The World's Major Languages

Second Edition

Edited by
Bernard Comrie

The World's Major Languages


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

The World's Major Languages

Second Edition

Edited by
Bernard Comrie

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First edition published 1987 by Croom Helm
Reprinted with revisions and additions in 1989 by Routledge
Second edition published 2009 by Routledge

Second edition first published in paperback 2011
by Routledge
2 Park Square, Milton Park, Abingdon, Oxfordshire OX14 4RN

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

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

First issued in hardback 2015

Typeset in Times New Roman by
Taylor & Francis Books

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

Library of Congress Cataloging in Publication Data
A catalog record for this book has been requested

ISBN 978-0-415-60902-9 (pbk)
ISBN 978-1-138-16993-7 (hbk)
ISBN 978-0-203-30152-4 (ebk)

Contents

Contributors

Daniel Abondolo
University of London

Robert Austerlitz
Columbia University, deceased

Philip Baldi
Pennsylvania State University

Michael Branch
University of London

Wayles Browne
Cornell University

George Cardona
University of Pennsylvania

Ross Clark
University of Auckland

R.G.G. Coleman
University of Cambridge, deceased

Bernard Comrie
Max Planck Institute for Evolutionary Anthropology and University of California Santa Barbara

Greville Corbett
University of Surrey

Scott DeLancey
University of Oregon

Edward Finegan
University of Southern California

John N. Green
University of Bradford

Einar Haugen
Harvard University, deceased

John A. Hawkins
University of California Davis; University of Cambridge

Robert Hetzron
University of California Santa Barbara, deceased

Thomas John Hudak
Arizona State University

Grover Hudson
Michigan State University

Brian D. Joseph
Ohio State University

Yamuna Kachru
University of Illinois at Urbana-Champaign

Alan S. Kaye
California State University Fullerton, deceased

M.H. Klaiman
Independent scholar

Jan G. Kooij
University of Amsterdam

Jaklin Kornfilt
Syracuse University

Charles N. Li
University of California Santa Barbara

Silvia Luraghi
University of Pavia

D.N. MacKenzie
University of Göttingen, deceased

Behrooz Mahmoodi-Bakhtiari
University of Tehran

Graham Mallinson
Independent scholar

Monique Monville-Burston
University of Cyprus

Nam-Kil Kim
University of Southern California

Đình-Hoà Nguyễn
Southern Illinois University at Carbondale, deceased

Paul Newman
University of Michigan

Michael P. Oakes
University of Sunderland

Ọlanikẹ Ọla Orie
Tulane University

Stephen Parkinson
University of Oxford

J.R. Payne
University of Manchester

Douglas Pulleyblank
University of British Columbia

Lawrence A. Reid
University of Hawaii

Paul Schachter
University of California Los Angeles

Masayoshi Shibatani
Rice University

David Short
University of London

Sanford B. Steever
Independent scholar

David Strecker
Independent scholar

Gerald Stone
University of Oxford

Uri Tadmor
Max Planck Institute for Evolutionary Anthropology

Sandra A. Thompson
University of California Los Angeles

Nigel Vincent
University of Manchester

Benji Wald
Independent scholar

Linda R. Waugh
University of Arizona

Julian K. Wheatley
Massachusetts Institute of Technology

Gernot L. Windfuhr
University of Michigan

Preface

In the preface to the first edition of this work, published in 1987, I noted that it represented the fruits of the collaboration of 44 scholars with international reputations ranging across a broad spectrum of the world's languages. This new edition adds two new languages (Javanese and Amharic) and increases the number of such scholars to 52: the authors of the two new chapters plus eight scholars who have either substantially revised or completely rewritten existing chapters. The second edition contains 52 chapters, each dealing with a single language, group of languages or language family, in addition to my Introduction. The chapters that are not completely new have been revised, at times substantially, from their original version published in 1987, although in 19 chapters this has been restricted primarily to my updating the bibliography.

Perhaps the most controversial problem that I, as editor, have continued to face in the second edition has been the choice of languages to be included. My main criterion has, admittedly, been a very subjective one: what languages do I think the volume's readership would expect to find included? In answering this question I have, of course, been guided by more objective criteria, such as the number of speakers of individual languages, whether they are official languages of independent states, whether they are widely used in more than one country, whether they are the bearers of long-standing literary traditions. These criteria often conflict – thus Latin, though long since deprived of native speakers, is included because of its immense cultural importance – and I bear full responsibility, as editor, for the final choice. I acknowledge that the criterion of readership expectation has led me to bias the choice of languages in favour of European languages, although over half of the volume is devoted to languages spoken outside Europe.

The notion of 'major language' is obviously primarily a social characterisation, and the fact that a language is not included in this volume implies no denigration of its importance as a language in its own right: every human language is a manifestation of our species' linguistic faculty and any human language may provide an important contribution to our understanding of Language as a general phenomenon. In the recent development of general linguistics, important contributions have come from the Australian Aboriginal languages Warlpiri and Dyirbal. My own research work has concentrated largely on languages that do not figure in this volume, such as Chukchi of

eastern Siberia, Huichol of Mexico, Maltese of the Mediterranean, Haruai of the New Guinea Highlands and Tsez and Bezhta of the North Caucasus. Other editors might well have come up with different selections of languages, or have used somewhat different criteria. When linguists learned of the death in 1989 of the last speaker of Kamassian, a Uralic language originally spoken in Siberia, who had kept her language alive for decades in her prayers – God being the only other speaker of her language – they may well have wondered whether, for Klavdija Plotnikova, *the* world's major language was not Kamassian.

Contributors were presented with early versions of my own chapters on Slavonic languages and Russian as models for their contributions, but I felt it inappropriate to lay down strict guidelines as to how each individual chapter should be written, although I did ask authors to include at least some material on both the structure of their language and its social background. The main criterion that I asked contributors to follow was: tell the reader what you consider to be the most interesting facts about your language. This has necessarily meant that different chapters highlight different phenomena, e.g. the chapter on English, the role of English as a world language; the chapter on Arabic, the writing system; the chapter on Turkish, the grammatical system. But I believe that this variety has lent strength to the volume, since within the space limitations of what has already grown to be quite a sizeable book it would have been impossible to do justice in a more comprehensive and homogeneous way to each of over 50 languages and language families.

The original impetus for the first edition, published by Croom Helm (since merged into Routledge/Taylor & Francis), came from a meeting with Jonathan Price, at that time the publisher's Linguistics Editor, who also worked with me editorially on the first edition and earned my eternal gratitude. This role was taken on for the second edition by Kate Aker, Director of Development at Routledge Reference, whom I thank not only for her editorial work but also for her patience during delays that were wished upon her. I would also like to thank Ulrike Swientek for her sterling efforts during the production phase of the volume. Finally, I am grateful to Routledge/Taylor & Francis for assuming the editorial costs involved in preparing the second edition.

Some of the authors who contributed to the first edition have since passed on: Robert Austerlitz, R.G.G. Coleman, Einar Haugen, Robert Hetzron, Alan S. Kaye (who completed his revisions shortly before his untimely death), D.N. MacKenzie, Dinh-Hoa Nguyen and D.J. Prentice. I dedicate this second edition to their memory.

Bernard Comrie
Leipzig/Santa Barbara, July 2008

Abbreviations

* The asterisk is used in discussion of historical reconstructions to indicate a reconstructed (non-attested) form. In synchronic discussions, it is used to indicate an ungrammatical item; (* X) means that inclusion of X makes the item ungrammatical; * (X) means that omission of X makes the item ungrammatical.
In the chapters on Tamil and Vietnamese, a subscript numeral *n* after a word in the English translation indicates that that word glosses the *n*th word in the Tamil or Vietnamese example.

In the chapters on the Romance languages, capitals are used to represent Latin or reconstructed Proto-Romance forms.

1	first person	asp.	aspirated
2	second person	aspc.	aspect
3	third person	athem.	athematic
abilit.	abilitative	aux.	auxiliary
abl.	ablative	Av.	Avestan
abs.	absolutive	ben.	beneficiary
abstr.	abstract	BH	Biblical Hebrew
acc.	accusative	BN	B-Norwegian
acr.	actor	Boh.	Bohemian
act.	active	BP	Brazilian Portuguese
act.n.	action nominal	Br.	British
adj.	adjective	c.	common
adv.	adverb	Cast.	Castilian
Alb.	Albanian	Cat.	Catalan
Am.	American	caus.	causative
anim.	animate	cc	class concord
aor.	aorist	Cent.	Central
Ar.	Arabic	circ.	second element of circumfix
Arm.	Armenian		
art.	article	cl.	class(ifier)
Ashk.	Ashkenazi(c)	clit.	clitic

cmpr.	comparative
comp.	complementiser
conj.	conjunction
conjug.	conjugation
conjv.	conjunctive
cont.	contemplated
conv.	converb
cop.	copula
cp	class prefix
crs.	currently relevant state
Cz.	Czech
Da.	Danish
dat.	dative
dbl.	double
decl.	declension
def.	definite
dem. prox.	proximal demonstrative
dem. dist.	distal demonstrative
dent.	dental
deriv. morph.	derivational morpheme
det.	determiner
de-v.	deverbal
dir.	direct
disj.	disjunctive
Dor.	Doric
drc.	directional
du.	dual
dur.	durative
d.v.	dynamic verb
E.	Eastern
Eng.	English
ENHG	Early New High German
EP	European Portuguese
erg.	ergative
ex.	existential-possessive
f.	feminine
fact.	factive
fact. n.	factive nominal
foc.	focus
Fr.	French
fut.	future tense
g.	gender
gen.	genitive
ger.	gerund(ive)
Gk.	Greek
gl.	glottalised
Gmc.	Germanic
Go.	Gothic
gr.	grade
GR	Gallo-Romance
gutt.	guttural
H	High
Hier. Hitt.	Hieroglyphic Hittite
Hitt.	Hittite
hon.	honorific
IE	Indo-European
imper.	imperative
imperf.	imperfect(ive)
inanim.	inanimate
incl.	inclusive
indef.	indefinite
indic.	indicative
indir.	indirect
infin.	infinitive
inst.	instrumental
inter.	interrogative
intr.	intransitive
inv.	inversion particle
irr.	irrational
It.	Italian
i.v.	intransitive verb
L	Low
lab.	labial
Lat.	Latin
Latv.	Latvian
LG	Low German
lig.	ligature
lingu.	lingual
lit.	literally
Lith.	Lithuanian
loc.	locative
m.	masculine
MBs.	Modern Burmese
ME	Middle English
med.	medio-passive
MH	Middle Hebrew
MHG	Middle High German
mid.	middle
MidFr.	Middle French
ModE	Modern English
ModFr.	Modern French
MoH	Modern Hebrew
Mor.	Moravian
MP	Middle Persian
n.	noun
necess.	necessitative

neg.	negative
NHG	New High German
nm.	nominal
NMLZ	noun-forming affix
NN	N-Norwegian
nom.	nominative
noms.	nominalisation
NP	New Persian
nt.	neuter
numb.	number
Nw.	Norwegian
O.	Oscan
OArm.	Old Armenian
obj.	object
obl.	oblique
OBs.	Old Burmese
Oc.	Occitan
OCS	Old Church Slavonic
OE	Old English
OFr.	Old French
OFri.	Old Frisian
OHG	Old High German
OIc.	Old Icelandic
OIr.	Old Irish
OIran.	Old Iranian
OLat.	Old Latin
OLith.	Old Lithuanian
ON	Old Norse
OP	Old Persian
opt.	optative
OPtg.	Old Portuguese
orig.	original(ly)
OS	Old Saxon
OV	object–verb
p.	person
pal.	palatal
part.	participle
pass.	passive
past.	past tense
pat.	patient
PDr.	Proto-Dravidian
perf.	perfect(ive)
pers.	person
PGmc.	Proto-Germanic
PIE	Proto-Indo-European
PIt.	Proto-Italic
Pkt.	Prakrit
pl.	plural
Po.	Polish
pol.	polite
pos.	position
poss.	possessive
pred.	predicate
prep.	preposition
prepl.	prepositional
pres.	present tense
pret.	preterit
prim.	primary
prog.	progressive
proh.	prohibitive
pron.	pronoun
Ptg.	Portuguese
Q	question
rat.	rational
recip.	reciprocal
red.	reduplication
refl.	reflexive
rel.	relative
rep.	reported
res.	result
Ru.	Runic
Rum.	Rumanian
Rus.	Russian
Sard.	Sardinian
SCr.	Serbo-Croat
sec.	secondary
Seph.	Sephardi(c)
sg.	singular
S-J	Sino-Japanese
Skt.	Sanskrit
Slk.	Slovak
s.o.	someone
SOV	subject–object–verb
Sp.	Spanish
spcf.	specifying form
spec.	species
s.t.	something
st.	standard
su.	subject
subj.	subjunctive
sup.	superlative
s.v.	stative verb
SVO	subject–verb–object
Sw.	Swedish
tap.	tense/aspect pronoun
temp.	temporal

them.	thematic	v.	verb
Tk.	Turkish	vd.	voiced
tns.	tense	Ved.	Vedic
Toch.	Tocharian	VL	Vulgar Latin
top.	topic	vls.	voiceless
tr.	transitive	v.n.	verbal noun
transg.	transgressive	VO	verb–object
t.v.	transitive verb	voc.	vocative
U.	Umbrian	VSO	verb–subject–object
unint.	unintentional	YNQ	yes/no question marker

Introduction

Bernard Comrie

1 Preliminary Notions

How many languages are there in the world? What language(s) do they speak in India? What languages have the most speakers? What languages were spoken in Australia, or in Mexico before European immigration? When did Latin stop being spoken, and when did French start being spoken? How did English become such an important world language? These and other similar questions are often asked by the interested layman. One aim of this volume – taking the Introduction and the individual chapters together – is to provide answers to these and related questions, or in certain cases to show why the questions cannot be answered as they stand. The chapters concentrate on an individual language or group of languages, and in this Introduction I want rather to present a linking essay which will provide a background against which the individual chapters can be appreciated.

