *otn*). In the case of 'foot', the Greek, Latin, and Germanic words all have different stem-vowels. In this instance the difference is not due to sound changes which have occurred in the daughter languages, but to slightly different etymons in proto-Indo-European: the Greek stem form *pod-* is from proto-Indo-European **pod-*, the Latin stem form *ped-* is from proto-Indo-European **ped-*, and the Germanic forms are from proto-Indo-European **pōd-*. These different etymons are all derived from the root **ped-* by a process known as ablaut which we will look at in section 4.4.1. This also explains the variation between **səd-* and **sād-* which we encountered above in the etymology of *sad*.

The operation of Grimm's Law thus explains why proto-Germanic **sāda-* < proto-Indo-European **sətō-* does not show medial **t*, but it does not explain why it shows **ð* rather than the expected **θ*. This is explained by another sound change known as Verner's Law, after the Danish philologist

Karl Verner (1846–96), by which the proto-Germanic voiceless fricatives became voiced whenever the accent did not fall on the immediately preceding syllable. (For an analogous situation in modern English, compare *ex'ert /ɛg'zə:t/* with *'exercise /'eksə'saɪz/*.) In the ancestor of *sad* the suffix, not the root, was stressed, and hence Verner's Law applied, giving voiced **ð*. Later, the accent shifted to the first syllable in all words in proto-Germanic, thus giving the pattern which we find reflected in all of the recorded Germanic languages. Hence, finally, we can explain how proto-Indo-European **sətō-* would give rise to proto-Germanic **sāda*, via the following stages: **sətō-* > **sā'ta* > **sa'θa* > **sa'ða* > **sāða*. We will not do so here, but pre-histories can similarly be reconstructed for classical Latin *sat*, *satis*, *satur*, Lithuanian *sotus*, and also ancient Greek *āatos*, and it is this (rather than vague resemblance in form and meaning) which gives substance to the hypothesis that all of these forms are ultimately cognate.

We will return to Grimm's Law and Verner's Law in a little more detail at the beginning of chapter 7, but for the time being there are one or two very important general observations which arise from this example. Note that in the preceding paragraph I said that proto-Indo-European **sətō-* 'would give rise to' proto-Germanic **sāda-*, and not 'could give rise to'. The merger of **ə*, **o*, and **a* as **a* in proto-Germanic, and the Grimm's Law and Verner's Law changes, are all regular processes, which apply in all cases (where not excluded by specific phonetic environments, which simply involve more precise statement of what the sound change was and in which environments it applied). The standard methodology of comparative linguistics does not permit us to say 'perhaps in this particular instance the merger simply did not happen' or 'perhaps Grimm's Law did not apply to this word' or 'perhaps in this instance an entirely unparalleled change of **ð* to **m* occurred'. As I have formulated it here, this is an oversimplification, but not a huge one. In chapter 7 we will look at the reasoning behind this in much more detail, and at some important qualifications, but for present purposes it is sufficient to be aware that comparative reconstruction depends upon the regularity of the correspondences and sound changes which are posited: this (as well as general phonetic plausibility, and the existence of parallels in the documented history of languages) is what gives a solid foundation to comparative etymological research.

A useful illustration of this principle is shown by the histories of the words *mother*, *father*, and *brother*. All three words show a voiced fricative /ð/ in modern English. However, in Old English the situation was

rather different: *brōðor* 'brother' showed a voiced fricative /θ/, but *mōðor* 'mother' and *fæder* 'father' both showed a voiced plosive /d/. In proto-Indo-European all three words in fact showed the same termination; *-tēr- (in the nominative case), which seems typical of terms for family kinship: *mātēr 'mother', *pātēr 'father', and *bhrātēr 'brother'.¹⁷ compare Latin *māter* 'mother', *pāter* 'father', *frāter* 'brother' (proto-Indo-European *b^h > f in word-initial position in Latin; compare also Sanskrit *bhrātar-*). The explanation for the different outcomes in Old English is the regular operation of Verner's Law. In the case of *mother* and *father* the stress in proto-Germanic fell on the second syllable, while in the case of *brother* it fell on the first syllable. Thus Verner's Law applied in the case of *mother* and *father*, but not in the case of *brother*, and so we find that proto-Germanic *brōþēr, with voiceless fricative *θ, corresponds to Latin *frāter*, but that proto-Germanic *mōðēr and *fāðēr, with voiced fricative *ð, correspond to Latin *māter* and *pāter*. In *mother* and *father* the proto-Germanic voiced fricative subsequently became a plosive in West Germanic, just as in the case of *sad*, hence Old English *mōðor* (or in fact more commonly *mōðor*, showing variation in the unstressed vowel of the second syllable) and *fæder*. In the case of *brother*, the medial voiceless fricative of proto-Germanic *brōþēr became voiced in intervocalic position in Old English, hence Old English *brōðer* (again in fact more commonly *brōðor*). Subsequently, in late Middle English, by another sound change, the voiced plosive of *moder* and *fæder* developed into a fricative before either /ər/ or syllabic /r/, resulting from reduction or loss of the vowel in the endings -or, -er. Thus, *mother* and *father* came to have the same voiced fricative as *brother*. So we can see that *mother*, *father*, and *brother* provide a very rare example of how subsequent sound changes can, very occasionally and entirely fortuitously, restore a formal resemblance which had been obscured by a much earlier sound change (figure 1.3). We have also now seen how *brother* and *friar*, discussed in section 1.2.1, are in fact cognate, both being ultimately from proto-Indo-European *bhrātēr. In the latter case the development was: *friar* < Old French *frere* < Latin *frāter* < proto-Indo-European *bhrātēr.

¹⁷ In the reconstructions *mātēr and *bhrātēr the *ā in the first syllable shows what is now generally considered to have been the output of earlier *eh₂, i.e. the vowel *e followed by a laryngeal which caused colouring and lengthening of the vowel. For a fuller explanation of this see section 4.4.1.

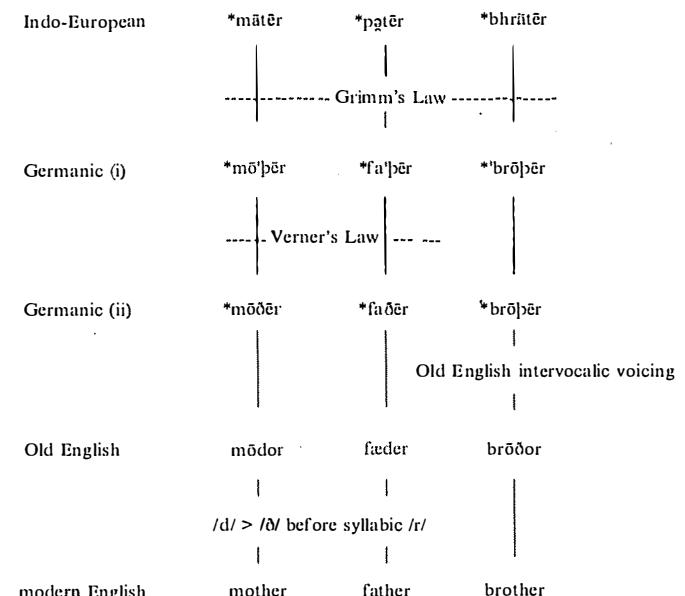


Fig 1.3 *mother*, *father*, and *brother* from proto-Indo-European to modern English

1.2.5 Summary

Our initial supposition about the meaning development of *sad* within English was supported by comparison with the meanings of its cognates in other Germanic languages, and ultimately also by the meanings of its cognates elsewhere in Indo-European.

In tracing the word's cognates at a great time depth we have seen the importance of regular sound correspondences and of regular sound changes in accounting for apparent discrepancies. We will return to this topic in more detail in chapter 7.

In the etymologies of both *friar* and *sad*, there is little or no connection between the processes of formal development and the processes of meaning development that we have examined. This is often the case, although there are also cases where form history and meaning history are very closely intertwined, and we will look closely at a number of such cases in chapters 7 and 8.

1.3 Why study etymology?

1.3.1 Etymology, historical and comparative grammars, and dictionaries

Etymology is an essential tool in reconstructing the history of a language, since a corpus of word histories provides a necessary basis for many other aspects of historical linguistic work. Conversely, each individual word history depends for its plausibility on the work that has been done in various subfields of historical linguistics. For instance, someone interested in historical semantics will want to look at the meaning histories of individual words which have been traced through the application of etymology, just as an etymologist will want to draw on the general observations about a whole body of meaning changes and their likely motivations which have been identified by specialists in historical semantics. Each activity informs and enriches the other in a mutually beneficial relationship.

Traditionally, etymology has been associated most closely with the construction of historical and comparative grammars. A historical grammar traces the developments in word forms which are found in the history of a language, often also extending into its pre-history. A comparative grammar relates the developments found in one language to those found in cognate languages, to explain the development of two or more languages from a common source using the technique of comparative reconstruction.

We have seen in the case of *friar* an example of how etymology interacts with the functions of a historical grammar:

- Etymological investigation suggests that *friar* shows the continuation of Middle English *frere*.
- A historical grammar identifies parallels such as *briar* and *choir* (themselves the result of other etymological investigations). Ideally, it will also supply an explanation for the unusual form history shown by such groups of words.

Our investigation of *sad* gave an insight into the world of comparative etymology and comparative reconstruction. The identification of regular sound correspondences depends at first upon the investigation of large numbers of potential etymological connections. This may make it possible to identify the regular processes of sound change. If so, our corpus of etymologies can be refined, and some at first apparently attractive connections can be discarded, at least until we can find a new explanation to account for them.

The best illustration of this may be to look at an example of how a sound method may enable us to identify a case of chance resemblance. If we start out, from an entirely uninformed perspective, by looking simply for words which are similar in form and meaning, English *care* and Latin *cūra* ‘care’ might seem attractive candidates for investigation: they overlap completely in their core meaning, and the consonants at least are the same. There is thus more resemblance in both form and meaning than there is between English *sad* and Latin *satis* ‘enough’ or Lithuanian *sotus* ‘filling, full, satisfied, substantial’. However, English *care* is an inherited Germanic word, with a good set of cognates from all branches of Germanic which enable us to reconstruct a proto-Germanic form **karō-*. If we remember Grimm’s Law, we will see that proto-Germanic /k/ is not going to correspond to Latin /k/, and in fact proto-Germanic **karō-* is usually referred to a proto-Indo-European root **gar-* with the meaning ‘to call, cry’. This same root is probably reflected also by Latin *garrire* ‘to chatter’ (ultimately the base of English *garrulous*). Latin *cūra* shows the regular development of an earlier form **koisā*, which can be reconstructed on the basis of forms in inscriptions and cognates from other Italic dialects; it has no generally accepted further etymology, but could not conceivably be connected with proto-Germanic **karō-*. In fact some doubts have been raised about the connection of proto-Germanic **karō-* with proto-Indo-European **gar-*.¹⁸ Revised or contested hypotheses are very common in etymological work at this sort of time depth. However, the important point is that a connection with Latin *cūra* remains impossible, even if we have no viable etymology for **karō-*: we do not need to have an alternative explanation in order to reject an impossible etymology.