After discussing some preliminary notions in Section 1, Section 2 of the Introduction provides a rapid survey of the languages spoken in the world today, concentrating on those not treated in the subsequent chapters, so that the reader can gain an overall impression of the extent of linguistic diversity that characterises the world in which we live. Since the notion of 'major language' is primarily a social notion – languages become major (such as English) or stop being major (such as Sumerian) not because of their grammatical structure, but because of social factors – Section 3 discusses some important sociolinguistic notions, in particular concerning the social interaction of languages.

1.1 How Many Languages?

Linguists are typically very hesitant to answer the first question posed above, namely: how many languages are spoken in the world today? Probably the best that one can say, with some hope of not being contradicted, is that at a very conservative estimate some 6,000 languages are spoken today. Laymen are often surprised that the figure should be so high, but I would emphasise that this is a conservative estimate. But why is it that linguists are not able to give a more accurate figure? There are several different reasons conspiring to prevent them from doing so, and these will be outlined below.

Even a couple of decades ago one could reasonably have observed that some parts of the world were simply insufficiently studied from a linguistic viewpoint, so that we did not know precisely what languages are spoken there. There are very few parts of the world where this still holds, since our knowledge of the linguistic situation in remote parts of the world has improved dramatically in recent years – New Guinea, for instance, has changed from being almost a blank linguistic map in the first half of the twentieth century to the stage where nearly all languages can be pinpointed with accuracy: since perhaps as many as one-fifth of the world's languages are spoken in New Guinea, this has radically changed any estimate of the total number of languages. Although new languages are still being discovered, this is no longer the major factor it would have been in the past.

A second and more important problem is that it is difficult or impossible in many cases to decide whether two related speech varieties should be considered different languages or merely different dialects of the same language. Native speakers of English are often surprised that there should be problems in delimiting languages from dialects, since present-day dialects of English are in general mutually intelligible (at least with some familiarisation), and even the language most closely related genetically to English, Frisian, is mutually unintelligible with English. With the languages of Europe more generally, there are in general established traditions of whether two speech varieties should be considered different languages or merely dialect variants, but these decisions have often been made more on political and social rather than on strictly linguistic grounds.

One criterion that is often advanced as a purely linguistic criterion is mutual intelligibility: if two speech varieties are mutually intelligible, they are different dialects of the same language, but if they are mutually unintelligible, they are different languages. But if applied to the languages of Europe, this criterion would radically alter our assessment of what the different languages of Europe are: the most northern dialects and the most southern dialects (in the traditional sense) of German are mutually unintelligible, while dialects of German spoken close to the Dutch border are mutually intelligible with dialects of Dutch spoken just across the border. In fact, our criterion for whether a dialect is Dutch or German relates in large measure to social factors – is the dialect spoken in an area where Dutch is the standard language or where German is the standard language? By the same criterion, the three nuclear Scandinavian languages (in the traditional sense), Danish, Norwegian and Swedish, would turn out to be dialects of one language, given their mutual intelligibility. While this criterion is often applied to non-European languages (so that nowadays linguists talk of the Chinese languages rather than the Chinese dialects, given the mutual unintelligibility of, for instance, Mandarin and Cantonese), it seems unfair that it should not be applied consistently to European languages as well.

In some cases, the intelligibility criterion actually leads to contradictory results, namely when we have a dialect chain, i.e. a string of dialects such that adjacent dialects are readily mutually intelligible, but dialects from the far ends of the chain are not mutually intelligible. A good illustration of this is the Dutch–German dialect complex. One could start from the far south of the German-speaking area and move to the far west of the Dutch-speaking area without encountering any sharp boundary across which mutual intelligibility is broken; but the two end points of this chain are speech varieties so different from one another that there is no mutual intelligibility possible. If one takes a simplified dialect chain A – B – C, where A and B are mutually intelligible,

as are B and C, but A and C are mutually unintelligible, then one arrives at the contradictory result that A and B are dialects of the same language, B and C are dialects of the same language, but A and C are different languages. There is in fact no way of resolving this contradiction if we maintain the traditional strict difference between language and dialects, and what such examples show is that this is not an all-or-nothing distinction, but rather a continuum. In this sense, it is not just difficult, but in principle impossible to answer the question how many languages are spoken in the world.

A further problem with the mutual intelligibility criterion is that mutual intelligibility itself is a matter of degree rather than a clear-cut opposition between intelligibility and unintelligibility. If mutual intelligibility were to mean 100 per cent mutual intelligibility of all utterances, then perhaps no two speech varieties would be classified as mere dialect variants; for instance, although speakers of British and American English can understand most of one another's speech, there are areas where intelligibility is likely to be minimum unless one speaker happens to have learned the linguistic forms used by the other, as with car (or auto) terms like British *boot*, *bonnet*, *mudguard* and their American equivalents *trunk*, *hood*, *fender*. Conversely, although speakers of different Slavonic languages are often unable to make full sense of a text in another Slavonic language, they can usually make good sense of parts of the text, because of the high percentage of shared vocabulary and forms.

Two further factors enter into the degree of mutual intelligibility between two speech varieties. One is that intelligibility can rise rapidly with increased familiarisation: when American films were first introduced into Britain they were initially considered difficult to understand, but increased exposure to American English has virtually removed this problem. Speakers of different dialects of Arabic often experience difficulty in understanding each other at first meeting, but soon adjust to the major differences between their respective dialects, and Egyptian Arabic, as the most widely diffused modern Arabic dialect, has rapidly gained in intelligibility throughout the Arab world. This can lead to 'one-way intelligibility', as when speakers of, say, Tunisian Arabic are more likely to understand Egyptian Arabic than vice versa, because Tunisian Arabic speakers are more often exposed to Egyptian Arabic than vice versa. The second factor is that intelligibility is to a certain extent a social and psychological phenomenon: it is easier to understand when you want to understand. A good example of this is the conflicting assessments different speakers of the same Slavonic language will often give about the intelligibility of some other Slavonic language, correlating in large measure with whether or not they feel well disposed to speakers of the other language.

The same problems as exist in delimiting dialects from languages arise, incidentally, on the historical plane too, where the question arises: at what point has a language changed sufficiently to be considered a different language? Again, traditional answers are often contradictory: Latin is considered a dead language, although its descendants, the Romance languages, live on, so at some time Latin must have changed sufficiently to be deemed no longer the same language, but a qualitatively different one. On the other hand, Greek is referred to as 'Greek' throughout its attested history (which is longer than that of Latin and the Romance languages combined), with merely the addition of different adjectives to identify different stages of its development (e.g. Ancient Greek, Byzantine Greek, Modern Greek). In the case of the history of the English language, there is even conflicting terminology: the oldest attested stages of English can be referred to either as Old English (which suggests an earlier stage of Modern English) or as Anglo-Saxon (which suggests a different language that is the

ancestor of English, perhaps justifiably so given the mutual unintelligibility of Old and Modern English).

A further reason why it is difficult to assess the number of languages spoken in the world today is that many languages are on the verge of extinction. While it has been the case throughout human history that languages have died out, recent social changes have considerably accelerated this process, as the languages of smaller speech communities are endangered by those with more speakers and more prestige. Economic factors often make it difficult for modern services (health, education, etc.) to be provided in the languages of smaller speech communities, and members of smaller speech communities interested in integrating into wider social networks often feel impelled to abandon their own language in favour of a language of wider currency or at least to encourage their children to do so. This is not just the ever increasing use of English as a result of globalisation. At a more local level, smaller languages are endangered in particular situations by French, Spanish, Indonesian, Swahili – even Tsamai (the last-mentioned a language of Ethiopia with around 10,000 speakers, to which the few remaining speakers of Birale are assimilating). Documentation and, where possible, preservation of endangered languages is one of the major tasks facing linguists in the twenty-first century.

1.2 Language Families and Genetic Classification

One of the basic organisational principles of this volume, both in Section 2 of the Introduction and in the arrangement of the individual chapters, is the classification of languages into language families. It is therefore important that some insight should be provided into what it means to say that two languages belong to the same language family (or equivalently: are genetically related).

It is probably intuitively clear to anyone who knows a few languages that some languages are closer to one another than are others. For instance, English and German are closer to one another than either is to Russian, while Russian and Polish are closer to one another than either is to English. This notion of similarity can be made more precise, as is done for instance in the chapter on the Indo-European languages, but for the moment the relatively informal notion will suffice. Starting in the late eighteenth century, a specific hypothesis was proposed to account for such similarities, a hypothesis which still forms the foundation of research into the history and relatedness of languages. This hypothesis is that where languages share some set of features in common, these features are to be attributed to their common ancestor. Let us take some examples from English and German.

In English and German we find a number of basic vocabulary items that have the same or almost the same form, e.g. English *man* and German *Mann*. Likewise, we find a number of bound morphemes (prefixes and suffixes) that have the same or almost the same form, such as the genitive suffix, as in English *man's* and German *Mann(e)s*. Although English and German are now clearly different languages, we may hypothesise that at an earlier period in history they had a common ancestor, in which the word for 'man' was something like *man* and the genitive suffix was something like *-s*. Thus English and German belong to the same language family, which is the same as saying that they share a common ancestor. We can readily add other languages to this family, since a word like *man* and a genitive suffix like *-s* are also found in Dutch, Frisian and the Scandinavian languages. The family to which these languages belong has been given

that the proto-language is not an attested language – although if written records had gone back far enough, we might well have had attestations of this language – but its postulation is the most plausible hypothesis explaining the remarkable similarities among the various Germanic languages.

Although not so obvious, similarities can be found among the Germanic languages and a number of other languages spoken in Europe and spreading across northern India as far as Bangladesh. These other languages share fewer similarities with the Germanic languages than individual Germanic languages do with one another, so that they are more remotely related. The overall language family to which all these languages belong is the Indo-European family, with its reconstructed ancestor language Proto-Indo-European. As is discussed in more detail in the chapter on Indo-European languages, the Indo-European family contains a number of branches (i.e. smaller language families, or subfamilies), such as Slavonic (including Russian and Polish), Iranian (including Persian and Pashto), and Celtic (including Irish and Welsh). The overall structure is therefore hierarchical: The most distant ancestor is Proto-Indo-European. At an intermediate point in the family tree, and therefore at a later period of history, we have such languages as Proto-Germanic and Proto-Celtic, which are descendants of Proto-Indo-European but ancestors of languages spoken today. Still later in history, we find the individual languages as they are spoken today or attested in recent history, such as English and German as descendants of Proto-Germanic and Irish and Welsh as descendants of Proto-Celtic. One typical property of language change that is represented accurately by this family-tree model is that, as time goes by, languages descending from a common ancestor tend to become less and less similar. For instance, Old English and Old High German (the ancestor of Modern German) were much closer to one another than are the modern languages – they may even have been mutually intelligible, at least to a large extent.

Although the family-tree model of language relatedness is an important foundation of all current work in historical and comparative linguistics, it is not without its problems, both in practice and in principle. Some of these will now be discussed.

We noted above that with the passage of time, genetically related languages will grow less and less similar. This follows from the fact that, once two languages have split off as separate languages from a common ancestor, each will innovate its own changes, different from changes that take place in the other language, so that the cumulative effect will be increasing divergence. With the passage of enough time, the divergence may come to be so great that it is no longer possible to tell, other than by directly examining the history, that the two languages do in fact come from a common ancestor. The best-established language families, such as Indo-European or Sino-Tibetan, are those where the passage of time has not been long enough to erase the obvious traces of genetic relatedness. (For language families that have a long written tradition, one can of course make use of earlier stages of the language, which provide more evidence of genetic relatedness.) In addition, there are many hypothesised language families for which the evidence is not sufficient to convince all, or even the majority, of scholars. For instance, the Turkic language family is a well-established language family, as is each of the Mongolic and Tungusic families. What is controversial, however, is whether or not these individual families are related as members of an even larger family. The possibility of an Altaic family, comprising Turkic, Mongolic, and Tungusic, is rather widely accepted, and some scholars would advocate increasing the size of this family by adding Korean and perhaps Japanese.

The attitudes of different linguists to problems of this kind have been characterised as an opposition between 'splitters' (who require the firmest evidence before they are prepared to acknowledge genetic relatedness) and 'clumpers' (who are ready to assign languages to the same family on the basis of quite restricted similarities). I should, incidentally, declare my own splitter bias, lest any of my own views that creep in be interpreted as generally accepted dogma. The most extreme clumper position would, of course, be to maintain that all languages of the world are genetically related, although there are less radical positions that would posit such 'macro-families' as Eurasiatic or Nostratic (including, inter alia, Indo-European, Uralic and Altaic), Dene-Caucasian (including, inter alia, Na-Dene, Sino-Tibetan, East Caucasian and West Caucasian), and Austric (including at least Austronesian and Austro-Asiatic). In the survey of the distribution of languages of the world in Section 2, I have basically retained my own splitter position, although for areas of great linguistic diversity and great controversy surrounding genetic relations (such as New Guinea and the Americas) I have simply refrained from detailed discussion.

While no linguist would doubt that some similarities among languages are due to genetic relatedness, there are several other possibilities for the explanation of any particular similarity, and before assuming genetic relatedness one must be able to exclude, at least with some degree of plausibility, these other possibilities. Unfortunately, in a great many cases it is not possible to reach a firm and convincing decision. Let us now examine some of the explanations other than genetic relatedness.