Latin *deus* ‘god’ and Greek *theós* ‘god’ are another pair of words which are synonymous and have a superficial resemblance in form, but which the methodology of comparative linguistics demonstrates have no etymological connection whatever: the first goes back to proto-Indo-European **deiwós* and the other probably to proto-Indo-European **dhesos*. We can thus make an important generalization: comparative reconstruction provides an essential tool for quickly eliminating very many cases of chance resemblance in form and meaning, just as it identifies many cognates which have little or no superficial resemblance in form or meaning.¹⁹ It also leaves us with

¹⁸ See for instance (in German) Rix (2001) 161.

¹⁹ For an excellent and much more detailed account of these and related issues see Campbell (2003).

very many rather doubtful cases, some examples of which we will examine later.

Sometimes 'etymology' has been seen as almost synonymous with 'comparative reconstruction', or at least it has been assumed that everything else which an etymologist has to consider is of secondary importance in comparison with the reconstruction of antecedent word forms and the identification of historical sound changes. This will not be entirely the approach adopted in this book, although it should not be forgotten that form history, as reflected in historical and comparative grammars, provides the backbone for nearly all etymological research: we will examine in detail in chapters 7 and 8 how and why it is that arguments based on word form usually provide by far the strongest foundation for etymologies.

Comparative reconstruction has a sister methodology known as internal reconstruction, in which reconstruction is based purely on the data provided by a single language. This is generally much more limited, and also less reliable, than comparative reconstruction, and it will not be a major topic in this book, although it should be noted that methods of internal reconstruction have contributed some important advances in knowledge even in areas such as Indo-European linguistics where the comparative data is relatively rich and plentiful. It tends to be most effective in tracing the origins of morphophonemic relationships, as between English *mouse* and *mice* (see section 7.2.4) or the contrast between voiceless and voiced consonants in German *Rad* and *Rädes* (section 2.1.1.3), although even here comparative data is often much more conclusive.²⁰ One very important and justly famous success of internal reconstruction was Ferdinand de Saussure's identification in the late nineteenth century of a series of hypothetical sounds in proto-Indo-European which he termed (in French) 'coefficients sonantiques'. These are now generally recognized as a series of so-called laryngeal sounds (although their exact quality is in fact unknown and the subject of much dispute). Hittite documents which began to be deciphered and studied in detail in the early twentieth century, long after Saussure's initial hypothesis based on internal reconstruction, provided crucial data which confirmed the reconstruction.²¹ We will return to this topic, and to its implications for the sound represented by * ϱ in the proto-Indo-European reconstructed forms given here, in section 4.4.1.

²⁰ For thorough accounts of internal reconstruction see Fox (1995) or Ringe (2003).

²¹ For short accounts of this see for example Fortson (2004) 75–6; also Hock (1991) 545–9, Clackson (2007) 53–61, or Millar (2007) 322–7.

Aside from historical and comparative grammars, etymology is also a crucial scholarly tool in historical lexicography. Historical dictionaries present in linear form the word histories which are treated thematically in grammars: in grammars we can see the connections between the developments shown by individual words, while in historical dictionaries we can see word histories whole and uninterrupted, together with the interplay between form history and meaning history, and at least some information on the influence of extralinguistic cultural and historical factors.

1.3.2 Historical relationships between words

A key function of etymology is that it illuminates the formal and semantic relationships between the words of a language. This is an area where a layman's interests may not be entirely dissimilar to those of a historical linguist, and thus it can be a very good entry point for people who are relatively new to the study of etymology. Indeed, this topic is of particular interest for speakers of a language like English which has seen a good deal of borrowing, and where the semantic relationship between for example *hand* and *manual* 'involving the hand, operated by hand, etc.' is obscured by the absence of any formal relationship between the two words. In this particular instance, the word *manual* is ultimately a derivative formation from a word meaning 'hand', but the word in question is Latin *manus* 'hand' (plus a Latin suffix *-alis* which forms adjectives with the meaning 'connected with') rather than English *hand*. Latin *manuālis* was borrowed into English (via French) as *manual* in the fifteenth century. For a time it competed with a word with the same meaning which did have a transparent formal relationship with *hand*, namely *handy*. This word today only has the specialized meanings 'convenient to handle or use', 'ready to hand', 'skilful, good with his or her hands', but in early use it also had the meaning 'done by hand, manual'. It is formed from *hand* and the suffix *-y* (which has a function similar to Latin *-alis*), although this is not the full story: *handy* probably originally arose as a result of reanalysis of the word *handiwork*, which was itself formed much earlier. *handiwork* is not (as we may at first assume) formed from *handy* and *work* but from *hand* and the obsolete noun *geweorc* 'work', which is a derivative of Old English *weorc* 'work' formed with a prefix *ge-* which had a collective meaning (thus 'work collectively') and which was pronounced with a palatal initial consonant /j/, thus /jeweɔrk/. In course of time phonetic reduction occurred in the unstressed medial syllable

of *handgeworc*, giving the form *handiwork*, which was then reanalysed as showing *hand*, -*y*, and *work*.

This small example illustrates some very important tendencies in word histories, which etymologists must always bear in mind. There will often be a formal relationship between words which have a semantic connection with one another. Thus, a word which means 'performed by hand' will very likely be related in form to a word meaning 'hand': in English we can imagine compound formations such as **hand-done* (compare *handmade*) or derivative formations such as **handish*, *handly*, or indeed *handy*. The asterisk here indicates entirely hypothetical word forms, rather than reconstructed word forms as we saw before with **sada-* in section 1.2.4. The word *handly* has no asterisk because it is in fact recorded several times in Middle English, and with precisely the meaning 'manual'. It was thus another synonym in competition with *manual* and *handy*.

This sort of relationship is called an iconic one: the word forms echo what seems to be the intuitive meaning relationship between the words. Such compound or derivative formations are called transparent when there is a clear form-and-meaning relationship between the complex word and its component parts. (We will look at transparency in more detail in chapter 2, and iconicity in chapter 4.)

Borrowing can disrupt these relationships, if, as typically happens, not all of the words in a related group are borrowed. In this particular instance so-called prestige borrowing of a relatively technical word has occurred, but the more basic word *hand* has not been replaced by a parallel borrowing of (Anglo-)French *main* or Latin *manus*. We will look at different sorts of borrowing situations, and their often unpredictable outcomes, in much more detail in chapters 5 and 6. For one example of the rather messy results of different borrowing processes compare the synonymous nouns *manual* and *handbook* in modern English. Both denote a book containing concise information readily to hand. *manual* shows borrowing from (Anglo-)French *manual*, which is itself from Latin *manuāle*. *handbook* was formed as a calque or loan translation (see section 5.1.2) on the model of Latin *manuāle*, although in modern use it owes its currency mostly to the influence of German *Handbuch* in the nineteenth century (which was also formed on the model of Latin *manuāle*).

handiwork shows another typical process, where the composition of a word has become obscured or opaque with the passage of time. Had Old English *geweorc* survived into Middle English it would have had the form

**iwork* (or more properly **iwerk*), and so it would have paralleled the formal changes shown by *handiwork*, but it did not survive, and *handiwork* became as it were an 'orphan', open to reinterpretation as showing *hand*, -*y*, and *work*. This reanalysis leads to the appearance of the adjective *handy*, and probably also to the remodelling of the word *handcraft* as *handicraft*. Thus, loss of other words in the linguistic system can lead to what were originally transparent relationships becoming opaque. Opacity can also result from many other factors, such as sound change. The great counter-force is analogy, in this case leading to reanalysis of *handiwork* and the formation of new words on the same pattern, thus setting up a new set of correspondences between form and meaning, albeit ones quite different from those found earlier in the word's history. (We will look at the workings of analogy in detail in chapter 7.)

We see here that an example of how etymology can help us to understand oddities in the modern-day structure of the vocabulary of a language has also brought us back to the interconnection of etymology with many other aspects of historical linguistics. This is one of the most fascinating aspects of etymology: we can move quite swiftly from interesting information which helps inform our understanding of the historical relationships between words in everyday use, to data that helps us to understand processes of historical linguistic change. Indeed, very often the same information serves both functions at once.

1.3.3 The etymological fallacy

It may seem odd to spend part of this chapter discussing what etymology is not for, but the misconceptions are very widespread, and colour many popular ideas about word histories. Additionally, of course, in examining what etymology is not about, we will uncover a good deal of what it really is about, and we will also see some further illustrations of how words change in both form and meaning over time.

The etymological fallacy is the idea that knowing about a word's origin, and particularly its original meaning, gives us the key to understanding its present-day use. Very frequently, this is combined with an assertion about how a word ought to be used today: certain uses are privileged as 'etymological' and hence 'valid', while others are regarded as 'unetymological' and hence 'invalid' (or at least 'less valid'). This attitude certainly has a venerable history: the word *etymology* is itself ultimately from ancient

Greek *etumologia*, which is formed from *éthmos* ‘true’ and *lógos* ‘word, speech’, hence denoting ‘the study of true meanings or forms’.²²

Perhaps the easiest way to illustrate the assumptions lying behind the etymological fallacy is to look at some verbal controversies of the relatively recent past. Today use of the word *meticulous* in the sense ‘painstakingly careful’ is perfectly normal and does not invite any negative reaction, but in the late nineteenth and early twentieth centuries it attracted a good deal of comment. The central ground of the objection was etymological. The word comes ultimately from Latin *metus* ‘fear’, and it first occurs in English (as also in French) in the sixteenth and seventeenth centuries in the sense ‘fearful’, for instance in the Older Scots writer William Stewart’s translation of Hector Boece’s *Chronicle of Scotland*, ‘Gif thou be... Meticulos, and dar nocht se blude drawin’ (‘if you are fearful, and do not dare see blood drawn’).²³ The word resurfaces in French in the early nineteenth century in the sense ‘overscrupulous’, with the connotation ‘fearful of making a mistake’, and it swiftly enters English in this sense, being found in 1827 in *Blackwood’s Magazine*: ‘He does many things which we ourselves, and we do not hold ourselves peculiarly meticulous, will not venture upon.’ However, the word subsequently developed more positive connotations in both French and English, as defined by the *OED*: ‘Subsequently usually in more positive sense: careful, punctilious, scrupulous, precise’. As we will see in chapter 8, this is a very far from unusual process of semantic change: the word’s meaning has first narrowed, and then it has developed more positive connotations or ameliorated – or in this particular instance, it would perhaps be more accurate to say that it has lost its negative connotations. But for many prescriptive commentators on English usage in the early twentieth century, this new sense was to be avoided, on the grounds that it was not sanctioned by the word’s history, and specifically by the meaning of the Latin word from which it was ultimately borrowed. (For a useful summary of such comment see *Webster’s Dictionary of English Usage* (1989) 634.)