First, two languages may happen purely by chance to share some feature in common. For instance, the word for *dog* in Mbabaram, an Australian Aboriginal language, happens to be *dog*. This Mbabaram word is not, incidentally, a borrowing from English, but is the regular development in Mbabaram of an ancestral form something like **gudaga*, which is found in forms similar to this reconstruction in other related languages (it is usual to prefix reconstructed forms with an asterisk). If anyone were tempted to assume on this basis, however, that English and Mbabaram are genetically related, examination of the rest of Mbabaram vocabulary and grammar would soon quash the genetic relatedness hypothesis, since there is otherwise minimal similarity between the two languages. In comparing English and German, by contrast, there are many similarities at all levels of linguistic analysis. Even sticking to vocabulary, the correspondence *man*: *Mann* can be matched by *wife*: *Weib*, *father*: *Vater*, *mother*: *Mutter*, *son*: *Sohn*, *daughter*: *Tochter*, etc. Given that other languages have radically different words for these concepts (e.g. Japanese *titi* 'father', *haha* 'mother', *musuko* 'son', *musume* 'daughter'), it can clearly not be merely the result of chance that English and German have so many similar items. But if the number of similar items in two languages is small, it may be difficult or impossible to distinguish between chance similarity and distant genetic relatedness.

Certain features shared by two languages might turn out to be manifestations of language universals, i.e. of features that are common to all languages or are inherently likely to occur in any language. Most discussions of language universals require a fair amount of theoretical linguistic background, but for present purposes I will take a simple, if not particular profound, example. In many languages across the world, the syllable *ma* or its reduplicated form *mama* or some other similar form is the word for 'mother'. The initial syllable *ma* enters into the Proto-Indo-European word for 'mother' that has given English *mother*, Spanish *madre*, Russian *mat′*, Sanskrit *mātā*. In Mandarin Chinese, the equivalent word is *mā*. Once again, examination of other features of Indo-European

languages and Chinese would soon dispel any possibility of assigning Chinese to the Indo-European language family. Presumably the frequency across languages of the syllable *ma* in the word for 'mother' simply reflects the fact that this is typically one of the first syllables that babies articulate clearly, and is therefore interpreted by adults as the word for mother. (In the South Caucasian language Georgian, incidentally, *mama* means 'father' – and 'mother' is *deda* – so that there are other ways of interpreting baby's first utterance.)

Somewhat similar to universals are patterns whereby certain linguistic features frequently co-occur in the same language, i.e. where the presence of one feature seems to require or at least to foster the presence of some other feature. For instance, the study of word universals by Greenberg (1966) showed that if a language has verb-final word order (i.e. if 'the man saw the woman' is expressed literally as 'the man the woman saw'), then it is highly probable that it will also have postpositions rather than prepositions (i.e. 'in the house' will be expressed as 'the house in') and that it will have genitives before the noun (i.e. the pattern 'cat's house' rather than 'house of cat'). Thus, if we find two languages that happen to share the features: verb-final word order, postpositions, prenominal genitives, then the co-occurrence of these features is not evidence for genetic relatedness. Many earlier attempts at establishing wide-ranging genetic relationships suffer precisely from failure to take this property of typological patterns into account. Thus the fact that Turkic languages, Mongolic languages, Tungusic languages, Korean and Japanese share all of these features is not evidence for their genetic relatedness (although there may, of course, be other similarities, not connected with recurrent typological patterns, that do establish genetic relatedness). If one were to accept just these features as evidence for an Altaic language family, then the family would have to be extended to include a variety of other languages with the same word order properties, such as the Dravidian languages of southern India and Quechua, spoken in South America.

Finally, two languages might share some feature in common because one of them has borrowed it from the other (or because they have both borrowed it from some third language). English, for instance, borrowed a huge number of words from French during the Middle Ages, to such an extent that an uncritical examination of English vocabulary might well lead to the conclusion that English is a Romance language, rather than a Germanic language. The term 'borrow', as used here, is the accepted linguistic term, although the terminology is rather strange, since 'borrow' suggests a relatively superficial acquisition, one which is moreover temporary. Linguistic borrowings may run quite deep, and there is of course no implication that they will ever be repaid. Among English loans from French, for instance, there are many basic vocabulary items, such as *very* (replacing the native Germanic *sore*, as in the biblical *sore afraid*). Examples from other languages show even more deep-seated loans: the Semitic language Amharic, for instance, has lost the typical Semitic word order patterns, in which the verb precedes its object and adjectives and genitives follow their noun, in favour of the order where the verb follows its object and adjectives and genitives precede their noun; Amharic is in close contact with Cushitic languages, and Cushitic languages typically have the order object–verb, adjective/genitive–noun, so that Amharic has in fact borrowed these word orders from neighbouring Cushitic languages.

It seems that whenever two languages come into close contact, they will borrow features from one another. In some cases the contact can be so intense among the languages in a given area that they come to share a significant number of common features, setting

this area off from adjacent languages, even languages that may happen to be more closely related genetically to languages within the area. The languages in an area of this kind are often said to belong to a sprachbund (German for 'language league'), and perhaps the most famous example of a sprachbund is the Balkan sprachbund, whose members (Modern Greek, Albanian, Bulgarian, Macedonian, Rumanian) share a number of striking features not shared by closely related languages like Ancient Greek, other Slavonic languages (Bulgarian is Slavonic), or other Romance languages (Rumanian is Romance). The most striking of these features is loss of the infinitive, so that instead of 'give me to drink' one says 'give me that I-drink' (Modern Greek *ðos mu na pjo*, Albanian *a-më të pi*, Bulgarian *daj mi da pija*, Rumanian *dă-mi să beau*; in all four languages the subject of the subordinate clause is encoded in the ending of the verb).

Since we happen to know a lot about the history of the Balkan languages, linguists were not deceived by these similarities into assigning a closer genetic relatedness to the Balkan languages than in fact holds (all are ultimately members of the Indo-European family, though from different branches). In other parts of the world, however, there is the danger of mistaking areal phenomena for evidence of genetic relatedness. In South-East Asia, for instance, many languages share very similar phonological and morphological patterns: in Chinese, Thai and Vietnamese words are typically monosyllabic, there is effectively no morphology (i.e. words do not change after the manner of English *dog*, *dogs* or *love*, *loves*, *loved*), syllable structure is very simple (only a few single consonants are permitted word-finally, while syllable-initially consonant clusters are either disallowed or highly restricted), and there is phonemic tone (thus Mandarin Chinese *mā*, with a high level tone, means 'mother', while *mǎ* with a falling–rising tone, means 'horse'), and moreover there are a number of shared lexical items. For these reasons, it was for a long time believed that Thai and Vietnamese were related genetically to Chinese. More recently, however, it has been established that these similarities are not the result of common ancestry, and Thai and Vietnamese are now generally acknowledged not to be genetically related to Chinese. The similarities are the results of areal contact. The shared vocabulary items are primarily the result of intensive Chinese cultural influence, especially on Vietnamese. The tones and simple syllable structures can often be shown to be the result of relatively recent developments, and indeed in one language that is genetically related to Chinese, namely Classical Tibetan, one finds complex consonant clusters but no phonemic tone, i.e. the similarities noted above are neither necessary nor sufficient conditions for genetic relatedness.

In practice, the most difficult task in establishing genetic relatedness is to distinguish between genuine cognates (i.e. forms going back to a common ancestor) and those that are the result of borrowing. It would therefore be helpful if one could distinguish between those features of a language that are borrowable and those that are not. Unfortunately, it seems that there is no feature that can absolutely be excluded from borrowing. Basic vocabulary can be borrowed, so that for instance Japanese has borrowed the whole set of numerals from Chinese, and even English borrowed its current set of third person plural pronouns (*they*, *them*, *their*) from Scandinavian. Bound morphemes can be borrowed: a good example is the agent suffix *-er* in English, with close cognates in the other Germanic languages; this is ultimately a loan from the Latin agentive suffix *-arius*, which has however become so entrenched in English that it is a productive morphological device applicable in principle to any verb to derive a corresponding agentive noun.

At one period in the recent history of comparative linguistics, it was believed that a certain basic vocabulary list could be isolated, constant across languages and cultures,

such that the words on this list would be replaced at a constant rate. Thus, if one assumes that the retention rate is around 86 per cent per millennium, this means that if a single language splits into two descendant languages, then after 1,000 years each language would retain about 86 per cent of the words in the list from the ancestor language, i.e. the two descendants would then share just over 70 per cent of the words in the list with each other. In some parts of the world, groupings based on this 'glottochronological' method still form the basis of the only available detailed and comprehensive attempt at establishing genetic relations. It must be emphasised that the number of clear counter-examples to the glottochronological method, i.e. instances where independent evidence contradicts the predictions of this approach, is so great that no reliance can be placed on its results.

It is, however, true that there are significant differences in the ease with which different features of a language can be borrowed. The thing that seems most easily borrowable is cultural vocabulary, and indeed it is quite normal for a community borrowing some concept (or artefact) from another community to borrow the foreign name along with the object. Another set of features that seem rather easily borrowable are general typological features, such as word order: in addition to the Amharic example cited above, one might note the fact that many Austronesian languages spoken in New Guinea have adopted the word order where the object is placed before the verb, whereas almost all other Austronesian languages place the object after the verb; this happened under the influence of Papuan languages, almost all of which are verb-final. Basic vocabulary and bound morphology are hardest to borrow. But even though it is difficult to borrow bound morphology, it is not impossible, so in arguments over genetic relatedness one cannot exclude a priori the possibility that even affixes may have been borrowed.

2 Distribution of the World's Languages

In this section, I wish to give a general survey of the distribution of the languages of the world, in terms of their genetic affiliation. I will therefore be talking primarily about the distribution of language families, although reference will be made to individual languages where appropriate. The discussion will concentrate on languages and language families not covered in individual chapters, and at appropriate places I have digressed to give a brief discussion of some interesting structural or sociological point in the language being treated.

2.1 Europe

Europe, taken here in the traditional cultural sense rather than in the current geographical sense of 'the land mass west of the Urals', is the almost exclusive preserve of the Indo-European family. This family covers not only almost the whole of Europe, but also extends through Armenia (in the Caucasus), Iran and Afghanistan into Central Asia (Tajikistan), with the easternmost outpost of this strand the Iranian language Sarikoli, spoken just inside China. Another strand spreads from Afghanistan across Pakistan, northern India and southern Nepal, to end with Bengali in eastern India and Bangladesh; an off-shoot from northern India, Sinhalese, is spoken in Sri Lanka, and the language of the Maldives is the closely related Maldivian.

In addition, the great population shifts that resulted from the voyages of exploration starting at the end of the fifteenth century have carried Indo-European languages to

many distant lands. The dominant languages of the Americas are now Indo-European (English, Spanish, Portuguese, French), as is the dominant language of Australia and New Zealand (English). While in some countries these languages are spoken by populations descended primarily from European settlers, there are also instances where a variety of the European language is spoken by a population of a different origin, perhaps the best-known example being the creolised forms of European languages (especially English, French and Portuguese) spoken by the descendants of African slaves in the Caribbean. It should be noted that these population shifts have not led exclusively to the spread of European languages, since many languages of India, both Indo-European and Dravidian, have also extended as a by-product, being spoken now by communities in the Caribbean area, in East Africa, and in the South Pacific (especially Fiji).

Of the few European languages not belonging to the Indo-European family, mention may first be made of Basque, a language isolate, with no established genetic relations to any other language. It is spoken in the Pyrenees near the French–Spanish border. Basque is perhaps most noted for its ergative construction, whereby instead of having a single case (nominative) for both subjects of intransitive verbs and subjects (agents) of transitive verbs, with a different case (accusative) for objects (patients) of transitive verbs, Basque uses one case (absolutive) for both intransitive subjects and objects of transitive verbs, and a different case (ergative) for subjects of transitive verbs, as in the following sentences:

Jon etorri da.	'John came.'
Neska-k gizona ikusi du.	'The girl saw the man.'

In the first sentence, *Jon* is intransitive subject, and stands in the absolutive (no inflection); in the second sentence, *neska-k* 'girl' is transitive subject, and therefore stands in the ergative (suffix *-k*), while *gizona* 'man' is transitive object, and therefore stands in the absolutive, with no inflection.

Some other languages of Europe belong to the Uralic family. These include Hungarian, Finnish, Estonian, and Saami (formerly also called Lappish – properly a group of mutually unintelligible languages rather than a single language), to which can be added a number of smaller languages closely related to Finnish and Estonian. Other members of the Uralic family are spoken on the Volga and in northern Eurasia on both sides of the Urals, stretching as far as southern Siberia.

Turkish as spoken in the Balkans represents the Turkic family in Europe, but this is primarily an Asian family, and will be treated in the next section. The same is true of Afroasiatic, represented in Europe by Maltese.

2.2 Asia

Having just mentioned Turkish, we may now turn to the Turkic family, which is spoken in Turkey, parts of the Caucasus, some areas on the Volga, most of Central Asia (and stretching down into north-western Iran), and large parts of southern Siberia, with one off-shoot, Yakut, in north-eastern Siberia. Turkic is perhaps to be joined in a single language family (Altaic) with the Mongolic and Tungusic families. The Mongolic languages are spoken predominantly in Mongolia and northern China, though there are also isolated Mongolic languages in Afghanistan (Moghol) and just to the north of the Caucasus mountains (Kalmyk); the main member of the family is the language Mongolian

(sometimes called Khalkha, after its principal dialect), which is the official language of Mongolia. The Tungusic languages are spoken by numerically small population groups in Siberia, spreading over into Mongolia and especially north-eastern China. The Tungusic language best known to history is Manchu, the native language of the Qing dynasty that ruled China from 1644 to 1911; the Manchu language is, however, now almost extinct, having been replaced by Chinese. Whether Korean or Japanese can be assigned to the Altaic family is a question of current debate.

This is a convenient point at which to discuss a number of other languages spoken in northern Asia. All are the languages of small communities (a few hundred or a few thousand). They are sometimes referred to collectively as Paleosiberian (or Paleoasiatic), although this is not a genetic grouping. Three of them are language isolates: Ket, spoken on the Yenisey river, and the sole survivor of the small Yeniseic family; Yukaghir, spoken on the Kolyma river; and Nivkh (Gilyak), spoken at the mouth of the Amur river and on Sakhalin island. The small Chukotko-Kamchatkan family comprises the indigenous languages of the Chukotka and Kamchatka peninsulas: Chukchi, Koryak, Kamchadal (Itelmen); it has been suggested that they may be related to Eskimo-Aleut, which we treat in Section 2.5 on the Americas. Finally, we may mention here the recently extinct Ainu, apparently a language isolate, whose last native speakers lived in Hokkaido, the most northerly Japanese island.

One of the geographic links between Europe and Asia, the Caucasus, has since antiquity been noted for the large number of clearly distinct languages spoken there; indeed it was referred to by the Arabs as the 'mountain of tongues'. Some of the languages spoken in the Caucasus belong to other families (e.g. Armenian and Ossetian to Indo-European, Azerbaijani to Turkic), but there are in addition a number of languages with no known affiliations to languages outside the Caucasus: these are the Caucasian languages. Even the internal genetic relations of the Caucasian languages are the subject of debate. Few scholars now accept the genetic relatedness of all Caucasian languages, but there is ongoing debate over whether West Caucasian and East Caucasian together form a single North Caucasian family. The Kartvelian or South Caucasian family includes Georgian, the Caucasian language with the largest number of speakers (over four million) and the only Caucasian language to have a long-standing literary tradition (dating back to the fifth century). The West (North-West) Caucasian languages are on and close to the Black Sea coast, though also in Turkey as a result of emigration since the mid-nineteenth century; one Caucasian language, Ubykh, which died out in Turkey towards the end of the twentieth century, is noteworthy for the large number of its consonant phonemes – at one time it was considered the world record-holder. The East (North-East) Caucasian (or Nakh-Daghestanian) languages are spoken mainly in Daghestan, Chechnya and Ingushetia in the Russian Federation; the best-known language is Chechen with about a million speakers, though the family also includes languages like Hinuq, spoken by about 500 people in a single village.

Turning now to south-western Asia, we may consider the Afroasiatic family, which, as its name suggests, is spoken in both Asia and Africa. In Asia its main focus is the Arab countries of the Middle East, although Hebrew and Aramaic are also Afroasiatic languages of Asia, belonging to the Semitic branch of Afroasiatic. In addition Arabic is, of course, the dominant language of North Africa, where Afroasiatic is represented not only by a number of other Semitic languages (those of Ethiopia, the major one being Amharic), but also by Berber, the Cushitic languages of the Horn of Africa (including Somali, the official language of Somalia), and the Chadic languages of

northern Nigeria and adjacent areas (including Hausa). One branch of Afroasiatic formerly spoken in Africa, Egyptian (by which is meant the language of ancient Egypt, not the dialect of Arabic currently spoken in Egypt), is now extinct.

In South Asia (the traditional 'Indian subcontinent'), four language families meet. Indo-European languages, more specifically languages of the Indo-Aryan branch of Indo-European, dominate in the north, while the south is the domain of the Dravidian languages (although some Dravidian languages are spoken further north, in particular Brahui, spoken in Pakistan). The northern fringe of the subcontinent is occupied by Sino-Tibetan languages, to which we return below. The fourth family is Austro-Asiatic. The languages in this family with most speakers are actually spoken in South-East Asia: Vietnamese in Vietnam and Khmer (Cambodian) in Cambodia, and they are the only languages of the family to have the status of national languages. Languages of the family are scattered from central India eastwards into Vietnam, Western Malaysia and the Nicobar Islands. In India itself, the Austro-Asiatic language with most speakers is Santali. It is only relatively recently that the assignment of Vietnamese to this family has gained widespread acceptance. In addition, there is one language isolate, Burushaski, spoken in northern Pakistan, while the genetic affiliations of the languages of the Andaman islands remain unclear.

We have already introduced a number of South-East Asian languages, and may now turn to the other two families represented in this area: Tai-Kadai (also called Kadai and Kam-Tai) and Sino-Tibetan. While the Tai-Kadai group of languages, which includes Thai (Siamese) and Lao, was earlier often considered a branch of Sino-Tibetan, this view has now been abandoned; Tai-Kadai languages are spoken in Thailand, Laos, southern China, and also in parts of Burma (Myanmar) and Vietnam. Sino-Tibetan contains the language with the largest number of native speakers in the world today, Chinese (and this remains true even if one divides Chinese into several different languages, in which case Mandarin occupies first position). The other Sino-Tibetan languages traditionally form the Tibeto-Burman branch, which includes Tibetan and Burmese, in addition to a vast number of languages spoken predominantly in southern China, Burma (Myanmar), northern India and Nepal. Finally, the languages of the Hmong-Mien (Miao-Yao) family are spoken in southern China and adjacent areas.

In East Asia we find Korean and Japanese (the latter together with the closely related Ryukyuan varieties), whose genetic affiliations to each other or to other languages (such as Altaic) remain the subject of at times heated debate.

The Austronesian family, though including some languages spoken on the Asian mainland, such as Malay of Western Malaysia and Cham spoken in Cambodia and Vietnam, is predominantly a language of the islands stretching eastwards from the South-East Asian mainland: even Malay–Indonesian has more speakers in insular South-East Asia than on the Malay peninsula. Austronesian languages are dominant on most of the islands from Sumatra in the west to Easter Island in the east, including the Philippines, but excluding New Guinea (where Austronesian languages are, however, spoken in many coastal areas); Malagasy, the language of Madagascar, is a western outlier of the family; Austronesian languages are also indigenous to Taiwan, though now very much in the minority relative to Chinese.

2.3 New Guinea and Australia

The island of New Guinea, which can be taken linguistically together with some of the smaller surrounding islands, is the most differentiated area linguistically in the whole

world. Papua New Guinea, which occupies the eastern half of the island, contains some 800 languages for a total population of about five and a half million, meaning that the average language has just under 7,000 speakers. In many of the coastal areas of New Guinea, Austronesian languages are spoken, but the other languages are radically different from these Austronesian languages. These other languages are referred to collectively as either 'non-Austronesian languages of New Guinea' or as 'Papuan languages', though it should be realised that this is a negatively characterised term, rather than a claim about genetic relatedness. Though much work remains to be done, there has been considerable recent progress in classifying the Papuan languages genetically; in particular, there is growing evidence for a Trans-New Guinea family containing a large number of languages running east–west across the middle of the main island and on to some of the smaller islands to the west.

One syntactic property that is widespread among the Highland Papuan languages is worthy of note, namely switch reference. In a language with a canonical switch reference system, a sentence may (and typically does) consist of several clauses, of which only one is an independent clause (i.e. could occur on its own as a free-standing sentence), all the others being dependent; each dependent clause is marked according to whether or not its subject is the same as or different from the subject of the clause on which it is dependent. The examples below are from Usan:

Ye nam su-ab, isomei. 'I cut the tree and went down.'
Ye nam su-ine, isorei. 'I cut the tree and it fell down.'

The independent verbs, *isomei* and *isorei*, are respectively first person singular and third person singular. The dependent verbs, *su-ab* and *su-ine*, have respectively the suffix for same subject and the suffix for different subject. In the first example, therefore, the subjects of the two clauses are the same (i.e. I cut the tree and I went/fell down), while in the second sentence they are different (i.e. I cut the tree and some other entity – from the context only the tree is available – went/fell down). The words *ye* and *nam* mean respectively 'I' and 'tree'. One effect of switch reference is that the speaker of a language with switch reference must plan a discourse ahead to a much greater extent than is required by languages lacking switch reference, since in switch reference languages it is nearly always the case that the dependent clause precedes the independent clause, i.e. in clause n one has to mark the co-reference relation that holds between the subject of clause n and the subject of clause $n + 1$. This should, incidentally, serve to dispel any lingering notions concerning the primitiveness or lack of grammar in the languages of other societies. Although switch reference is found in many other parts of the world (e.g. in many indigenous languages of the Americas), it is particularly characteristic of the languages of the New Guinea Highlands.

Although the genetic classification of the indigenous languages of Australia, which numbered over 250 at the time of contact with Europeans, is the subject of at times acrimonious debate, there is a general consensus that a large Pama-Nyungan family can be identified, comprising most languages spoken in the south and centre and some in the north, while the other languages of the north form a number of small families and language isolates.

The Australian languages overall are characterised by an unusual consonant system, from the viewpoint of the kinds of consonant systems that are found most frequently across the languages of the world. Most Australian languages have no fricatives, and no

voice opposition among their stops. However, they distinguish a large number of places of articulation, especially in terms of lingual articulations: thus most languages have, in addition to labial and velar stops, all of palatal, alveolar, and retroflex stops, while many languages add a further series of phonemically distinct dentals. The same number of distinctions is usually found with the nasals, and some languages extend this number of contrasts in the lingual stops to the laterals as well. One result of this is that Europeans usually fail to perceive (or produce, should they try to do so) phonemic oppositions that are crucial in Aboriginal languages, while conversely speakers of Australian languages fail to perceive or produce phonemic oppositions that are crucial in English (such as the distinction among *pit*, *bit*, *bid*).

One Australian language, Dyirbal, spoken in the Cairns rainforest in northern Queensland, has played an important role in recent discussions of general linguistic typology, and it will be useful to make a short digression to look at the relevant unique, or at least unusual, features of Dyirbal – though it should be emphasised that these features are not particularly typical of Australian languages overall.

In English, one of the pieces of evidence for saying that intransitive and transitive subjects are just subtypes of the overall notion 'subject' is that they behave alike with respect to a number of different syntactic processes. For instance, a rule of English syntax allows one to omit the subject of the second conjunct of a coordinate sentence if it is co-referential with the subject of the first conjunct, i.e. one can abbreviate the first sentence below to the second one:

I hit you and I came here.
I hit you and came here.

It is not possible to carry out a similar abbreviation of the next sentence below, since its subjects are not co-referential, even though the object of the first conjunct is co-referential with the subject of the second conjunct:

I hit you and you came here.

In the above examples, the first clause is transitive and the second clause intransitive, but the notion of subject applies equally to both clauses. If we think not so much of grammatical labels like subject and object, but rather of semantic labels like agent and patient, then we can say that in English it is the agent of a transitive clause that behaves as subject. In the corresponding Dyirbal sentences, however, it is the patient that behaves as subject, as can be seen in the following sentences:

Ngaja nginuna balgan, ngaja baninyu.	'I hit you and I came here.'
Ngaja nginuna balgan, nginda baninyu.	'I hit you and you came here.'
Ngaja nginuna balgan, baninyu.	'I hit you and you came here.'

In these sentences, *ngaja* is the nominative form for 'I', while *nginuna* is the accusative form for 'you'; the verbs are *balgan* 'hit' (transitive) and *baninyu* 'come here' (intransitive). In the third sentence, where the intransitive subject is omitted, it must be interpreted as co-referential with the patient, not the agent, of the first clause. In Section 2.1 we mentioned ergativity in connection with Basque case marking. These Dyirbal examples show that Dyirbal has ergativity in its syntactic system: patients of transitive verbs,

rather than agents of transitive verbs, are treated as subjects, i.e. are treated in the same way as intransitive subjects. Note that in this sense Dyirbal grammar is certainly different from English grammar, but it is no less well defined.

Another unusual feature of Dyirbal is sociolinguistic. In many, if not all languages there are different choices of lexical item depending on differences in social situation, such as the difference between English *father* and *dad(dy)*. What is unusual about Dyirbal is that a difference of this kind exists for every single lexical item in the language. Under certain circumstances, in particular in the presence of a taboo relative (e.g. a parent-in-law), every lexical item of ordinary language (Guwal) must be replaced by the corresponding lexical item from avoidance style (Jalnguy). No doubt in part for functional reasons, to ease the memory load, it is usual for several semantically related words of Guwal to correspond to a single Jalnguy word, as when the various Guwal names for different species of lizard are all subsumed by the one Jalnguy word *jijan*.

The surviving textual materials in the Tasmanian languages, extinct since the end of the nineteenth century, are insufficient in scope or reliability to allow any accurate assessment of the genetic affiliations of these languages – certainly none is immediately apparent.

2.4 Africa

Africa north of the Sahara is the preserve of Afroasiatic languages, which have already been treated in Section 2.2. This section will therefore concentrate on the sub-Saharan languages, though excluding languages introduced into Africa by external colonisation (though one such language, Afrikaans, a descendant of colonial Dutch, is a language of Africa by virtue of its geographic distribution), and also Malagasy, the Austronesian language of Madagascar. It is useful to take as a starting point the classification of the sub-Saharan African languages into three groups as proposed by J.H. Greenberg in the mid-1960s – Niger-Congo, Nilo-Saharan and Khoisan – while keeping in mind that some of these groupings remain controversial, either in general or in particular details.

The Niger-Congo (Niger-Kordofanian) family covers most of sub-Saharan Africa, and includes not only major languages of West Africa such as Yoruba but also, as a low-level node on the family tree, the Bantu languages, dominant in most of East, Central and southern Africa. The assignment of some groups of languages to Niger-Congo, such as the Kordofanian languages spoken in the Kordofan mountains of Sudan and the Mande languages of western West Africa, remains controversial.

More controversial is the proposed Nilo-Saharan family, which would include languages spoken in a number of geographically discontinuous areas of northern sub-Saharan Africa including parts of southern Sudan running through northern Uganda and western Kenya to northern Tanzania, northern Chad and neighbouring areas and the bend of the Niger river in West Africa. While the languages of the first group form a well-defined language family, Nilotic, at the opposite extreme inclusion of Songhay on the bend of the Niger river is widely rejected.