²² On the early history of the word and the concept see the short sketch in Lass (2007) §8.1.1 and further references there, and also the discussion in the four chronological volumes of Lepschy (1994a), (1994b), (1998), and Morpurgo Davies (1998). On the study of the etymology of English words up to 1882, when the first fascicle of the *OED* appeared, see Görlich (2002b) 71–136. On etymology in the twentieth century see especially Malkiel (1993).

²³ See *OED3* at *meticulous* adj., as also for the quotation from *Blackwood’s Magazine* below.

Similarly, the word *obnoxious* comes ultimately from Latin *obnoxius*, which is formed from the preposition *ob* ‘in front of, in view of’ and the noun *noxa* ‘hurt, injury’ (compare modern English *noxious*, used frequently of harmful substances, especially gases). The Latin adjective had the meanings ‘exposed to harm, liable, answerable, submissive, subject to punishment’, and it is broadly these meanings which are commonest from the word’s first occurrence in English in the sixteenth century down to the nineteenth century. As late as 1902 we find in William James *Varieties of Religious Experience*: ‘The impulse... is... far too immediate and spontaneous an expression of self-despair and anxiety to be obnoxious to any such reproach.’²⁴ However, from the late seventeenth century onwards we find a sense which the *OED* defines as: ‘Offensive, objectionable, odious, highly disagreeable. Now esp. (of a person): giving offence, acting objectionably; extremely unpleasant, highly dislikeable.’ This results from association with *noxious*, and has become the usual sense in modern English (indeed it is the only one for which the *OED* records any examples later than 1902), but in the nineteenth century use in this sense was a matter of contention, and again the focus of debate was the word’s etymology. (For a summary see again *Webster’s Dictionary of English Usage* (1989) 676.)

These are both complex words, and their original meaning is to some extent guessable for people who know some Latin because the composition of each word is transparent. It is notable that in English attempts to determine usage by recourse to etymology very often involve words of Latin origin, and particularly words which remain reasonably close in form to their Latin etymons, so that the historical connection between the two is fairly obvious, as in the cases of *meticulous* or *obnoxious*. We can see an interesting cultural phenomenon in action here, where the authority of an ancient language is taken to be an effective arbiter of usage even in a quite different language some two thousand years later. However, so far as the scientific study of language is concerned, such assertions about the authority of ‘etymological meanings’ are quite irrelevant; or rather, if they are relevant to anyone, it is to people studying attitudes towards language use, rather than to etymologists. It is one of the linguistic facts of life that words change both in form and in meaning. Predicting exactly what those changes will be and when they will occur is normally impossible, although

²⁴ See *OED3* at *obnoxious* adj.

describing and explaining changes which have occurred in the past is a much more achievable goal, and forms the main focus of this book.

The changes in meaning shown by *meticulous* or *obnoxious* look very minor when compared with some much more dramatic changes in meaning which have occurred during the recorded history of English, but which tend to be noticed only by linguistic historians and by people reading texts from earlier periods.

To take a much cited example, the English word *deer* originally denoted any animal, as its cognates Dutch *dier* and German *Tier* still do today. However, in the course of the Middle English period the word came to be applied more and more often specifically to the deer, and in early modern English the broader sense ‘animal’ was lost completely, so that whenever the word occurred it had the narrowed sense ‘deer’. Explaining why this happened is much more difficult, and in spite of the popularity of this example in the literature, there is no generally accepted explanation.²⁵

To take another example, the word *treacle* originally (from the fourteenth century) denoted a kind of medicine, as it did also in its donor language French and in the other Romance languages; in an extended figurative meaning it could denote anything with healing effects. Its transferred use to denote a type of sugar product dates only from the end of the seventeenth century, but now is the only one which remains in current use (except when this sense is itself used figuratively, especially of compliments or praise).

We will look in more detail at the mechanisms of meaning change in chapter 8, but we should already be able to put the etymological fallacy to one side if we consider how foolish it would be to assert that English *deer* should be used in the sense ‘animal’ (and another word be used in the meaning ‘deer’) because of its history and the modern meanings of its cognates Dutch *dier* and German *Tier*, or that *treacle* should revert to the meaning ‘medicine’ because of its history (its ultimate etymon in Greek in fact means an antidote against a venomous bite). Earlier in this chapter we

²⁵ For one attempt see Samuels (1972) 73–4, who examines the relationships between the terms *beast*, *hart*, and *deer* in Middle English, and suggests that the homophony between *hart* and *heart* may have blocked adoption of *hart* as a general term for the deer, while partial homophony between *deer* and the adjective *dear* may have been a pressure against continued use of *deer* to denote more ferocious wild animals. Such arguments based on what is often termed ‘dangerous homophony’ are controversial, especially in cases where, as in this instance, genuine ambiguity must rarely if ever have occurred. See further discussion of arguments of this type in section 3.8.

saw a similarly dramatic semantic development in the word *sad*: it would be absurd to suggest today that *sad* should be used only in the sense ‘satisfied’ because of its etymology.

1.4 What an etymologist does

Our initial investigation of the comparative method has given a first illustration of the methodology of an etymologist. Various aspects of this methodology will take up most of the rest of this book. We will end this first chapter by considering some of the typical activities that characterize etymological research. In any (hypothetical) day of etymological research a lot of what happens will depend upon the particular circumstances of the language or period being studied, reflecting such factors as how much data is available, and what form that data takes. However, some things are almost certain to be true: there will be few, if any, blinding flashes of insight, and any that do occur will be the result of a good deal of painstaking work. Gathering data together (from important source texts, from corpora, from dictionaries, or from the work of previous researchers) is likely to figure largely, along with the careful analysis of this data. Frequently this analysis will involve approaching the same material time and again from different points of view, testing out one hypothesis after another, and probably discarding most of them as they run aground in insuperable difficulties. When real progress is made, it is most likely that it will emerge slowly, as the etymologist attempts to approach the same set of data with (yet) another hypothesis, to find that on this occasion the hypothesis does not collapse, but holds up against all of the challenges that one can think of to test it with. And then very probably one puts the hypothesis to one side for a little while and comes back to it another day, to see whether one had overlooked an obvious difficulty. Only then may one begin to feel that perhaps some real progress has been made.

Whenever we try to establish a link between two pieces of data, we must remember to check how plausible this link is from a variety of different perspectives. Is there any difficulty semantically? Can we find parallels for any changes in meaning that we assume? Is the connection acceptable phonologically? If phonological changes are posited, are they plausible, and do we have parallels for them? Are any morphological relationships which are posited plausible, and are they supported by parallels? Finally, is this

hypothesis demonstrably preferable to any others which have been proposed or which we can formulate?

More often than not, the word history which emerges from this process will reflect the work of more than one researcher. A lot of etymological research involves taking up the threads of past investigations, carefully going through the work of previous researchers (who perhaps worked generations ago), and seeing whether new data or new insights help reinforce and confirm a hypothesis suggested by earlier research, or instead challenge this hypothesis, or even suggest a new one. Fortunately, a lot of etymological work ages rather well. Of course, we must always be very careful when revisiting older scholarship to take note of any places where it rests on outdated assumptions, and to investigate it rigorously by applying modern methodologies and procedures. But so long as due caution is exercised, a great deal of scholarship dating from at least as far back as the late nineteenth century is still an excellent foundation for further work. There is, of course, a good reason for this: as we have already noted in discussing Verner's Law, many of the most important advances in the development of linguistic reconstruction and the comparative method belong to the late nineteenth century, and although there have been very important methodological advances since then, much of the scholarship of that period still does not appear to be in a completely alien scholarly 'language'.

Finally, words form part of a system, the lexis of a language, with numerous links to its grammar also. Any change in our understanding of one part of that system may have echoes or repercussions in another, possibly quite distantly removed, part of the same large system, and we must always be alert to such implications in our own or others' work. Sometimes, one changed etymology can open the way to a whole set of new solutions to old problems. One should bear in mind the adage of the great French comparative linguist and etymologist Antoine Meillet that a language is 'un système ... où tout se tient', 'a system where everything is connected' (Meillet (1921) 16; also cited in similar form at many other points in Meillet's work: see Koerner (1999)). Some linguists would reformulate this as 'a system where many things are connected', but still we should be alert to the implications that one etymology may have for many other word histories. Additionally, we must never forget that words and languages are spoken by real people, living in a particular society at a particular point in history, and it is in the usage of individual speakers that changes in word form and word meaning arise and develop. In order to understand the words of the past we must

often immerse ourselves in its material and intellectual culture, in order to trace connections between words and concepts which may seem quite unrelated from a modern perspective. We should also give consideration to the many different registers and styles of language, and the specialist vocabularies of different groups and communities. When we take account of such issues, we are likely to produce much better etymologies, and we may also make some important discoveries about social and cultural history.

As we have seen, a lot of argumentation in etymology, whether it concerns form history or meaning history, works on the basis of establishing parallels, in order to identify regular patterns of language change which lend support to individual etymologies. However, if we also have a reasonable explanation for why a change may have occurred, this is inherently much more satisfying, and more productive for work in historical linguistics in general. Additionally, if we have a plausible explanation for why a change is likely to have happened in one case, we can assess whether similar circumstances are likely to have existed in a hypothetical parallel case.

The task of an etymologist is thus a very large one. It was described with characteristic boldness by one of the great etymologists of the twentieth century, Walther von Wartburg:

Today the task of etymology is no longer solely to look for the root of a word or group of words. It must follow the group in question throughout the whole period during which it belongs to the language, in all its ramifications and all its relations to other groups, constantly asking the questions appropriate to etymology in the strict sense of the word.

(von Wartburg, tr. Reid (1969) 121)²⁶

We may not always be able to answer all of the questions that such an investigation poses, and sometimes there may be so little evidence that we can barely establish any trace even of a word's existence, but we should still not lose sight of this ultimate aim.

²⁶ Die Erforschung des Radix eines Wortes oder einer Wortgruppe ist heute nicht mehr die einzige Aufgabe der Etymologie. Sie hat die zu betrachtende Wortgruppe in ihrer Verästelung und mit all ihren Beziehungen zu anderen Gruppen während der ganzen Zeit, da sie einer Sprache angehört, zu verfolgen, ohne jemals die etymologisierende Fragestellung aufzugeben.

(von Wartburg (1962) 120-1)

2

What is a word? Which words need etymologies?

2.1 What are words? 34	2.5 Apparent reversals of the process 56
2.2 How new words arise 43	2.6 Cranberry morphs 56
2.3 Lexicalization 49	2.7 Which words need etymologies? 58
2.4 Examples of lexicalization 51	

In chapter 1 we encountered some of the main characteristics of etymology, its aims, and some important features of its methodology. We considered some examples of change in word meaning and change in word form, and began to look at some of the mechanisms by which both of these occur. We will return to these topics in more detail later. In this chapter and the next we will take a closer look at the main objects of study in etymological research, words. In etymological dictionaries a ‘word’ stands at the head of each dictionary entry, and the status and selection of these words can seem to be a given. However, the identification of words as coherent entities for study raises a number of quite complex questions. Additionally, selection of which words to concentrate on is a far from trivial matter.

2.1 What are words?

2.1.1 Problems of definition

So far in this book I have taken the term ‘word’ rather for granted, as being a self-evident one which any reader will readily understand. The concept is very familiar to a non-specialist, and the term forms part of general vocabulary and so does not have to be learnt by beginners in linguistics,

unlike phoneme, morpheme, etc. In literate societies lay conceptions of word boundaries (i.e. where one word ends and another begins) are often very much bound up with literacy and the rules of various writing systems, but there is also at least some evidence that non-literate speakers of languages with no written form also have intuitions about word boundaries, as do children who have not yet learned to read in literate societies.¹ However, it is also notoriously difficult to define a ‘word’ in a way which makes sense consistently at all levels of linguistic analysis. Specialists in morphology and also in phonology often grapple with this particular problem, and a full discussion would take up much more space than is available to us here. The discussion that follows will be brief, and will focus on those aspects which most affect etymological research.²

2.1.1.1 *Spelling* A non-specialist from most modern literate societies who is asked what a word is will probably say that the words in a sentence are the things written with a space on either side. This definition is unsatisfactory for linguists for various reasons. Firstly, not all languages have a written form, and even when they do they do not necessarily separate words. Certainly, the way that many languages are written tells us something about writers’ intuitions about what constitute words, but a definition on this basis runs the risk of circularity, and is also detached from any analysis of linguistic structure: by this criterion, words are the things that people write as separate words (i.e. with spaces between them) because they perceive them as separate words (whatever that may mean).

Written language also tends to be rather inconsistent in its treatment of certain kinds of units. Any survey even of published written English will show very considerable variation in whether some combinations of two nouns are written as a solid, or with a hyphen, or with a space between the two elements. Thus *lunchbox* can also appear as either *lunch box* or *lunch-box*, and even dictionaries do not agree on which to list as a preferred spelling. We would have to resort to some very odd reasoning to argue that *lunchbox* is one word but *lunch box* is two: both have the same meaning and behave the same way syntactically, as does *lunch-box*, and in the spoken language the pronunciation is the same for all three. This leads to the

¹ See further Bauer (2003) 57, Sapir (1921) 34–5.

² For detailed discussion of most of the points in this section see e.g. Bauer (2003), which I have largely followed here, or (with some slightly different perspectives) Adams (2001: 2–5), Booij (2007: 281–94), Plag (2003: 4–9).

fairly obvious conclusion that we are looking at three different spellings of precisely the same linguistic unit.

2.1.1.2 Meaning One useful and conventional way of thinking about words as linguistic units is that a word is a linguistic sign which has both form and meaning. (We will come to the very important concept of the arbitrariness of this linguistic sign in chapter 4.) Linguistic meaning is expressed by the combination of units in a sentence. This might seem to give us a shortcut to a definition of a word: words are minimal units of meaning in a sentence. However, a little reflection will present us with some major problems. It is not always possible to infer the established, conventional or institutionalized meaning of phrases from their constituent words: consider idioms like *it's raining cats and dogs* (and see further section 2.1.5 below). There is also ample evidence that people often analyse the morphological composition of unfamiliar complex words as and when they hear them in order to interpret their meaning, and that they do this as part of their general competence as speakers of a language. For instance, if someone knows the word *vinaceous* 'of the colour of red wine' they are unlikely to have any more difficulty in understanding the derivative formations *vinaceousness* or *vinaceously* than the phrase *very vinaceous*, although they will probably never have encountered these particular derivative words before. (Both words are extremely rare, and even a Google search shows only a couple of examples of each.)

2.1.1.3 Phonological criteria Phonological criteria can provide very useful evidence about word boundaries. In some languages, probably including proto-Germanic at one point in its history, stress regularly falls at the beginning of a word. (In proto-Germanic more accurately on the first syllable of a lexical root, rather than on prefixes.) In some other languages, such as modern English, each word has a particular syllable on which the main stress will normally fall if that word is stressed in a sentence (e.g. *'kindness*, *in'eptitude*, *incon'solable*); but this is not true of all languages.

Some phonological processes apply only at particular positions in a word. In the history of German a sound change occurred by which obstruents were devoiced when they occurred word-finally, but not when they occurred medially or initially, giving rise to a situation in modern German where e.g. *Rat* 'counsel' and *Rad* 'wheel' are homophonous in the nominative singular (both /ra:t/) but not in inflected case forms in which an inflectional ending

follows the obstruent (e.g. genitive singular *Rates* /ra:təs/ 'of counsel' beside *Rades* /ra:dəs/ 'of a wheel').³ Some phonological processes, especially vowel harmony, typically operate across syllable boundaries within a word, but not across word boundaries. (See for example section 7.2.4 on *i*-mutation in the history of English.) However, other processes do apply across word boundaries, such as the assimilatory devoicing in English /haftu:/ as a realization of *have to*. This is usually called external sandhi, following the terminology of the ancient Sanskrit grammarians.

2.1.1.4 Morphological criteria A commonly cited morphological criterion is that words are uninterruptible units, although there are exceptions, as for instance when expletives are inserted in the middle of a word in English, e.g. *absobloominglutely*.

2.1.2 Problems of analysis

In addition to there being no generally accepted and completely satisfactory definition of what constitutes a word, there is also considerable scholarly disagreement about whether some particular linguistic units should be regarded as words or as phrases, i.e. syntactic combinations of more than one word. In English it is notoriously difficult to define what constitutes a compound and what constitutes a phrase. To begin with an unproblematic example, it would normally be accepted that *blackbird* is a compound, and *a black bird* is a noun phrase. *blackbird* has reference to a particular variety of bird, and if someone calls a crow a *blackbird* they will be using the English language in an idiosyncratic way that is unlikely to be understood by anyone else. However, if someone refers to a crow as *a black bird*, then they will be making a simple factual statement, and in grammatical terms we will analyse their utterance as a noun phrase showing *bird* as a head modified by the adjective *black*. Conversely, female and younger male blackbirds are mostly brown. Even white blackbirds sometimes occur, and they are still *blackbirds*, albeit uncharacteristic ones, although they are not *black birds*. However, if we try to extrapolate from this unproblematic example precisely what it is that distinguishes a compound from a phrase, we start to encounter some real difficulties:

³ For discussion of this particular phenomenon from a number of different theoretical standpoints see Lass (1984).

- *blackbird* has a meaning not predictable from its component parts, whereas *black bird* refers very predictably to any bird which is black. But many phrases and idioms also have unpredictable meanings.
- *blackbird* is written without any spaces, *black bird* is written with a space. But compare again *lunchbox*, *lunch-box*, *lunch box*.
- In some languages an adjective will show agreement with a noun in a phrase but will show a bare stem form in a compound, giving a clear morphological criterion for telling phrases from adjective-noun compounds, but this is not the case in other languages such as modern English.
- *blackbird* shows stress on the first element, while *black bird* shows stress on *bird*, the head of the phrase. But consider *blackcurrant*, in American English typically *'blackcurrant*, but in British English typically *black 'currant* (except sometimes as the first element in a compound, when the stress may be shifted, e.g. *'blackcurrant bush*). Consider also idiosyncratic cases, such as street names ending in *street* (e.g. *'Downing Street*, *Coronation Street*, *'Ship Street*) as opposed to those ending in *road*, *lane*, *avenue*, etc. (e.g. *Station Road*, *Cemetery Road*, *Park Lane*, *Shaftesbury Avenue*).⁴

This last point in particular is the subject of much debate, but it is sufficient for our purposes to know that there is as yet no clear consensus.⁵ In the case of adjective-noun compounds, gradability of the adjective can be a safer test, at least if the adjective is gradable:

- We may talk about *a very black bird*, or indeed *a very black blackbird*, but not **a very blackbird*.

However, this criterion often conflicts with what we might predict from the position of the stress. *red admiral*, the name of a type of butterfly, has stress on the second element, suggesting phrasal status, but we cannot speak of *a very red admiral* or *the reddest admiral* (at least, not if we are speaking about the butterfly; either phrase would be perfectly plausible if referring to the left-wing politics or the flushed face of a naval officer).

⁴ For a useful discussion of these see Plag (2005).

⁵ For a recent summary see Bauer (2006a), and also Bauer (1998a); for a sample of rather different views see Booij (2007) or Giegerich (2004).

2.1.3 Why these are not major problems for etymology

I have introduced these issues largely to show that the use of 'word' and 'compound' is not always uncontroversial, and because it is important to realize that the simple statement 'etymologists study the origins of words' may not really be so simple as it at first sounds.

Whatever definition of the term 'word' we adopt, etymologists cannot avoid interesting themselves very closely in many units much larger than the word. Very many phrases have complex meanings and complex histories which require etymological explanation. Furthermore, many single words have their origin in what is sometimes termed the univerbation of what were originally phrasal units consisting of more than one word, e.g.:

- *upon* < *up and on*
- *goodbye* shows a contraction of *God be with you*, with remodelling of the first element after *good day*, *good night*, etc.
- the phrase *at one* > the adverb *atone*, on which the abstract noun *atonement* is formed, which in turn gives rise to the verb *to atone*

In some other languages, such as French, lexicalized phrases frequently occur in meanings which are typically realized by compounds in English, for instance French *sac à main* beside English *handbag*. We can also examine the etymologies of units smaller than the word, for example derivational affixes such as *pre-*, *un-*, *-ness*, etc., and even morphological inflections, although these do raise some rather different issues, which we will explore in chapter 4.

Conversely, if we are studying a contemporary language, or even a past stage which has a large corpus of surviving evidence, then we cannot possibly pay attention to the etymology of every word ever uttered, or even every word ever recorded, in that language, and nor would we want to. As we will see in section 2.2.4, the lexicon of every language is constantly open to new words, formed according to the productive word-forming patterns of that language. Nearly all such new words are immediately transparent in meaning (when heard in the appropriate context) to other speakers of that language. Additionally, nearly all such words fail to enter more extensive usage, and remain 'one-offs' or nonce formations (although the same word may well be formed again, quite separately, by other speakers on other occasions).

2.1.4 Word forms and word meanings

If we return to the expression of meaning by words, we can observe that some words, like *a* or *the*, have grammatical content but no other meaning content. Other words, like *haddock* or *ankle*, have clear meaning content. Many words have multiple established meanings, i.e. they are polysemous, and we can only tell which meaning is intended from the context of a particular utterance. For example, we can speak of someone working in an *office* (a physical place) or holding an elected *office* (an abstract social role), or we can say that a container is *full* (there is no room left in it) or that the moon is *full* (none of the side turned towards the earth is in shade). In fact, meanings are often stretched or extended in particular contexts. It is only when particular new or extended meanings of words in particular contexts become institutionalized, i.e. used fairly frequently by different speakers of a language, and perhaps extended to other contexts, that they begin to be recorded in dictionaries. We will return to this point and its importance for etymological research in chapter 8.