Finally, the three main groups within Khoisan as proposed by Greenberg, namely Northern, Central (Khoe) and Southern, all spoken in southern, mainly south-western Africa, plus two geographically isolated languages of Tanzania, Hadza and Sandawe, are individually well-defined language families (or isolates). However, grouping them together as a single larger family is generally considered at least premature. Typologically, Khoisan languages are most noted for having click sounds as part of their regular

phoneme inventory, a sound type that has also been borrowed into some neighbouring Bantu languages such as Zulu and Xhosa.

2.5 The Americas

The genetic classification of the indigenous languages of the Americas is overall the most contentious, with proposals ranging from a single family covering nearly all these languages, associated especially with the name of J.H. Greenberg, to around 200 distinct families and isolates. Even widely cited intermediate proposals, such as the Hokan and Penutian families, remain controversial. In what follows, I have concentrated on some of the more widespread established families and on some of the other languages with relatively large numbers of speakers.

Two population groups of North America are distinct ethnically from the remainder, namely the Eskimos (Inuit, although this latter term properly only refers to part of the Eskimos overall) and Aleuts. The Eskimo-Aleut family contains two branches, Aleut and Eskimo. Eskimo is properly a number of different languages rather than a single language, and is spoken from the eastern tip of Siberia through Alaska and northern Canada to Greenland; in Greenland it is, under the name Greenlandic, an official language.

Another language family centred in Alaska is the Athapaskan family (more properly: Athapaskan-Eyak, with inclusion of the Athapaskan or Athabaskan languages and the single language Eyak as the two branches of the family). Most of the Athapaskan languages are spoken in Alaska and north-western Canada, though the Athapaskan language with most speakers, Navajo, is spoken in Arizona and adjacent areas. Navajo is the indigenous language of North America (Canada and the USA) with the largest number of speakers, about 150,000. Athapaskan-Aleut is related to Tlingit, together forming a grouping often referred to as Na-Dene, although this term is also used to include Haida, which may well rather be a language isolate.

Among the other major families of North America are Iroquoian (around Lakes Ontario and Erie), Siouan (the Great Plains), and Algonquian (much of the north-eastern USA and eastern and central Canada, though also extending into the Great Plains with Arapaho and Cheyenne). One interesting feature of the Algonquian languages to which it is worth devoting a short digression is obviation. In Algonquian languages, a distinction is made between two kinds of third person, namely proximate and obviative, so that where English just has one set of third person pronouns (e.g. *he*, *she*, *it*, *they*) and morphology (e.g. the third person singular present tense ending *-s*), Algonquian languages distinguish two sets. In a given text span (which must be at least a clause, but may be longer), one of the third person noun phrases is selected as proximate (the one which is in some sense the most salient at that part of the text), all other third person participants are obviative. In the remainder of the text span, the proximate participant is always referred to by proximate morphology, while other participants are referred to by obviative morphology. In this way, the ambiguity of an English sentence like *John saw Bill as he was leaving* (was it John that was leaving, or Bill?) is avoided. The following examples are from Cree:

Nāpēw atim-wa wāpam-ē-w, ē-sipwēhtē-t.
'The man saw the dog as he (the man) left.'

Nāpēw atim-wa wāpam-ē-w, ē-sipwēhtē-ýit.
'The man saw the dog as it (the dog) was leaving.'

In both sentences, 'the man' is proximate (indicated by the absence of any affix on *nā pēw* 'man'), and 'the dog' is obviative (indicated by the suffix *-wa* on *atim-wa* 'dog'). The morphology of the verb *wāpam-ē-w* 'he sees him' indicates that the agent is proximate and the patient obviative (this is important, since the word order can be varied). The prefix *ē-* on the second verb indicates that it is subordinate ('conjunct', in Algonquianist terminology). In the first sentence, the suffix *-t* on this second verb indicates a proximate subject, i.e. the subject must be the proximate participant of the preceding clause, namely *the man*. In the second sentence, the suffix *-ýit* indicates an obviative subject, i.e. the subject of this verb must be an obviative participant of the preceding clause, in this sentence the only candidate being *the dog*.

Another important family, Uto-Aztecan, includes languages spoken in both North America (the South-West) and Central America. Its Aztecan branch includes Nahuatl, whose varieties have in total over a million speakers. The ancestor of the modern dialects, Classical Nahuatl, was the language of the Aztec civilisation which flourished in Central Mexico before the arrival of the Spanish. Spoken to the south of Nahuatl entirely within Central America, the Mayan family has an equally glorious past, because of its association with the ancient Mayan civilisation. Mayan languages are spoken in southern Mexico and Guatemala, with some overspill into neighbouring Central American countries; the Mayan language with the largest number of speakers is Yucatec, with about 700,000, although several others have speaker numbers in the hundreds of thousands.

The major families of South America include Carib, Arawakan and Tupi. These language families do not occupy geographically continuous areas: Carib languages are spoken to the north of the Amazon, and predominate in the eastern part of this region; Arawakan languages, once also spoken in the West Indies, dominate further west and are also found well south of the Amazon; while Tupi languages are spoken over much of Brazil south of the Amazon and in Paraguay. One Tupi language, (Paraguayan) Guarani, with about five million speakers, is a co-official language of Paraguay and is unique among indigenous languages of the Americas in that most of its speakers are non-Indians. Hixkaryana, a Carib language spoken by about 600 people on the Nhamundá river, a tributary of the Amazon, has become famous in the linguistic literature as the first clear attestation of a language in which the word order is object–verb–subject, as in the following sentence:

Toto yahosɨye kamara. 'The jaguar grabbed the man.'

In Hixkaryana, *toto* means 'man', *kamara* means 'jaguar', while the verb *yahosɨye* has the lexical meaning 'grab' and specifies that both subject and object are third person singular. Since there is no case marking on the nouns, and since the verb morphology is compatible with either noun as subject or object, the word order is crucial to understanding of this Hixkaryana sentence (which cannot mean 'the man grabbed the jaguar'), just as the different subject–verb–object word order is crucial in English.

Quechua – properly a family of often mutually unintelligible languages rather than a single language – has about ten million speakers, primarily in Peru and Bolivia, though with offshoots north into Ecuador and Colombia and south into Chile and Argentina. It is of uncertain genetic affiliation, though often claimed to be related to the neighbouring Aymara language. Quechua was the language of the Inca civilisation, centred on Cuzco in what is now Peru.

3 The Social Interaction of Languages

As was indicated in the Preface, the notion of 'major language' is defined in social terms, so it is now time to look somewhat more consistently at some notions relating to the social side of language, in particular the social interaction of languages. Whether a language is a major language or not has nothing to do with its structure or with its genetic affiliation, and the fact that so many of the world's major languages are Indo-European is a mere accident of history.

First, we may look in more detail at the criteria that serve to define a language as being major. One of the most obvious criteria is the number of speakers, and certainly in making my choice of languages to be given individual chapters in this volume number of speakers was one of my main criteria. However, number of speakers is equally clearly not the sole criterion.

An interesting comparison to make here is between Chinese (or even more specifically, Mandarin) and English. Mandarin has far more native speakers than English, yet still English is generally considered a more useful language in the world at large than is Mandarin, as seen in the much larger number of people studying English as a second language than studying Mandarin as a second language. One of the reasons for this is that English is an international language, understood by a large number of people in many different parts of the world; Mandarin, by contrast, is by and large confined to China, and even taking all Chinese dialects (or languages) together, the extension of Chinese goes little beyond China and overseas Chinese communities. English is not only the native language of sizable populations in different parts of the world (especially the British Isles, North America and Australia and New Zealand) but is also spoken as a second language in even more countries, as is discussed in more detail in the chapter on English. English happens also to be the language of some of the technologically most advanced countries (in particular of the USA), so that English is the basic medium for access to current technological developments. Thus factors other than mere number of speakers are relevant in determining the social importance of a language.

Indeed, some of the languages given individual chapters in this volume have relatively few native speakers. Some of them are important not so much by virtue of the number of native speakers but rather because of the extent to which they are used as a lingua franca, as a second language among people who do not share a common first language. Good examples here are Swahili and Indonesian. Swahili is the native language of a relatively small population, perhaps a couple of million, primarily on the coast of East Africa, but its use as a lingua franca has spread through much of East Africa (especially Kenya and Tanzania) and beyond, so that the language is used by a total of perhaps around 50 million people. The Indonesian variety of Malay–Indonesian is the native language of perhaps 23 million, but is used as a second language by about 140,000,000 in Indonesia. In many instances, in my choice of languages I have been guided by this factor rather than by raw statistics. Among the Philippine languages, for instance, Tagalog does not have the largest number of native speakers, but I selected it because it is both the national language of the Philippines and used as a lingua franca across much of the country. A number of Indo-Aryan languages would surely have qualified for inclusion in terms of number of speakers, but they have not been assigned individual chapters because in social terms the major languages of the northern part of South Asia are clearly Hindi–Urdu and Bengali.

Another important criterion is the cultural importance of a language, in terms of the age and influence of its cultural heritage. An example in point is provided by the Dravidian

languages. Tamil does not have the largest number of native speakers; it is, however, the oldest Dravidian literary language, and for this reason my choice rested with Tamil. I am aware that many of these decisions are in part subjective, and in part contentious. As I emphasise in the Preface, the thing furthest from my mind is to intend any slight to speakers of languages that are not considered major in the contents of this volume; much of our knowledge of Language as a general characteristic of the human species comes precisely from the study of smaller, often endangered languages.

Certain languages are major even despite the absence of native speakers, as with Latin and Sanskrit. Latin has provided a major contribution to all European languages, as can be seen most superficially in the extent to which words of Latin origin are used in European languages. But even those languages that have tried to avoid the appearance of Latinity by creating their own vocabulary have often fallen back on Latin models: German *Gewissen* 'conscience', for instance, contains the prefix *ge-*, meaning 'with', the stem *wiss-*, meaning 'know', and the suffix *-en* to form an abstract noun – an exact copy of the Latin *con-sci-entia*; borrowings that follow the structure rather than the form in this way are known as calques or loan translations. Sanskrit has played a similar role in relation to the languages of India, including Hindi. Hebrew is included not because of the number of its speakers – as noted in the chapter on Hebrew, this has never been large – but because of the contribution of Hebrew and its culture to European and Middle Eastern society.

A language can thus have influence beyond the areas where it is the native or second language. A good example to illustrate this is Arabic. Arabic loans form a large part of the vocabulary of many languages spoken by Islamic peoples, even of languages that are genetically only distantly related to Arabic (e.g. Hausa) or that are genetically totally unrelated (e.g. Turkish, Persian and Urdu). The influence of Arabic can also be seen in the adoption of the Arabic writing system by many Islamic peoples. Similarly, Chinese loan words form an important part of the vocabulary of some East Asian languages, in particular Vietnamese, Japanese and Korean; the use of written Chinese characters has also spread to Japan and Korea, and in earlier times also to Vietnam.

It is important to note also that the status of a language as a major language is far from immutable. Indeed, as we go back into history we find many significant changes. For instance, the possibility of characterising English as the major language of the world is an innovation of the twentieth century. One of the most important shifts in the distribution of major languages resulted from the expansion of European languages, especially English, Spanish, Portuguese, and to a lesser extent French as a result of the colonisation of the Americas: English, Spanish and Portuguese all now have far more native speakers in the New World than in Britain, Spain or Portugal. Indeed, in the Middle Ages one would hardly have imagined that English, confined to an island off the coast of Europe, would have become a major international language.

In medieval Europe, Latin was clearly the major language, since, despite the lack of native speakers, it was the lingua franca of those who needed to communicate across linguistic boundaries. Yet the rise of Latin to such pre-eminence – which includes the fact that Latin and its descendants have ousted virtually all other languages from south-western Europe – could hardly have been foreseen from its inauspicious beginnings confined to the area around Rome. Equally spectacular has been the spread of Arabic, in the wake of the spread of Islam, from being confined to the Arabian peninsula to being the dominant language of the Middle East and North Africa.

In addition to languages that have become major languages, there are equally languages that have lost this status. The earliest records from Mesopotamia, often considered the

cradle of civilisation, are in two languages: Sumerian and Akkadian (the latter the language of the Assyrian and Babylonian empires); Akkadian belongs to the Semitic branch of Afroasiatic, while Sumerian is as far as we can tell unrelated to any other known language. Even at the time of attested Sumerian inscriptions, the language was probably already approaching extinction, and continued to be used in deference to tradition (as with Latin in medieval Europe). The dominant language of the area was to become Akkadian, but in the intervening period this too has died out, leaving no direct descendants. Gone too is Ancient Egyptian, the language of the Pharaohs and whose earliest texts are roughly contemporaneous with those of Sumerian. The linguistic picture of the Mediterranean and Near East in the year nought was very different from that which we observe today.

Social factors and social attitudes can even bring about apparent reversals in the family-tree model of language relatedness. At the time of the earliest texts from Germany, two distinct Germanic languages are recognised: Old Saxon and Old High German. Old Saxon is the ancestor of the modern Low German (Plattdeutsch) dialects, while Old High German is the ancestor of the modern High German dialects and of the standard language. Because of social changes – such as the decline of the Hanseatic League, the economic mainstay of northern Germany – High German gained social ascendancy over Low German. Since the standard language, based on High German, is now recognised as the standard in both northern and southern Germany, both Low and High German dialects are now considered dialects of a single German language, and the social relations between a given Low German dialect and standard German are in practice no different from those between a High German dialect and standard German.

One of the most interesting developments to have arisen from language contact is the development of pidgin and creole languages. A pidgin language arises from a very practical situation: speakers of different languages need to communicate with one another to carry out some practical task, but do not speak any language in common and moreover do not have the opportunity to learn each other's language properly. What arises in such a situation is, initially, an unstable pidgin, or jargon, with highly variable structure – considerably simplified relative to the native languages of the people involved in its creation – and just enough vocabulary to permit practical tasks to be carried out reasonably successfully. The clearest examples of the development of such pidgins arose from European colonisation, in particular from the Atlantic slave trade and from indenturing labourers in the South Pacific. These pidgins take most of their vocabulary from the colonising language, although their structure is often very different from those of the colonising language.

At a later stage, the jargon may expand, particularly when its usefulness as a lingua franca is recognised among the speakers of non-European origin, leading to a stabilised pidgin, such as Tok Pisin, the major lingua franca of Papua New Guinea. This expansion is on several planes: the range of functions is expanded, since the pidgin is no longer restricted to uses of language essential to practical tasks; the vocabulary is expanded as a function of this greater range of functions, new words often being created internally to the pidgin rather than borrowed from some other language (as with Tok Pisin *maus gras* 'moustache', literally 'mouth grass'); the structure becomes stabilised, i.e. the language has a well-defined grammar.

Probably at any stage in this development, from inception to post-stabilisation, the pidgin can 'acquire native speakers', i.e. become the native language of part or all of the community. For instance, if native speakers of different languages marry and have

the pidgin as their only common language, then this will be the language of their household and will become the first language of their children. Once a pidgin has acquired native speakers, it is referred to as a creole. The native language of many inhabitants of the Caribbean islands is a creole, for instance the English-based creole of Jamaica, the French-based creole of Haiti, and the Spanish- and/or Portuguese-based creole Papiamento (Papiamentu) of the Netherlands Antilles and Aruba. At an even later stage, social improvements and education may bring the creole back into close contact with the European language that originally contributed much of its vocabulary. In this situation, the two languages may interact and the creole, or some of its varieties, may start approaching the standard language. This gives rise to the so-called post-creole continuum, in which one finds a continuous scale of varieties of speech from forms close to the original creole (basilect) through intermediate forms (mesolect) up to a slightly regionally coloured version of the standard language (acrolect). Jamaican English is a good example of a post-creole continuum.

Even with hindsight, as we saw above, it would have been difficult to predict the present-day distribution of major languages in the world. It is equally impossible to predict the future. In terms of number of native speakers, it is clear that a major shift is underway in favour of non-European languages: the rate of population increase is much higher outside Europe than in Europe, and while some European languages draw some benefit from this (such as Spanish and Portuguese in Latin America), the main beneficiaries are the indigenous languages of southern Asia and Africa. It might well be that a later version of this volume would include fewer of the European languages that are restricted to a single country, and devote more space to non-European languages. Another factor is the increase in the range of functions of many non-European languages: during the colonial period European languages (primarily English and French) were used for most official purposes and also for education in much of Asia and Africa, but the winning of independence has meant that many countries have turned more to their own languages, using these as official language and medium of education. The extent to which this will lead to increase in their status as major languages is difficult to predict – at present, access to the frontiers of scholarship and technology is still primarily through European languages, especially English; but one should not forget that the use of English, French and German as vehicles for science was gained only through a prolonged struggle against what then seemed the obvious language for such writing: Latin. (The process may go back indefinitely: Cicero was criticised for writing philosophical treatises in Latin by those who thought he should have used Greek.) But at least I hope to have shown the reader that the social interaction of languages is a dynamic process, one that is moreover exciting to follow.

Bibliography

The most comprehensive and up-to-date index of the world's languages, with genetic classification, is Gordon (2005); while some data are no doubt questionable, this is certainly the most reliable such index available. For a more splitter-oriented classification, see Dryer (2005); for a more clumper-oriented one, Ruhlen (1991). Two series dealing with particular language families are the Cambridge Language Surveys (http://www.cup.cam.ac.uk/series/sSeries.asp?code=CLS) and the Routledge Language Family Series (http://www.routledge.com/books/series/Routledge_Language_Family_Series); the former concentrates on general properties of the language group in question, while the latter, initially inspired

by the first edition of the present work, provides sketches of individual languages. Though outside these series, Heine and Nurse (2000) provides a good overview of languages of Africa. Among several recent publications on endangered languages are Abley (2003) and Harrison (2007); see also http://www.ethnologue.com/nearly_extinct.asp for a list of 'nearly extinct' languages, i.e. for which 'only a few elderly speakers are still living'.

Readers wanting to delve deeper into problems of genetic classification should consult a good introduction to historical and comparative linguistics, such as Campbell (2004). For discussions of language universals and typology, reference may be made to Comrie (1989) or Croft (2002). A good introduction to language contact is Thomason (2001).

References

Abley, Mark. 2003. *Spoken Here: Travels among Threatened Languages* (Heinemann, London; Houghton Mifflin, Boston).

Campbell, L. 2004. *Historical Linguistics*, 2nd edn (Edinburgh University Press, Edinburgh).

Comrie, B. 1989. *Language Universals and Linguistic Typology*, 2nd edn (Basil Blackwell, Oxford; University of Chicago Press, Chicago).

Croft, W. 2002. *Typology and Universals*, 2nd edn (Cambridge University Press, Cambridge).

Dryer, M.S. 2005. 'Genealogical Language List', in M. Haspelmath, M.S. Dryer, D. Gil and B. Comrie (eds) *The World Atlas of Language Structures* (Oxford University Press, Oxford), pp. 584–644.

Gordon, R.G., Jr (ed.) 2005. *Ethnologue: Languages of the World*, 15th edn (SIL International, Dallas). Online version http://www.ethnologue.com/

Greenberg, J.H. 1966. 'Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements', in J. H. Greenberg (ed.) *Universals of Language*, rev. edn (MIT Press, Cambridge, MA), pp. 73–112.

Harrison, K. David. 2007. *When Languages Die: The Extinction of the World's Languages and the Erosion of Human Knowledge* (Oxford University Press, Oxford).

Heine, B. and Nurse, D. (eds) 2000. *African Languages: An Introduction* (Cambridge University Press, Cambridge).

Ruhlen, M. 1991. *A Guide to the World's Languages, Vol. I: Classification*, rev. edn (Stanford University Press, Stanford).

Thomason, S.G. 2001. *Language Contact: An Introduction* (Edinburgh University Press: Edinburgh).

Sources

I owe the Mbabaram example to R.M.W. Dixon. The Basque examples are from R. Etxepare (2003), 'Valency and Argument Structure in the Basque Verb', p. 364, in J.I. Hualde and J. Ortiz de Urbina (eds) *A Grammar of Basque* (Mouton de Gruyter, Berlin), pp. 363–426; the system is somewhat more complex than indicated in my text. The Usan examples are to be found in Ger P. Reesink (1983), 'Switch Reference and Topicality Hierarchies', *Studies in Language*, vol. 7, pp. 215–46. The discussion of Dyirbal is based on R.M.W. Dixon (1972), *The Dyirbal Language of North Queensland* (Cambridge University Press, Cambridge), especially Section 5.2.2 and Chapter 8. The Cree examples are from H. C. Wolfart and J.F. Carroll (1981), *Meet Cree*, 2nd edn (University of Alberta Press, Edmonton), p. 26. The Hixkaryana example is taken from D.C. Derbyshire (1985), *Hixkaryana and Linguistic Typology* (Summer Institute of Linguistics, Dallas; University of Texas at Arlington, Arlington) p. 32.

1

Indo-European Languages

Philip Baldi

1 Introduction

By the term *Indo-European* we are referring to a family of languages which by about 1000 BCE were spoken over a large part of Europe and parts of southwestern and southern Asia. Indo-European is essentially a geographical term: it refers to the easternmost (India) and westernmost (Europe) pre-colonial expansion of the family at the time it was proven to be a linguistic group by scholars of the eighteenth and nineteenth centuries (the term was first used in 1813). Of course, modern developments which have spread Indo-European languages around the world now suggest another name for the family, but the term *Indo-European* (German *Indogermanisch*) is now well rooted in the scholarly tradition.

Establishing languages as members of linguistic families is a process which must be accomplished using proven methods and principles of scientific analysis. During the approximately two centuries in which the interrelationships within the Indo-European family have been systematically studied, techniques to confirm and quantify genetic affiliations among its members have been developed with great success. Chief among these is the comparative method, which takes shared features among languages as its data and provides procedures for establishing protoforms (reconstruction). The comparative method is supplemented by the method of internal reconstruction and the application of principles of typological inference, which can be utilised together with the comparative method to achieve reliable reconstructions. But since space is limited and the focus of this chapter is Indo-European and not methods of reconstruction, we will restrict ourselves here to a brief review of the comparative method as it applies under normal conditions, using only data from Indo-European languages, though it should be pointed out that the method is generally applicable to the world's languages, regardless of family affiliation.

When we claim that two or more languages are genetically related, we are also claiming that they share common ancestry. And if we make such a claim about common ancestry, then our methods should provide us with a means of recovering the ancestral system, attested or not. The initial demonstration of relatedness is only a first step;

establishing well-motivated intermediate and ancestral forms is somewhat more complex. Among the difficulties are: which features in which of the languages being compared are older? which are innovations? which are the result of contact? how many shared similarities are enough to prove relatedness conclusively, and how are they weighted for significance? what assumptions do we make about the relative importance of lexical, morphological, syntactic and phonological features, and about directions of language change?

With these questions in mind, we begin the reconstruction process with the following assumption: if two or more languages share a feature which is unlikely to have arisen by accident, borrowing or as the result of some typological tendency or language universal, then *under normal circumstances* (i.e. in contrast with the rare instances of language mixing), the feature is assumed to have arisen only once and to have been transmitted to the two or more languages from a common source. The more such features are discovered and securely identified, the firmer the relationship.

In determining genetic relationship and reconstructing proto-forms using the comparative method, we usually start with vocabulary. Table 1.1 contains a number of words from various Indo-European languages which demonstrate a common core of lexical items too large and too basic to be explained either by accident or borrowing. A list of possible cognates which is likely to produce a maximum number of common inheritance items, known as the basic vocabulary list, provides many of the words we might investigate, such as basic kinship terms, pronouns, major body parts, lower numerals and other lexical fields which have proven to be resistant to borrowing in this family. From these and other data we seek to establish sets of equations known as *correspondences*, which represent statements that in a given environment X phoneme of one language will correspond to Y phoneme of another language *consistently* and *systematically* if the two languages are descended from a common ancestor.

Table 1.1 Some Basic Indo-European Terms

A NUMERALS	*one*	*two*	*three*	*four*	*five*	*six*	*seven*	*eight*	*nine*
Skt.	éka-	dvắ, dváu	tráya-	catvāra-	páñca	ṣáṭ-	saptá-	aṣṭá(u)	náva
Gk.	oînos ‘ace’	dú(w)ō	treîs	téttares, téssares	pénte, pémpe	héks	heptá	oktṓ	enné(w)a
Lat.	ūnus	duo	trēs	quattuor	quīnque	sex	septem	octō	novem
Hitt.		dā-	tēri-	Hier. Luv. paⁿta			šipta-		
Toch. A		wu	tre	śtwar	päñ	ṣäk	ṣpät	okät	ñu
B		wi	trai	śtwer	piś	ṣkas	ṣuk(t)	okt	ñu
OIr.	ōen	dāu, do	trī	ceth(a)ir	cōic	sē	secht	ocht	noī
Go.	ains	twai	þreis	fidwōr	fimf	saíhs	sibun	ahtau	niun
OCS	(jed)inŭ	dŭva	trĭje	četyre	pętĭ	šestĭ	sedmĭ	osmĭ	devętĭ
Lith.	víenas	dù	trỹs	keturì	penkì	šešì	septynì	aštuonì	devynì
Arm.	mi	erku	erek‘	čork‘	hing	vec‘	evt‘n	ut‘	inn
Alb.	nji, një	dü	tre, tri	katër	pesë	gjashtë	shtatë	tetë	nëndë

Table continued on next page.

Table 1.1 (*continued*)

B	ANIMAL NAMES	*mouse*	*wolf*	*cow*	*sheep*	*pig*	*dog*	*horse*
	Skt.	mū́ṣ-	vṛ́ka-	gáv-	ávi-	sūkará-	śvá-	áśva-
	Gk.	mûs	lúkos	boûs	ó(w)is	hûs	kúōn	híppos
	Lat.	mūs	lupus	bōs	ovis	sūs	canis	equus
	Hitt.							
	Toch. A			ko			ku	yuk
	B		walkwe	keu	eye	suwo	ku	yakwe
	OIr.		olc ‘evil’	bō	oī		cū	ech
	OHG	mūs	wulfs OIc.	kȳr OHG	ouwi	swein	hunds OE	eoh
	OCS	myšĭ	vlĭkŭ	gomŭno ‘threshing floor’	ovĭnŭ	svinija	Russ. súka; ‘bitch’	
	Lith.		viĨkas Latv.	gùovs Lith.	avìs Latv.	suvęns	šuō (OLith.)	ešvà, ašvà, ‘mare’
	Arm.	mukn		kov	hoviw ‘shepherd’		šun	ēš
	Alb.	mī	ulk			thi		

C	BODY PARTS	*foot*	*heart*	*eye*	*tongue*
	Skt.	pád-	śrád-dhā- ‘put the heart in trust’	ákṣi-	jihvā́
	Gk.	poús (gen. podós)	kardíā	óp-somai ‘I will see’	
	Lat.	pēs (gen. pedis)	cor (gen. cordis)	oculus	lingua
	Hitt.	pāta-(Luw.)	kard-		
	Toch. A	pe	kri ‘will’	ak	käntu
	B	paiyye	kāryāñ (pl.)	ek	kantwo
	OIr.	īs ‘below’	cride	enech	teng
	Go.	fōtus	haírtō	augō	tuggō
	OCS	pĕšĭ ‘on foot’	srŭdĭce	oko	języ̨kŭ
	Lith.	pādas ‘sole’	širdìs	akìs	liežùvis
	Arm.	otn	sirt	akn	lezu
	Alb.	(për)posh ‘under’		sü	

D	KINSHIP TERMS	*mother*	*father*	*sister*	*brother*
	Skt.	mātár-	pitár-	svásar-	bhrā́tar-
	Gk. (Dor.)	mā́tēr	patḗr	eór (voc.) (Dor.)	phrā́tēr < ‘member of a brotherhood’
	Lat.	māter	pater	soror	frāter
	Hitt.				
	Toch. A	mācar	pācar		pracar
	B	mācer	pācer		procer
	OIr.	māthair	athair	siur	brāthair
	OIc.	mōðir Go.	fadar	swistar	Go. brōþar
	OCS	mati		sestra	bratrŭ, bratŭ
	Lith.	mótė ‘woman’		sesuō	brólis
	Arm.	mair	hair	k‘oir	ełbair
	Alb.	motrë			

Table continued on next page.