Additionally, we need to distinguish between different homonyms, i.e. quite separate words which happen to be identical in form. For instance, distinct homonyms are shown by *file* 'type of metal tool' (of Germanic origin) and *file* 'set of documents' (a borrowing from French). In this instance the words are distinct from a synchronic point of view, since there is no semantic common ground between the meanings which they realize, and also from a diachronic point of view, since they have different histories. However, these two criteria do not always coincide, as we will explore in detail in section 3.3.

Meaning is also expressed by the inflections of a word, e.g. in the singular/plural distinction between *giraffe/giraffes*, *board/boards*, *fish/fishes*, *man/men*, etc. Technically, these inflected forms are distinct word forms, which belong to a single unit called a lexeme. In order to identify the lexeme to which the word forms *giraffe* and *giraffes* both correspond, we normally use what is called the citation form, i.e. the form that we can look up in a dictionary. So *giraffe* is the citation form of the lexeme which has the word forms *giraffe* and *giraffes* (also *giraffe's*, *giraffes'*). Sometimes small capitals are used to identify lexemes, e.g. **GIRAFFE**, **MAN**. Note that in the case of *man/men* the morphological relationship is realized by variation in the stem vowel, rather than by an inflectional affix (see further section 4.4.1).

Sometimes we find the phenomenon known as suppletion, where word forms of different historical origins stand in the same sort of relationship, within a grammatical paradigm, as inflected forms like *giraffe* and *giraffes* do to one another. Thus, *was* and *is* are not inflected forms of *be* (they are of a quite different historical origin), but they stand in the same paradigmatic relationship to it as *opened* and *opens* do to *open*. Similarly, *worse* and *worst* stand in the same paradigmatic relationship to *bad* as *poorer* and *poorest* do to *poor*. We can say that *be*, *was*, and *is* (and also *are*) are word forms of the lexeme **BE**, and that *worse* and *worst* are word forms of the lexeme **BAD** (and also of the lexeme **BADLY**). Interestingly, in the case of *worse* and *worst* this pattern is relatively modern. Both forms go back to the Old English period (Old English *wyrsa* and *wyrst*), and they have been the antonyms of *better* and *best* (Old English *betra* and *betsi*) throughout their history in English, but the adjective in the general sense 'bad' to which they correspond (again suppletively) as comparative and superlative in Old English is *yfel* (modern English *evil*). In early Middle English we find a new adjective *ill* in many of the same senses as *evil*, and *worse* and *worst* are also found as its comparative and superlative. Finally, *bad* becomes increasingly common in senses formerly expressed by *evil* and *ill*, and gradually *worse* and *worst* become established as its comparative and superlative forms. However, there is a long transitional period in which *worse* and *worst* are found in paradigmatic relationships with all of these three words, e.g. we find examples of *from evil to worse, from ill to worse, and from bad to worse*. Thus patterns of suppletion can vary over time, and can also vary in the usage of particular individuals or speech communities within a particular period.

Suppletion is quite different from the phenomenon where different variants realize the same grammatical form of a single lexical item. Modern standardized written languages do much to disguise this sort of variation, but consider the regional differences in pronunciation between for example /tu:θ/ *tooth* in the English West Midlands as against /tu:θ/ elsewhere, or the variation in the pronunciation of *either* as /i:ðə/ or /aɪðə/ in the speech of different individuals in both Britain and the US. This is an issue that we will look at in much more detail in chapter 3.

In this book, I will normally use 'word' rather loosely in the sense 'lexeme', and I will refer to words by their citation forms. This is not normally a problem in etymological work, so long as we have a more sophisticated terminology available for instances where we need to tease the various

distinctions apart more carefully, and so long as we remain aware of the bundle of different forms and meanings which a single word may show.

2.1.5 Idioms

As we have noted, units larger than a single word also often have conventional or institutionalized meaning which is not predictable from their component parts. Idioms are by their nature constructions which are stored in one's memory and form part of one's competence in speaking a particular language, even if this only involves selection of the correct preposition or adverb in verbal constructions such as *to sober up*, or selection between for example *to engage in* 'to participate in' and *to engage with* 'to establish a meaningful contact or connection with'. In these particular cases it might be possible to interpret the meaning of the expression correctly even if one has not encountered it before, i.e. to apprehend it passively even if it lies outside one's active competence, but it is questionable how far most speakers ever stop to analyse idiomatic expressions such as *to catch up on*, *to give (something) up*, *to leave off (doing something)*, *on the one hand...on the other hand*, *to run (someone) to ground*.

There is thus a very strong case for listing idiomatic expressions in dictionaries, so long as they are in sufficiently common use. They are often denoted technically by the broader term lexical item, as distinct from individual words or lexemes. However, not every lexical item that is listed in a dictionary automatically requires etymological investigation. We may feel that constructions such as *to engage in* and *to engage with* will normally be outside the scope of etymological research. However, some of the examples given above are less clear-cut. Understanding of the origin of the idiom *on the one hand...on the other hand* is helped by knowing that *hand* in earlier use had the senses 'side of the body' and more generally 'side, direction' (e.g. in an example from 1548 'on the other hand or side of the gate'⁶). The origin of *to run (someone) to ground* is understandable only when one realizes it originated in the specialist language of fox-hunting, referring to hounds running a fox to its burrow or earth. Many other idioms similarly rely on conventional metaphors which may or may not become opaque as a result of technological or cultural change, e.g. *to run out of steam* 'to lose impetus or enthusiasm' (which originated in the age of the steam engine) or

⁶ See *OED* at *hand* n.¹ sense B.4.

to have shot one's bolt 'to have done all that one could do' (which originated in the age of the crossbow), while others reflect otherwise obsolete or near-obsolete senses of words, e.g. *to cut a caper* 'to make a playful, skipping movement, to act ridiculously' (showing *cut* in the sense 'to perform or execute' and *caper* 'a frolicsome leap, especially in dancing'). Some originate in quotations, e.g. biblical quotations or paraphrase such as *to turn the other cheek* or *to take someone's name in vain*, or quotations from Shakespeare such as *the milk of human kindness* or *the world's your oyster*. (This last example becomes rather less opaque when the metaphor is heard in its original fuller context: *The Merry Wives of Windsor* II. ii. Falstaff: *I will not lend thee a penny*. Pistol: *Why then, the world's mine oyster Which I with sword will open*.) We will take up the difficult issues that such cases raise about the role of non-linguistic, encyclopedic knowledge in etymological research in chapters 8 and 9. Some idioms remain stubbornly resistant to all attempts to explain their origin, e.g. *Bob's your uncle* 'there you are' (said in a situation where a task becomes easy to complete) or *the full monty* 'everything which is necessary, appropriate, or possible, the works'.

Sometimes idioms arise from remodelling of earlier expressions. For instance, the rather opaque expression *to have another thing coming* (as in, *If you think you can get away with that, you have another thing coming*) becomes much more readily explicable when a little etymological research reveals that it is an alteration of earlier *to have another think coming*, in which *think* 'action of thinking' has been replaced by the commoner word *thing* (perhaps as a result of homophony in casual speech), even though the outcome is an idiom which is semantically much more opaque.

2.2 How new words arise

As well as looking at word forms and how they realize meaning, we can look at structure within the word, and in a book on etymology it makes most sense to do this primarily from the point of view of word origins, and thus to take a preliminary look at how new words enter a language.

2.2.1 Monomorphemic words and complex words

An important initial distinction is between monomorphemic words and complex words. As the name implies, monomorphemic words are composed of only a single morpheme or meaningful unit. Examples which we

encountered in chapter 1 include *friar*, *sad*, and *deer*: at least in modern English, these words are unanalysable units, and if we understand them it must either be because they are stored as meaningful units in our memory or because a given context in which they appear makes their meaning obvious. Other words are clearly analysable, such as *happiness*, *steadiness*, *freshness*, or *closeness*, although compare *highness*, which is analysable but not transparent, at least not in its use as an honorific title. It is important to note that it is not necessarily the case that these words are not also stored in our memory; but we can analyse all of them from their component parts (*happy*, *steady*, *fresh*, *close*, *high*, and the suffix *-ness*), and all except *highness* are semantically transparent. Throughout this section we will return often to the following questions:

- (i) Do words of this type need to be included in an etymological dictionary?
- (ii) Are words of this type interesting to etymologists?

We can immediately conclude that any monomorphemic words in a language will need to be included in any etymological dictionary which claims to be at all comprehensive, and that they will be of obvious interest to etymologists: from the point of view of the contemporary language they are stand-alone items which must have an origin and history which we will want to trace. A good case can also be made for including all affixes which are found in analysable words. (We will return to the etymologies of affixes in chapter 4.) The situation is much less clear-cut with words which are analysable, and we will need to look at a number of issues before we will be in any position to address this question.

2.2.2 Borrowed words

Words which have been borrowed from another language are typically monomorphemic, such as *friar* in chapter 1. However, some are analysable, usually because each of the elements of which they are composed have also been borrowed. For instance, English *municipality* is a borrowing from French *municipalité*, but it is analysable, because *municipal* has also been borrowed, and the ending *-ity* is familiar as the ending of a great many abstract nouns borrowed from French nouns in *-ité* (and/or Latin nouns in *-itās*) and has also become productive within English. Often it is difficult to determine whether complex words of this type show borrowing at all:

we will examine some of the issues concerned in sections 5.1 and 6.8. At a greater time depth, or where there is little data, borrowing generally becomes much more difficult to detect, and we will look at some of the implications of this in chapter 7.

Lexical borrowing is probably found to at least some extent in all languages, although the extent varies greatly (see chapter 5). We may fairly safely conclude that all words which have been borrowed will be of some interest to an etymologist, since we will want to find out how, when, and from which other language they have been borrowed. As we will see in chapters 5 and 6, these are very often difficult questions to answer, because of lack of evidence and/or difficulties of analysis. If we are even reasonably inquisitive about the ulterior histories of words, we will also want to delve further than this, and discover whether the word in the donor language is itself analysable and what its history is.

It may thus seem that all borrowed words will automatically need to be included in any etymological dictionary which attempts to be comprehensive. However, this presents some problems, both of a practical and of a theoretical nature. Fundamentally, words are borrowed, just as they are used, by individuals, not by 'languages', and we may find that very different selections of borrowed words belong to the vocabularies of particular social groups, geographical areas, etc., and even to the vocabularies of individuals within those groups, areas, etc.