Table 1.1 (*continued*)

E	GENERAL TERMS	*full*	*race, kind*	*month*	*die, death*	*old*	*vomit*
	Skt.	pūrṇá-	jánas-	mā́s-	mṛtá-	sána-	vámiti
	Gk.	plḗrēs	génos	mḗn	ámbrotos ‘immortal’	hénos ‘last year’s’	eméō
	Lat.	plēnus	genus	mēnsis	mortuus	senex	vomō
	Hitt.				mer-		
	Toch. A			mañ			
	B			meñe			
	OIr.	lān	gein ‘birth’	mī	marb	sen	
	Go.	fulls	kuni	mēna, mēnōþs	maúrþr	sineigs OIc	vāma ‘sickness’
	OCS	plŭnŭ		OBulg. měsęcĭ	mĭrǫ		vémti
	Lith.	pìlnas		mė́nuo	miřtis	sẽnas	
	Arm.	li	cin ‘birth’	amis	mard ‘mortal’	hin	
	Alb.	plot		muai			

In order to illustrate the comparative method we will briefly and selectively choose a few items from Tables 1.1 and 1.2, restricting our data to fairly clear cases.

	mouse		*mother*		*nine*
Skt.	mū́ṣ-		mātā́r		náva
Gk.	mûs	(Dor.)	mā́tēr		enné(w)a
Lat.	mūs		māter		novem
OHG	mūs	OIc.	mōðir	Go.	niun
	dead		*dog*		*race, kind*
Skt.	mṛtá-		śvá-		jánas-
Gk.	ámbrotos ‘immortal’		kúōn		génos
Lat.	mortuus		canis		genus
Go.	maúrþr ‘murder’		hunds		kuni
	I am		*vomit*		*old*
Skt.	ásmi		vámiti		sána-
Gk.	eimí		eméō		hénos ‘last year’s’
Lat.	sum		vomō		senex
Go.	im	OIc.	vāma ‘sickness’	Go.	sineigs

We will first look only at the nasals *m* and *n*. Lined up for comparative analysis they look like this:

	mouse	*mother*	*nine*	*dead*	*dog*	*race, kind*	*I am*	*vomit*	*old*
Skt.	m-	m-	n-	m-	-∅-	-n-	-m-	-m-	-n-
Gk.	m-	m-	-nn-	-m(b)-	-n	-n-	-m-	-m-	-n-
Lat.	m-	m-	n-	m-	-n-	-n-	-m	-m-	-n-
Gmc.	m-	m-	n-	m-	-n-	-n-	-m	-m-	-n-

Before we begin reconstructing we must be sure that we are comparing the appropriate segments. It is clear that this is the case in ‘mouse’, ‘mother’, ‘dog’, ‘race, kind’, ‘I am’, ‘vomit’ and ‘old’, but less clear in ‘nine’, ‘dead’ and ‘dog’. What of the double *n* in Gk. *enné(w)a*? Internal reconstruction reveals that *en-* is either a prefix or the

Table 1.2 Inflectional Regularities in Indo-European Languages

A *Examples of Verb Inflection*

		I am	*he, she is*
Skt.		ásmi	ásti
Gk.		eimí	estí
Lat.		sum	est
Hitt.		ēšmi	ēšzi
Toch.	A		
	B		ste
OIr.		am	is
Go.		im	ist
OCS		jesmĭ	jestĭ
OLith.		esmì	êsti
Arm.		em	ē
Alb.		jam	është

B *Examples of Noun Inflection*

		tooth				
		Skt.	*Gk.*	*Lat.*	*Go.*	*Lith.*
Sg.	nom.	dán	odṓn	dēns	*tunþus	dantìs
	gen.	datás	odóntos	dentis	*tunþáus	dantiẽs
	dat.	daté	odónti	dentī	tunþáu	dañčiui
	acc.	dántam	odónta	dentem	tunþu	dañtį
	abl.	datás		dente		
	loc.	datí				dantyjè
	inst.	datā́				dantimì
Pl.	nom.	dántas	odóntes	dentēs	*tunþjus	dañtys
	gen.	datā́m	odóntōn	dentium	tunþiwē	dantų̃
	dat.	dadbhyás	odoũsi	dentibus	tunþum	dantìms
	acc.	datás	odóntas	dentēs	tunþuns	dantìs
	abl.	dadbhyás		dentibus		
	loc.	datsú				dantysè
	inst.	dadbhís				dantimìs

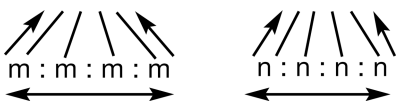
C *Examples of Pronoun Inflection*

		I, me					
		Skt.	*Gk.*	*Lat.*	*Hitt.*	*Go.*	*OCS*
	nom.	ahám	egṓ	ego	uk	ik	azŭ
	gen.	máma(me)	emoû(mou)	meī	ammēl	meina	mene
	dat.	máhyam(me)	emoí(moi)	mihī	ammuk	mis	mĭnē(mi)
	acc.	mā́m(mā)	emé(me)	mē(d)	ammuk	mik	mene(mę)
	abl.	mát		mē(d)	ammēdaz		
	loc.	máyi			ammuk		mĭnē
	inst.	máyā					mŭnojǫ
		you (sg.)					
		Skt.	*Gk.*	*Lat.*	*Hitt.*	*Go.*	*OCS*
	nom.	tvám	sú	tū	zik	þu	ty
	gen.	táva(te)	soú(sou)	tuī, tīs	tuēl	þeina	tebe
	dat.	túbhyam(te)	soí(soi)	tibī	tuk	þus	tebē(ti)
	acc.	tvā́m(tvā)	sé(se)	tē(d)	tuk	þuk	tebe(tę)
	abl.	tvát		tē(d)	tuēdaz		
	loc.	tváyi			tuk		tebē
	inst.	tváyā					tobojǫ

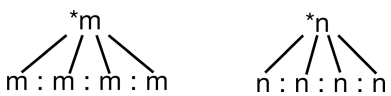
Note: Forms in parentheses are enclitic variants.

outcome of some pre-Greek phonological process; in either case, the first *n* is definitely outside the comparative equation. Likewise with Gk. *ámbrotos* 'immortal': the *a-* is a prefix meaning 'not' (= Lat. *in-*, Go. *un-*, etc.), and the *b* results from a rule of Greek in which the sequence *-mr-* (a-mrotos) results in *-mbr-*, with excrescent *b* (cf. Lat. *camera* > Fr. *chambre*). Similarly with Skt. *śvá-* 'dog', which has no *n* itself, but reveals an *n* in the stem of the genitive case form *śúnas*. So the nasals in these sets do indeed align, leaving us with consistent *m* and *n* correspondences in their respective sets:

These alignments represent the horizontal or comparative dimension of the reconstruction process. Next we 'triangulate' the segments, adding the vertical, or historical dimension:



Finally, after checking all the relevant data and investigating their distributional patterns, we make a hypothesis concerning the proto-sound. In these two cases there is only one reasonable solution, namely **m* and **n*:

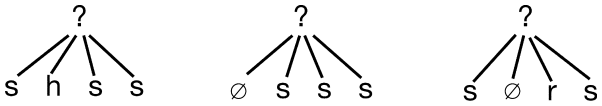


At this stage of the analysis we invert the equation, so that it reads '**m* > (develops into) *m* and **n* > *n*' in the specified daughter languages.

Neat correspondences such as these are more the exception than the rule in comparative analysis. It is far more common to find sets in which only a few of the members have identical segments. But the comparative method does not require that segments **match**, only that they **correspond** systematically. Consider the following data from Table 1.1, supplemented by some additional material:

	pig	*old*	*race, kind* (gen. case)	*be*
Skt.	sūkará-	sána-	jánasas	ástu 'let him be!'
Gk.	hûs	hénos 'last year's'	géneos (génous)	éō̃ (ễ) 'I might be'
Lat.	sūs	senex	generis	erō 'I will be'
Go.	swein	sineigs	(OCS slovese 'word')	ist 'he/she is'

We are concentrating here on the correspondences which include *s*, *h* and *r*. In 'pig' and 'old' we have the set *s* : *h* : *s* : *s* initially (cf. also 'six' and 'seven'). In final position we find ∅ : *s* : *s* : *s* in 'pig' and 'old' (cf. also 'one', 'three', 'mouse' and 'wolf', among others). And in medial position we have *s* : ∅ : *r* : *s* in 'race, kind' (gen.) and 'be'. What is (are) the proto-sound(s)?



A brief look at the languages in question takes us straight to **s* for all three correspondences: **s* > *h* in Greek initially (weakens), and disappears completely medially (*éō*), yielding a phonetically common pattern of *s* > *h* > ∅; Sanskrit final ∅ in 'old' is only the result of citing the Sanskrit words in their root forms; the full nominative form (as in the other languages) would contain *s* as well (*ékas*, etc.); and the medial Latin *r* is the result of rhotacism, whereby Latin consistently converts intervocalic *s* to *r* (cf. *es-* 'be', *erō* 'I will be'; (nom.) *flōs* 'flower' (gen.) *flōris*).

From these few, admittedly simplified examples we can get an idea of the comparative method which, when supplemented by adequate information about the internal structure of the languages in question and by a consideration of all the relevant data, can produce consistent reconstructions of ancestral forms. It is with such methods that Proto-Indo-European (PIE) has been reconstructed.

2 The Languages of the Indo-European Family

The Indo-European languages are classified into eleven major groups (ten if Baltic and Slavonic are considered together as Balto-Slavonic). Some of these groups have many members, while some others have only one. Of the eleven major groups, nine have modern spoken representatives while two, Anatolian and Tocharian, are extinct.

2.1 Indo-Iranian

The Indo-Iranian group has two main subdivisions, Indo-Aryan (Indic) and Iranian. The similarities between the two subdivisions are so consistent that there is no question about the status of Indo-Iranian intermediate between Proto-Indo-European and the Indic and Iranian subgroups. The Indo-Aryan migrations into the Indian area took place some time in the second millennium BCE.

2.1.1 Indo-Aryan (or Indic)

(See Chapter 20.)

2.1.2 Iranian

(See Chapter 24.)

2.2 Hellenic (Greek)

(See Chapter 19.)

2.3 Italic

(See Chapter 7.)

2.4 Anatolian

The Anatolian languages were unknown to modern scholars until archaeological excavations during the first part of the last century in Boğazköy, Turkey, yielded texts which were written primarily in Hittite, the principal language of the Anatolian group. The texts, which date from approximately the seventeenth to the thirteenth centuries BCE, were written in cuneiform script and contained not only Hittite, but Akkadian and Sumerian as well. Decipherment proceeded quickly and it was demonstrated by B. Hrozný in 1915 that the Hittite in the texts was an Indo-European language. It was later shown that Hittite contained a large number of archaic features not found in other Indo-European languages, which resulted in revised reconstructions of the proto-language. Now totally extinct, the Anatolian group contains, in addition to the amply attested Hittite, Luwian (Cuneiform and Hieroglyphic), Lycian, Palaic, Lydian, and possibly Carian.

2.5 Tocharian

Around the turn of the twentieth century a large amount of material written in an unknown language was discovered in the Chinese Turkestan (Tarim Basin) region of Central Asia. The language represented in these texts is now known as Tocharian; it was quickly recognised as Indo-European, despite its extreme eastern location. The Tocharian documents are chiefly of a religious nature, but also contain commercial documents, caravan passes and medical and magical texts. There are two dialects of Tocharian: Tocharian A, also known as East Tocharian or Turfan, and Tocharian B, also known as West Tocharian or Kuchean. The texts found in Chinese Turkestan are all from the period CE 500 to 1000, so this language has not played the same role as other twentieth-century discoveries like Hittite and Mycenaean Greek in the shaping of reconstructed Proto-Indo-European.

2.6 Celtic

The Celtic languages are largely unknown until the modern period, though it is clear from inscriptional information and place and river names that Celtic languages were once spread over a fairly wide section of Europe in the pre-Christian era. The Celtic languages are commonly classified into two groups: the Continental group, comprising the extinct Celtiberian (Hispano-Celtic), Gaulish and Galatian; and the Insular group, which contains two subdivisions, namely the Gaelic (Goidelic) group, made up of Irish, Scots Gaelic and the extinct Manx, and the Brythonic (Brittanic) group, made up of Welsh, Breton and the extinct Cornish. The oldest records of Celtic are some sepulchral inscriptions from the fourth century CE, and Old Irish manuscripts which date from the late seventh to early eighth century CE.