Lexical borrowing is one of the many areas in which we can observe the open-ended nature of the lexicon of a language. Even if we restrict our focus to the usage of monolingual speakers, individuals have different interests or pursuits which will bring them into contact with different words from other languages. For example, very often people will have different enthusiasms for different cuisines, and accordingly they will have slightly different (active or passive) vocabularies of food terms. The Italian bread name *focaccia* has reasonable currency in contemporary British English, and also in many other varieties of English. The *OED* has an entry for this word as an English borrowing from Italian, with illustrative quotations dating back to 1881. However, the early quotations given in the *OED* present the word as an unusual item which authors feel the need to explain to their readers, and it is not until relatively recent years that we find examples reflecting more general currency of the word.

This particular example of a food term imported from another culture may seem an obvious symptom of modern cosmopolitanism and hence not

encountered in chapter 1 include *friar*, *sad*, and *deer*: at least in modern English, these words are unanalysable units, and if we understand them it must either be because they are stored as meaningful units in our memory or because a given context in which they appear makes their meaning obvious. Other words are clearly analysable, such as *happiness*, *steadiness*, *freshness*, or *closeness*, although compare *highness*, which is analysable but not transparent, at least not in its use as an honorific title. It is important to note that it is not necessarily the case that these words are not also stored in our memory; but we can analyse all of them from their component parts (*happy*, *steady*, *fresh*, *close*, *high*, and the suffix *-ness*), and all except *highness* are semantically transparent. Throughout this section we will return often to the following questions:

- (i) Do words of this type need to be included in an etymological dictionary?
- (ii) Are words of this type interesting to etymologists?

We can immediately conclude that any monomorphemic words in a language will need to be included in any etymological dictionary which claims to be at all comprehensive, and that they will be of obvious interest to etymologists: from the point of view of the contemporary language they are stand-alone items which must have an origin and history which we will want to trace. A good case can also be made for including all affixes which are found in analysable words. (We will return to the etymologies of affixes in chapter 4.) The situation is much less clear-cut with words which are analysable, and we will need to look at a number of issues before we will be in any position to address this question.

2.2.2 Borrowed words

Words which have been borrowed from another language are typically monomorphemic, such as *friar* in chapter 1. However, some are analysable, usually because each of the elements of which they are composed have also been borrowed. For instance, English *municipality* is a borrowing from French *municipalité*, but it is analysable, because *municipal* has also been borrowed, and the ending *-ity* is familiar as the ending of a great many abstract nouns borrowed from French nouns in *-ité* (and/or Latin nouns in *-itās*) and has also become productive within English. Often it is difficult to determine whether complex words of this type show borrowing at all:

we will examine some of the issues concerned in sections 5.1 and 6.8. At a greater time depth, or where there is little data, borrowing generally becomes much more difficult to detect, and we will look at some of the implications of this in chapter 7.

Lexical borrowing is probably found to at least some extent in all languages, although the extent varies greatly (see chapter 5). We may fairly safely conclude that all words which have been borrowed will be of some interest to an etymologist, since we will want to find out how, when, and from which other language they have been borrowed. As we will see in chapters 5 and 6, these are very often difficult questions to answer, because of lack of evidence and/or difficulties of analysis. If we are even reasonably inquisitive about the ulterior histories of words, we will also want to delve further than this, and discover whether the word in the donor language is itself analysable and what its history is.

It may thus seem that all borrowed words will automatically need to be included in any etymological dictionary which attempts to be comprehensive. However, this presents some problems, both of a practical and of a theoretical nature. Fundamentally, words are borrowed, just as they are used, by individuals, not by 'languages', and we may find that very different selections of borrowed words belong to the vocabularies of particular social groups, geographical areas, etc., and even to the vocabularies of individuals within those groups, areas, etc.

Lexical borrowing is one of the many areas in which we can observe the open-ended nature of the lexicon of a language. Even if we restrict our focus to the usage of monolingual speakers, individuals have different interests or pursuits which will bring them into contact with different words from other languages. For example, very often people will have different enthusiasms for different cuisines, and accordingly they will have slightly different (active or passive) vocabularies of food terms. The Italian bread name *focaccia* has reasonable currency in contemporary British English, and also in many other varieties of English. The *OED* has an entry for this word as an English borrowing from Italian, with illustrative quotations dating back to 1881. However, the early quotations given in the *OED* present the word as an unusual item which authors feel the need to explain to their readers, and it is not until relatively recent years that we find examples reflecting more general currency of the word.

This particular example of a food term imported from another culture may seem an obvious symptom of modern cosmopolitanism and hence not

applicable to earlier historical periods, but in fact we find that imported items (foodstuffs, items of manufacture, etc.) are a very frequent source of new borrowings in almost all cultures and almost all historical periods. Inevitably, whenever we have a reasonably large body of historical data, we can ask, but not necessarily answer, the same sorts of questions about precisely whose vocabulary particular borrowed words may or may not have belonged to in a given place and time.⁷

Additionally, we should remember that mobility of individuals or groups between different speech communities is hardly a modern innovation, and much recent work in linguistics has highlighted just how typical (and indeed normal) bilingualism and multilingualism are in many parts of the world today and have probably been at all times in the past. We will look in chapter 6 at the rather vexed question of whether switches between languages by bilingual speakers actually show borrowing at all, and if not how great the connection between the processes is. However, as soon as we are dealing with a situation where people speak more than one language, it is fairly certain that there will be some interchange of lexis between the two languages, even if this is restricted to technical or specialist registers.

We can thus see that in any language a core of well-established borrowings is likely to be surrounded by a periphery of much less well-established ones. Wherever there is a language contact situation, any large sample of actual usage is likely to include nonce, one-off, borrowings which do not show more general adoption (although the same word may well occur as a nonce borrowing on multiple separate occasions).

The open-ended nature of the lexicon of any language becomes yet more apparent if we now consider new words which are formed within a language rather than borrowed from another language.

2.2.3 New formations: aspects of affixation and compounding

One very common method of forming new words is by affixation (or derivation). Both prefixes (which involve addition of material at the beginning of a base, e.g. *un-*, *in-*, *pre-*) and suffixes (which involve addition of material at the end of a base, e.g. *-ness*, *-ment*, *-ly*) are common in very many languages. We will look at both in detail in chapter 4. Much more rarely infixes are found,

⁷ For a detailed discussion of the general importance in etymological research of paying attention to how words can shift between specialist vocabularies and general usage see von Wartburg (1969) 107–14.

which interrupt a morphological base; in its inflectional morphology (rather than its derivational morphology) proto-Indo-European probably had an infix **-n-* which formed present stems as part of its verbal system, reflected in for instance English *stand*.⁸ We sometimes also find circumfixes, which involve addition simultaneously of material at the beginning and the end of a base; by some analyses a circumfix is shown by the *ge- -t* which is added to the stem of weak verbs in modern German to form the past participle, as e.g. *gefragt* ‘asked’, past participle of *fragen* ‘to ask’ (stem *frag-*), although again this belongs to inflectional rather than derivational morphology (unless we take the past participle to be an adjective formed on a verbal stem).⁹ In section 4.4.1 we will look at ablaut, the systematic employment of variation in a stem vowel to mark different morphological or derivational categories.

Another very common process is, as we have seen, compounding. One important thing that compounding and affixation have in common is that the resulting word is ‘bigger’ than the elements from which it is formed. The word form thus enacts the semantic relationship between a base word and a compound or derivative. When we encounter a new compound or derivative, we recognize that it contains a base word plus something else (an affix or another base word). This suggests to us that the new word will have a meaning related to that of the base word but modified in some way. This sort of relationship between word form and word meaning is termed iconic. (See further section 4.5.)

2.2.4 Productivity

If an affix is productive, i.e. capable of forming new words, it can sometimes generate an enormous number of new word forms.¹⁰ The process may be open-ended; this is particularly clearly illustrated by affixes which

⁸ See Plag (2003) 101–4 for an argument that derivational infixation is shown in modern English in expletive insertion of the sort shown by *absobloominghately* (see also section 2.1.1.4). On the distinction between derivation and inflection see Plag (2003) 14–16.

⁹ Circumfixation should be distinguished from the simultaneous addition of both a prefix and a suffix in cases like *decaffeinate* < *de-* + *caffein* + *-ate*, where *de-* and *-ate* remain distinct affixes with distinct meaning and function. Such formations are normally called parasynthetic.

¹⁰ For a detailed analysis of morphological productivity see Bauer (2001); a useful account, with further references, is also given by Plag (2006). Productivity is a difficult and somewhat disputed term, and is not used in exactly the same way by all scholars.

can attach to names to form new lexical items, like *-ism* in *Thatcherism*, *Stalinism*, etc. New derivational formations may be formed at almost any time within the context of a particular utterance, and be understood within the context of that utterance. An influential study in this area is Baayen and Renouf (1996), in which the authors looked at frequencies of word forms with the affixes *-ly*, *-ness*, *-ity*, *un-*, and *in-* in the British newspaper *The Times* over a period between 1989 and 1993. They found very large numbers of forms which occurred only once in this corpus, and which were not recorded in any dictionaries.¹¹ Their findings point strongly to very many of these formations being genuinely one-off nonce uses (examples include *archdukely*, *composerly*, *conductory*), which readers of the newspaper process effortlessly by means of their knowledge of the productive word-forming patterns of the language. These words are not stored in the reader's memory, and yet they pose no problems for interpretation. Baayen and Renouf concentrated on words formed with derivational suffixes, but we can find just as great if not greater facility in the production of new compounds in English, which will be readily interpreted and understood by a hearer even if they are being encountered for the first time. (Of course, as noted in section 2.1.2, some scholars would anyway interpret at least some of these as showing phrases rather than compounds.)

Many words can be processed as they are encountered in context, drawing on the hearer's or reader's knowledge of the word-forming rules of the language. We can compare this to the way that any of an almost infinite number of different possible sentences can be interpreted (normally quite unconsciously) through the hearer/reader's knowledge of the syntactic patterns of a language. Other words are stored in our memory, including some which are perfectly transparent and analysable. Some people will encounter and/or use some words regularly which some other people never encounter: Baayen and Renouf's *composerly*, *conductory*, and even *archdukely* may be part of everyday discourse for some people. Many linguists invoke the concept of a mental lexicon, which will probably differ at least slightly for each individual speaker of a language.¹²

If we take the view that an etymologist's task is to account for the origin and development of the lexicon of a language, then this begins to appear

¹¹ Additionally, they found that formations with the native, non-borrowed affixes *-ly*, *-ness*, and *un-* appeared to be much more frequent than would be suggested if one worked simply from the wordlists of dictionaries.

¹² For an overview of this topic see Aitchison (2003).

an impossible endeavour if new words are continually arising in the speech or writing of different individual speakers and writers on a daily basis, and if different individuals will have different lexical items stored in their memories. A more useful framework for defining the main focus of an etymologist's work is provided by the concepts of transparent and opaque (and also analysable and unanalysable) meanings and word forms which we have already encountered, and by the diachronic processes of institutionalization and lexicalization by which these commonly come about.

2.3 Lexicalization

A distinction is often made between nonce formations, institutionalized words, and lexicalized words. (More strictly, we should speak of lexical items here, so as to allow phrases to be included in the same framework.) Some scholars regard these as stages in a process which words may (but need not) undergo:¹³

nonce formation > institutionalization > lexicalization

Nonce formations are ad hoc coinages by individuals in particular circumstances, the majority of which will never gain any wider currency, such as the words encountered in the Baayen and Renouf study which we looked at in the preceding section. Institutionalized words, while they remain (at least relatively) transparent, are used conventionally within a certain speech community in a given context or with a fairly specific meaning. Lexicalized words are opaque – in meaning, or composition, or both.

lunchbox is, compositionally, a transparent compound of *lunch* and *box*, and we are not surprised to find that it denotes a box for transporting one's lunch. However, the definition in the *OED* suggests that it has some more conventional meaning characteristics than this:

A container designed to carry a packed lunch (or other meal). Formerly, any of various types and sizes of receptacle, sometimes also carrying crockery, etc., but now usually a small lidded box for food.

From the accompanying illustrative quotations in the *OED* we see that the modern use is most often specifically to denote such a box used for

¹³ See for example Bauer (1983) 45–50. For a thorough overview of this field see Brinton and Traugott (2005).

transporting lunch to a workplace or, especially, to school. This suggests that it is an institutionalized word for this item. If someone called the same thing a **foodbox* or a **lunchcarrier* we might understand from context what was meant, but it would strike us as not being the right word: in fact, it would be a nonce formation which we would interpret from the context in which it occurred, and we would soon conclude that it was intended as a synonym of the institutionalized word *lunchbox*. To take another example from the same semantic field, not many decades ago many British workers, particularly miners, carried their lunch in a metal container, usually called a *snap-tin*. *Snap* was a word for a light meal, and hence the compound was transparent, if institutionalized. However, today *snap-tins* (i.e. the physical objects) tend to be encountered only as collectables or museum pieces, and the word itself is encountered either as the name associated with these artefacts or in recollections of a bygone world. Internet discussions sometimes speculate on the meaning of *snap* in the compound, or feel the need to explain the word's origin. In fact it shows *snap* 'light or packed lunch', itself a metaphorical use of *snap* 'quick or sudden closing of the jaws or teeth in biting' (compare *a bite to eat*), which is in turn related to the verb *snap*. We could imagine an alternative scenario in which *snap-tin* was formed directly from the verb *snap*, perhaps because of its lid snapping shut when closing, and in which *snap* 'light or packed lunch' was so called because it was carried in a *snap-tin*; it is the historical record that shows us otherwise, rather than anything that we can intuit from the modern use of the word. Hence we see that for some speakers at least the term is not just institutionalized but lexicalized: they call this sort of box a *snap-tin*, but at least some of them are not sure why.

Lexicalization is an important process in any study of etymology, because it is key to explaining many word histories. In the case of *snap-tin* it is both the meaning and the composition of the word that have become not just institutionalized but opaque: someone encountering the word *lunchbox* for the first time will have a good idea of what a *lunchbox* is simply from the composition of the word (even though they may miss some of the nuances of the institutionalized meaning), but someone encountering the word *snap-tin* for the first time is going to need to make careful use of information from the context of the wider utterance in order to work out what the word denotes, and will have little idea which out of numerous possible meanings *snap* shows in this word.

There are various different processes by which a word may become lexicalized. The most typical are:

- (1) Semantic change occurs, either in the lexicalized word or in one or more of its constituent elements (i.e. the words, affixes, etc. from which it is composed)
- (2) The word may become 'orphaned' as a result of one or more of its constituent elements becoming obsolete
- (3) Changes in word form (typically through the operation of sound change) may obscure the relationship between the word and its constituent elements

Often, more than one of these processes is found in a single word history, and it is sometimes hard to tell in what order they occurred. It is also often difficult to tell when a word became opaque, and a word may well remain transparent for some speakers when it is already opaque for others. Any change which results in the original morphological composition of a word becoming opaque is sometimes referred to as demorphemization or demorphologization (see e.g. Brinton and Traugott (2005) 52–4): for instance, in the case of *handiwork* which we encountered in section 1.3.2, the prefix *ge-* in the medial syllable has become opaque, as a result of loss of *i-* (< *ge-*) where it occurred word initially. (For further discussion of the prefix *ge-* see section 4.1.2.)

2.4 Examples of lexicalization

So far we have looked at *lunchbox*, a word which has an institutionalized meaning but is of transparent composition, and *snap-tin*, which is opaque for some speakers, but is also now a rather rare word. However, very many perfectly common words have shown a historical development from being analysable and transparent to being completely unanalysable and opaque.

husband is a word with something of a 'disguised' history. As a modern English word it is unanalysable and indisputably monomorphemic, but this is not true at all points in its history. It occurs in its modern sense 'a man joined to a woman by marriage' from the thirteenth century. The word first appears, as late Old English *hūsbonda*, in the eleventh century, in the sense 'the master of a house, the male head of a household'. It is a borrowing from Old Norse *hūsbondi* (with assimilation to the class of weak masculine

nouns, hence the ending *-a* in the nominative case in Old English). However, the composition of *hūsbonda* would have been transparent to speakers of Old English, since the first element *hūs* is identical in form and meaning to its Old English cognate *hūs* 'house', and the second element *bōndi* 'peasant owning his own house and land, freeholder, franklin, yeoman' was also borrowed into late Old English as *bonda* (i.e. again with assimilation to the class of weak nouns). Indeed, as with many borrowed compounds, it could alternatively be argued that *hūsbonda* was formed in Old English from *hūs* and *bonda* on the model of Old Norse *hūsbōndi* (see section 5.1 for discussion of this topic, and also 5.2 for terminological complications to do with the term 'Old Norse'). In the Middle English period the vowel in the first syllable of the English word was shortened as part of a regular process of shortening before consonant clusters. Consequently it did not participate in the Great Vowel Shift affecting long vowels, as *house* did, with the result that the first element of the word became opaque, since *hus-* /huz/ (later /hʌz/ or /həz/) showed no obvious relation to *house* /haus/. (We will return to the Great Vowel Shift in section 7.2.3.) Old English *bonda* is continued by Middle English and early modern English *bonde*, *bond*, but the word is now obsolete. *husband* has thus become opaque as a result of:

- semantic specialization
- formal change in its first syllable (and different formal change in the parent word *house*)
- obsolescence of the word which forms its second element

As is typical in such cases, it would be very difficult to identify exactly when the word ceased to be transparent. If we consider that a language is something spoken by large numbers of individuals, we can see that it will be impossible ever to pin down a precise moment when change occurred, because the relevant changes in word form and word meaning will not have occurred for all speakers at the same time. In fact, the evidence of spelling forms and recorded meanings in the *OED* suggests considerable overlap both between different meanings and between different forms in the history of this word, just as we find in a great many other cases as well. Additionally, if we are trying to assess whether people in the past perceived a word as a transparent compound, we will always be engaging in guesswork to some degree: we can show that in such and such a period the language contained relevant word forms, so that someone so minded could make the connection between simplex word and compound word, but we cannot demonstrate

that this actually happened. Thus we may in some cases be able to show when a word ceased to be analysable, but we cannot show when it ceased to be analysed. (It can be difficult to gauge whether a word is perceived as a transparent compound even by contemporary speakers.)

To take another example, English *lord* was also originally a compound, even though in modern English it is both monomorphemic and monosyllabic. It is recorded in Old English most commonly in the form *hlāfword*, but also once in the form *hlāfweard*. It has a range of meanings in Old English, including 'master', 'prince', 'chief', 'sovereign', 'feudal superior', and even 'husband', but probably its original meaning was 'the male head of a household'. Although poorly attested, *hlāfweard* is almost certainly the earlier form of the word, showing a compound of *hlāf* (modern English *loaf*) and *weard* 'keeper' (modern English *ward*); the original meaning was thus metaphorical, referring to the role of the head of a household as owner and provider of the food eaten by his servants and dependants. In the more usual Old English form *hlāfword* with reduced second syllable the connection with *weard* is already obscured, and very possibly no connection with *hlāf* was felt either. Certainly, all formal connection with *loaf* is lost in the reduced monosyllabic form *lord* which becomes the usual form from the middle of the Middle English period. *lady* (Old English *hlāfdige*) probably shows a similar origin, <*hlāf* + an otherwise unrecorded word with the meaning 'kneader' ultimately related to *dough*. (In this instance *hlāf* in the Old English word form shows the sound change known as *i*-mutation: see section 7.2.4.)

In each of these cases changes in word form have played a major part in making the etymologies and early meanings of the words opaque, i.e. demorphologization has occurred. In other cases change in meaning is much more important than change in word form. The word *handsome* is formed from *hand* and the suffix *-some*. This suffix seldom produces new words in modern English: it has become unproductive and now only occurs in occasional analogous nonce formations. The words in which it survives are a rather complex set of lexicalized words in which the suffix shows a number of different relationships with the base word, e.g. *quarrelsome*, *bothersome*, *loathsome*, *fearsome*, *wholesome*, *cumbersome*. However, in all of these cases it remains clear that e.g. *quarrelsome* has some connection with *quarrel* or *quarrelling*, and *bothersome* with *bother* or *bothering*, even if a particular speaker is unfamiliar with the lexicalized meanings 'given to or characterized by quarrelling', 'annoying, causing bother', etc. In some

other cases the parent word has simply become obsolete, as in the case of *winsome* (from Old English *wynn* ‘joy’); viewed synchronically, it has become unanalysable and so a unique morph (more commonly called a cranberry morph, for reasons we will see in section 2.6). In the case of *handsome* the situation is rather different. The first element is *hand*, and this is still very clear from the written form of the word. There is often no /d/ in the spoken form, but careful listening shows that the same applies to *handshake*, *handsaw*, *hands-off*, *hands-on*, *handstand*, and other words with a similar sequence of sounds, as pronouncing dictionaries will confirm, and yet in all of these cases the relationship with *hand* remains perfectly obvious. The crucial difference in the case of *handsome* is the development in meaning that the word has shown. When first found in the fifteenth century the word meant ‘easy to handle or manipulate, or to wield, deal with, or use in any way’, and in the early sixteenth century also ‘handy, ready at hand, convenient, suitable’ (we may compare the semantic history of *handy* already investigated in chapter 1). But these senses are now obsolete in most varieties of English, and the word has passed via the senses ‘apt’, ‘proper’, ‘fitting’ to the core modern senses ‘(especially of a man) good-looking’, ‘(of a number, sum of money, etc.) substantial’. In consequence all semantic connection with *hand* has been lost, and the word has become opaque.

penknife presents an interesting case of a word which is perhaps rather less far down the route of lexicalization. It obviously and transparently denotes a type of knife. However, to the vast majority of modern speakers, it does not have any obvious or transparent connection with pens. The *Oxford Dictionary of English* (revised edition, 2005), a dictionary which takes a synchronic (i.e. non-historical) approach based on a corpus of contemporary usage, boldly defines *penknife* as ‘a small knife with a blade which folds into the handle’. It also offers no etymology for the word, and in my view this could conceivably leave some readers confused about its origin; they might guess wrongly at some connection with *pen* ‘small enclosure for animals’ (reasoning that penknives have some sort of basic out-of-doors function), or they might assume that this kind of folding pocket knife was invented by someone with the surname *Pen* or *Penn*. Or perhaps they will alight on the right *pen*, but with the wrong reasoning, assuming that a *penknife* is a knife which is taken to resemble a pen when folded away. This is perhaps a little unlikely, but most people will probably need to engage in a little lateral historical thinking to arrive at the right answer. It is much more likely that in the ordinary course of events they will give the matter no thought at

all, and regard *penknife* as the specific but inherently uninformative name of a type of knife. The historically based definition in the *OED* (third edition, entry published 2005) informs anyone about the history of the word *penknife* immediately: ‘Originally: a small knife for use in making and mending quill pens (now rare). Now usu.: a pocket knife with one or more blades (and occas. other tools) designed to fold back into the handle when not in use.’ And to avoid any lingering confusion, a brief formal etymology is provided, identifying that the word is indeed a compound of *pen* ‘writing implement’ and *knife*; hence *pen* has in formal terms an objective relation to *knife*, denoting the thing which the knife is (or rather was) used to sharpen. In this case it is the changing use of the denotatum, i.e. technological change in the non-linguistic world, which has been the driving force leading to lexicalization.

A final example will introduce some further themes which we will explore more fully later in this book. The word *acorn* is clearly monomorphemic and unanalysable in modern English. Furthermore it has a satisfying meaning relationship with an easily identified and very tangible entity in the real world. If someone asks us what the word *acorn* means (or more likely, what an acorn is) we can point to an acorn and say ‘it means one of these’. (Although a botanist may note that different types of oak tree in fact have different types of acorns.) However, etymologically the word *acorn* is almost certainly related ultimately to the word *acre*, the modern reflex of Old English *acer* ‘field’. It probably originally had the meaning ‘fruit of the unenclosed land, natural produce of the forest’, although by the date of its earliest recorded appearance in English (in the form *aceren*) its sense has become restricted to ‘acorn’, the fruit of the oak tree, to which the authoritative *Dictionary of Old English* adds ‘perhaps other fruit of similar form, mast’ (that is to say, the fruit of woodland trees, such as acorns, beech mast, etc.). The meaning development, and the relationship between *acorn* and *acre*, become clearer when we look at some of *acorn*’s cognates in other Germanic languages: Dutch *aker* ‘acorn’, Old Norse *akarn* ‘acorn’, Old High German *ackeran* ‘oak or beech mast’, Gothic *akran* ‘fruit’. We have no real way of knowing for certain whether the Anglo-Saxons connected the word with *acre*, but the restricted meaning, and the lack of any metalinguistic comments to the contrary, would suggest quite strongly that they did not. In modern English both the word’s meaning and its form disguise the etymological connection with *acre*, and etymological investigation is required to establish the connection and to trace how the two words subsequently diverged. Interestingly, the word has been subject to various

folk-etymological alterations during its history in English, indicating a desire on the part of language users to establish iconic relationships with other words in the language. (See further section 7.4.5, and also 4.5 on iconicity.) In the seventeenth century we find the form *oke-corn*, in which the word has been remodelled after *oke*, a variant of *oak*, and *corn*. Thus the word's form has been altered in such a way as to make transparent a perceived basic meaning 'corn (or fruit) of the oak', which certainly reflects what an acorn is, but this does not coincide with the word's historical composition. The modern form *acorn* (rather than **akern*) results from this same folk-etymological association with *corn*.

2.5 Apparent reversals of the process

Very occasionally the interaction between the written language and the spoken language may lead to apparent reversal of the lexicalization process. This typically happens in languages which have a standard and long-settled written form. The written language may therefore not reflect changes in word form which have occurred since. Thus *breakfast*, *blackguard*, or *boatswain* all reflect their composition transparently in the written form, but not in the spoken form (/brekfəst/, /blækɡəd/, /bəʊsən/), although since *blackguard* and *boatswain* are both now relatively rare words 'spelling pronunciations' are sometimes heard for each of these, hence /blækga:d/ or /bəʊtsweɪn/ (but /bəʊtsweɪn/ would never occur as the spoken realization of the adapted spelling *bosun*). Such spelling pronunciations can sometimes completely oust an older pronunciation which shows demorphologization, hence /weɪstkəʊt/ rather than /weskit/ is now usual for *waistcoat*, and /fɔ:hed/ is becoming more common than /fɔ:rid/ for *forehead*. We will look in section 7.4 at various other processes such as folk etymology which run counter to lexicalization, since they lead to an increase in compositionality and analysability, and which are therefore sometimes described as showing anti-lexicalization.¹⁴

2.6 Cranberry morphs

If compounds and derivatives are common in a language (as they certainly are in English), this can lead to a certain degree of tolerance of words which have the appearance of being compounds or derivatives but

¹⁴ See for example Brinton and Traugott (2005) 102–3.

in which one of the elements is not analysable. The first element of the word *cranberry* is totally opaque to a speaker of modern English who does not know something about the history of the word, and morphologists often refer to unanalysable morphemes of this kind as cranberry morphs (or alternatively, and less colourfully, as unique morphs).¹⁵ In fact, the word *cranberry* has been opaque for all of its history in English. It shows a seventeenth-century North American English borrowing from another Germanic language, probably Low German, in which the word ultimately shows a cognate of the bird name *crane* and a cognate of *berry*; compare the forms Low German *kranebere*, High German *Kranbeere*. In English, the second element of the word has been remodelled after, or perhaps assimilated to, the English cognate *berry*. As a result the word belongs to a family of words denoting types of (relatively) soft fruit, which also includes such transparent formations as *blackberry* and *blueberry* which both have fairly clear reference to the characteristic appearance of the fruit, although both are clearly institutionalized names. (Someone might hypothetically perceive blueberries as being more black than blue in colour, but that person could not then reasonably expect to be understood if she began to refer to blueberries as *blackberries* without making it very clear that she was making a deliberate departure from conventional linguistic usage.) Various shrubs of the genus *Symporicarpus* (most of them originally native to North America) are normally called *snowberry* in English. Many of these have white berries, and this might seem the obvious reason for the name, but some others have red berries. The name may simply have been transferred from the white-berried type to the red-berried type, and indeed the white-berried type do appear to have been the first to be given this name. However, most snowberries, regardless of colour, bear their berries in winter, and this might suggest a quite different motivation for the name, or alternatively explain how the name could easily be transferred from the white-berried to the red-berried type, if reanalysed as referring to the season when the plants bear their berries. The reason for the *strawberry* being so called is far from obvious; it is normally considered by etymologists that it shows the word *straw* 'stem(s) or stalk(s) of various cereal plants', but various explanations have been suggested to account for this, such as the appearance of the plant's runners, or the appearance of the small seeds on the surface of the fruit, or perhaps the name reflects the cultivation of strawberries

¹⁵ See e.g. Bauer (2003) 48, 50; Booij (2007) 30–1.

on beds of straw to keep the berries off the ground. *Raspberry* is almost certainly a compound of the earlier word *rasp* denoting a raspberry, but without a knowledge of linguistic history we may just as well think that *rasp* is a clipping (or shortening; see section 4.4.3) of *raspberry*; compare some fruiters' use of *straws* for *strawberries*. Thus we see that within this group of words we have a cline of different degrees of analysability: *blackberry* and *blueberry* are obvious descriptive names; *snowberry* may be a less certain case; *strawberry* may be analysable if we stop to think about it, but is hardly likely to be apprehended as a descriptive name in everyday use; *raspberry* may be a longer alternative name for *rasp*, but in synchronic terms the two words are merely synonyms and *rasp* is of no aid in explaining *raspberry* since we do not know the origin of *rasp*; *cranberry*, so far as its existence in English is concerned, is evidently a type of berry, but has a first element with no connections elsewhere in the language, unless we happen to know its further etymology in Low German and work backwards from that to the English cognate *crane*, but that is purely extralinguistic knowledge. If we consider the different types of fruit which these various plants have, it also becomes clear that the concept denoted by *berry* in these formations is not a very precise one; we will return to this point when we consider prototype semantics in section 8.2. Nonetheless, the group of words ending in *-berry* has acquired new members through folk etymology: *naseberry* denoting the sapodilla (a type of fruit which grows on a tree) in fact shows a borrowing from either Spanish *néspera* or Portuguese *nêspera*, with the ending remodelled by folk-etymological association with words ending in *-berry*. (On this etymology compare sections 7.4.5 and 8.8.1. For some further *berry* names see section 9.7. A further interesting example to pursue is *gooseberry*.)

2.7 Which words need etymologies?

We have seen that the lexicon of any language will be extended by speakers in an ad hoc way, as new words are formed by productive word-forming processes such as derivation or compounding. These will normally be understood very easily by other users of the language from their transparent composition and from clues in the context of the utterance which help to explain the meaning. Only a tiny percentage of such introductions are likely to be adopted more widely. If we are working on a dead language or an ear-

lier historical stage of a living language which has a relatively small corpus of surviving material, then we may decide nonetheless to include all of the surviving words in any etymological dictionary or corpus of etymologies, on the not unreasonable assumption that not enough evidence survives to enable us to see which formations are completely trivial and transparent and which are not, and it is therefore much better to be safe than to be sorry. If we are working on a contemporary language, we will certainly not have this luxury. Since the lexicon is almost infinitely extendible, it will be impossible for us to compile a comprehensive list of all of its words, let alone etymologize all of them. But this poses a problem for etymologists: as we will see in subsequent chapters, investigating almost any word history involves either implicitly or explicitly drawing parallels with other word histories, and we will not want to run the risk of neglecting words which may provide crucial information in explaining another etymology.

A useful framework for deciding which words to concentrate our energies on is provided by the concepts of transparency, opacity, and analysability, and by the insights provided by observing the diachronic processes of institutionalization and lexicalization. We might decide that our ideal etymological coverage of a language will include:

- any monomorphemic words (although we may need to reconsider this in the case of languages where variation of the stem vowel is a productive method of realizing derivational relationships: see section 4.4.1)
- any word containing a *cranberry* morph
- any word which has a form which is not explicable by the productive word-formation processes of the language
- any word which is formally analysable but semantically opaque, e.g. *handsome*, *handy*, or for some speakers *penknife*; also idioms such as *to cut a caper*

The last category is particularly difficult to define, since what is opaque for one speaker may not be for another. In each of these categories, our etymological investigations will in many cases show that the current status of a word results from earlier lexicalization, as e.g. *lord*, *lady*, *acorn*, *strawberry*.

We may also decide to add:

- all remaining words with a non-predictable, institutionalized meaning
- all phrases and constructions with institutionalized meanings not readily predictable from the meanings of the words of which they consist