Many specialists believe that the Celtic and Italic languages have a remote relationship (Italo-Celtic) intermediate between the disintegration of Proto-Indo-European and the establishment of the separate Celtic and Italic groups.

2.7 Germanic

(See Chapter 2.)

2.8 Slavonic

(See Chapter 14.)

2.9 Baltic

This highly conservative group of Indo-European languages has played a significant role in Indo-European studies. Despite the fact that the oldest useful recorded material from Baltic dates from the mid-fourteenth century CE, Baltic has preserved many archaic features, especially in morphology, which scholars believe existed in Proto-Indo-European.

Only two Baltic languages are spoken today, Lithuanian and Latvian (or Lettish). Many others are now extinct, including Semigallian, Selonian, Curonian, Yotvingian and Old Prussian. Old Prussian is the most important of these; it became extinct in the early eighteenth century, but provides us with our oldest written documentation of the Baltic group.

The Baltic languages are considered by many specialists to be in a special relationship with the Slavonic languages. Those who follow such a scheme posit a stage intermediate between Proto-Indo-European and Baltic and Slavonic called Balto-Slavonic.

2.10 Armenian

Armenian was probably established as a language in its historic homeland in the southern Caucasus and western Turkey by the sixth century BCE. The first records of the language are from the fifth century CE, and it shows considerable influence from Greek, Arabic, Syriac and especially Persian. In fact, so extreme is the foreign influence on Armenian that it was at first thought to be a radical dialect of Persian rather than a language in its own right. Written in an alphabet developed in the fifth century, the language is quite conservative in many of its structural features, especially inflectional morphology and, by some accounts, consonantal phonology.

2.11 Albanian

The remote history of Albanian is unknown, and although there are references to Albanians by Greek historians in the first century CE, we have no record of the language until the fifteenth century. Much influenced by neighbouring languages, Albanian has proven to be of marginal value in the reconstruction of Proto-Indo-European. There are two principal dialects of Albanian: Gheg, spoken in the north and in parts of the former Yugoslavia, and Tosk, spoken in southern Albania and various communities in Greece and Italy.

2.12 Fragmentary Languages

In addition to the eleven major groups, there remain a number of 'minor' Indo-European languages which are known only in fragments, glosses, inscriptions and other unpredictable sources. Though there is some dispute about the Indo-European character of some of these languages, scholars generally agree on the following as Indo-European: Ligurian (Mediterranean region), Lepontic (possibly affiliated with Celtic), Sicel (possibly affiliated with Italic), Thraco-Phrygian (frequently connected with Armenian and Albanian), Illyrian (especially prevalent along the Dalmatian coast), Messapic (with uncertain Italic or Albanian connections), and Venetic (probably connected with Italic). None of these languages exists in sufficient material detail to be of systematic value in the reconstruction of Proto-Indo-European.

3 The Structure of Proto-Indo-European

There have been many attempts to reconstruct Proto-Indo-European from the evidence of the daughter languages. The discoveries of Hittite, Tocharian and Mycenaean Greek in the last century have modified the data base of Indo-European studies, so it is not surprising that there have been frequent changes in views on Proto-Indo-European. Also, there have been a refinement of technique and an expansion of knowledge about language structure and language change which have modified views of the protolanguage. In this section we will briefly review past and present thinking on Proto-Indo-European phonology, and we will then discuss commonly held positions on the morphological and syntactic structure of the proto-language.

3.1 Phonology

3.1.1 Segmental Phonology

The first systematic attempt to reconstruct the sound system of Proto-Indo-European was by A. Schleicher in the first edition of his *Compendium der vergleichenden Grammatik der indogermanischen Sprachen* in 1861. Using the sound correspondences worked out by his predecessors, Schleicher proposed the consonant system as in Table 1.3 (from the 1876 edition, p. 10).

Schleicher's three-vowel system of *a*, *i* and *u* was based primarily on the pattern found in Sanskrit whereby 'basic vowels' are modified by combinatory processes which the Indian grammarians called *guṇa* 'secondary quality' and *vr̥ddhi* 'growth, increment'. By these processes the basic three-vowel system is modified by the prefixation of *a* as follows (1876: 11):

Basic Vowel	*First Increment*	*Second Increment*
a	a + a → aa	a + aa → āa
i	a + i → ai (e)	a + ai → āi
u	a + u → au (o)	a + au → āu

The reconstructed PIE system is not identical to the Sanskrit system; it is, however, patterned on it.

Schleicher's model soon gave way to the one proposed by the Neogrammarians, a group of younger scholars centred at Leipzig who had quite different views about Proto-Indo-European, and about language change generally, from their predecessors.

Table 1.3 Schleicher's Reconstructed System

	unaspirated		*aspirated*	*spirants*		*nasals*	*r*
	vls.	*vd.*	*vd.*	*vls.*	*vd.*	*vd.*	*vd.*
Guttural	k	g	gh				
Palatal					j		
Lingual							r
Dental	t	d	dh	s		n	
Labial	p	b	bh		v	m	

The Neogrammarian system is embodied in the classic work of K. Brugmann (1903: 52), as in Table 1.4.

Brugmann's system is much more elaborate than Schleicher's in almost every respect: there are more occlusives (stops), more fricatives, diphthongs, etc. But probably the most significant difference is in the vowel system. Brugmann proposes a six short, five long vowel system which is much more like that of Greek or Latin than that of Sanskrit. This change was brought about by the discovery that a change had taken place whereby Sanskrit collapsed PIE *ĕ, *ŏ, *ă into ă (cf. Lat. *sequor*, Gk. *hépomai*, Skt. *sáce* 'I follow' (**e*); Lat. *ovis*, Gk. *ó(w)is*, Skt. *ávi-* 'sheep' (**o*); Lat. *ager*, Gk. *agrós*, Skt. *ájra-* 'field, plain' (**a*)). From this it could be seen that Sanskrit was not to be considered closest to the proto-language in all respects.

The Neogrammarian system, which in modified form still finds adherents today, was put to the test by the theories of Saussure (1879) and the findings of Kuryłowicz (1927) and others. Based on the irregular behaviour of certain sounds in the daughter languages, Saussure proposed that Proto-Indo-European had contained sounds of uncertain phonetic value which he called 'coefficients sonantiques'. According to Saussure, these sounds were lost in the daughter languages but not before they left traces of their former presence on the sounds which had surrounded them. For example, there is no regular explanation for the difference in vowel length between the two forms of Gk. *hístāmi* 'I stand' and *stătos* 'stood'. Saussure theorised that originally the root had been **steA* (A = a coefficient sonantique). The A had coloured the *e* to *a* and had lengthened it to *ā* in *hístāmi* before disappearing. The major changes ascribed to the action of these sounds include changing *e* to *o*, *e* to *a* and lengthening preceding vowels.

This new theory, based on abstract principles, was put to use to explain a wide range of phonological and morphological phenomena in various Indo-European languages. It came to be called the 'laryngeal theory', since it is thought that these sounds may have had a laryngeal articulation. Proposals were made to explain facts of Indo-European

Table 1.4 Brugmann's Reconstructed System

Consonants										
Occlusives:	p	ph	b	bh	(labial)					
	t	th	d	dh	(dental)					
	k̂	k̂h	ĝ	ĝh	(palatal)					
	q	qh	g	gh	(velar)					
	qᵘ	qᵘh	gᵘ	gᵘh	(labio-velar)					
Fricatives:	s	sh	z	zh	þ	þh	ð	ðh	(j)	
Nasals:	m	n	n̂	ŋ						
Liquids:	r	l								
Semi-vowels:		i̯	u̯							
Vowels										
A	Vowels:	e	o	a	i	u	ə			
		ē	ō	ā	ī	ū				
B	Diphthongs:	ei̯	oi̯	ai̯	əi̯		eu̯	ou̯	au̯	əu̯
C	Syllabic Liquids and Nasals:	l̥	r̥	m̥	n̥	n̥̂	ŋ̥			
		l̥̄	r̥̄	m̥̄	n̥̄	n̥̂̄	ŋ̥̄			

Source: Brugmann 1903: 67, 89, 122–38.

root structure, ablaut relations (see Section 3.2.2) and other aspects of PIE phonology and morphology. Many proposals concerning the exact number of laryngeals, and their effects, were made. Some scholars worked with one, others with as many as ten or twelve. It remained an unverified theory until 1927, when Kuryłowicz demonstrated that Hittite preserved laryngeal-like sounds (written as ḫ or ḫḫ) precisely in those positions where Saussure had theorised they had existed in Proto-Indo-European. Some examples: Hitt. *ḫanti* 'front': Lat. *ante*; Hitt. *ḫarkiš-* 'white': Gk. *argḗs* Hitt. *palḫiš*; 'broad': Lat. *plañus*; Hitt. *meḫur* 'time': Go. *mēl*; Hitt. *u̯aḫanzi* 'they turn': Skt. *vāya-* 'weaving'; Hitt. *newaḫḫ-* 'renew': Lat. *novāre.*

The empirical confirmation that Hittite provided for Saussure's theories led to a reworking of the Proto-Indo-European sound system. We may take the scheme proposed by W. Lehmann (1952: 99) as representative of these developments as in Table 1.5.

There are many differences between Lehmann's system and that of Brugmann. Note in particular the postulation of only one fricative, *s*, the lack of palatals, diphthongs, voiceless aspirates and schwa. These were all given alternative analyses, partly based on the four laryngeals which Lehmann assumed.

In a further refinement, O. Szemerényi (1999: 150) proposed the system in Table 1.6, which is a bit more robust than that of Lehmann, say, but not as elaborate as that of Brugmann. It is a system that many specialists feel best represents the facts of the IE languages. Note the absence of laryngeals in Szemerényi's account of this stage of PIE.

Table 1.5 Lehmann's Reconstructed System

Obstruents:	p	t	k	k^w			
	b	d	g	g^w			
	b^h	d^h	g^h	g^{wh}			
		s					
Resonants:	m	n					
	w	r	l	y			
Vowels:		e	a	o	e		
	i·	e·	a·	o·	u·		
Laryngeals:			x	γ	h		ʔ

Table 1.6 Szemerényi's Reconstructed System

Obstruents:	p	p^h	b	b^h			
	t	t^h	d	d^h			
	(k′	k'^h	g′	g'^hʔ)			
	k	k^h	g	g^h			
	k^w	k^{wh}	g^w	g^{wh}			
Nasals	n	m					
Semivowels	y	w					
Liquids	l	r					
Fricatives	s	h					
Syllabic Liquids and Nasals:			n̥	m̥	n̥̄	m̥̄	
			l̥	r̥	l̥̄	r̥̄	
Vowels:							
i		u			ī	ū	
e	ə	o			ē	ō	
	a				ā		

Criticisms of the traditional system centre on the typological naturalness of the overall system. While faithful to the comparative method, such a system seems to be in conflict with known patterns of phonological structure in attested languages. One problem lies in the presence of the voiced aspirated stops without a corresponding series of voiceless aspirates. A principle of typological inference stipulates that the presence of a marked member of a correlative pair implies the presence of the unmarked member of that pair: thus *bh* ⊃ *ph*. And as T. Gamkrelidze puts it (1981: 591): 'Reconstructed systems should be characterised by the same regularities which are found in any historical system.'

Pursuing the dicta of typological structure and dependency, some scholars have recently followed a different approach to Indo-European sound structure. The focus of the effort has been the obstruent system of the proto-language, which has long presented special challenges to Indo-European scholars. Chief among the problems are the following:

(a) The traditional system without voiceless aspirates is in violation of certain markedness principles. But the solution which calls for a voiceless aspirated series only begs the question, since only one language (Sanskrit) has the four-way distinction of voiced/voiceless, aspirated/unaspirated. Thus the elaborate Proto-Indo-European system seems to rely far too heavily on Sanskrit, and is unjustified for the other groups.
(b) There has always been a problem with **b*. It is extremely rare, and those few examples which point to **b* (e.g. Lith. *dubùs*, Go. *diups* 'deep') are by no means secure.
(c) There are complicated restrictions on the co-occurrence of obstruents in Proto-Indo-European roots (called 'morpheme' or 'root structure' conditions) which are only imperfectly handled with traditional reconstructions. They are that a root cannot begin and end with a plain voiced stop, and a root cannot begin with a plain voiceless stop and end with a voiced aspirate, or vice versa.
(d) Plain voiced stops as traditionally reconstructed almost never occur in reconstructed inflectional affixes, in which Proto-Indo-European was rich. This is a distributional irregularity which cannot be explained under the traditionally reconstructed system.
(e) It has long been a curiosity to Indo-European scholars that both Germanic and Armenian underwent similar obstruent shifts (the Germanic one came to be celebrated as 'Grimm's Law', and forms the backbone of much pre- and post-Neogrammarian thinking on sound change):

'Grimm's Law' and the Armenian Consonant Shift

PIE					*Gmc.*				*Arm.*			
*p	t	k	k^w	>	f	þ	h	h^w	h	t‘	s	k‘
*b	d	g	g^w	>	p	t	k	k^w/k	p	t	c	k
*bh	dh	gh	gh^w	>	b	d	g	g^w/g	b	d	j	g

In the revised reconstruction of the obstruent system (commonly known as the 'glottalic theory'), the pattern in the occlusives is based on a three-way distinction of voiceless stops/voiced aspirates/glottalised stops (see Hopper 1981, Gamkrelidze 1981, Gamkrelidze and Ivanov 1995). The traditional plain voiced stops are now interpreted as glottalised stops (ejectives).

Reconstructed Obstruents in the 'Glottalic Theory'

	I *Glottalised*	*II* *Voiced Aspirates/ Voiced Stops*	*III* *Voiceless Aspirates/ Voiceless Stops*
Labial	(p')	b^h/b	p^h/p
Dental	t'	d^h/d	t^h/t
Velar	k'	g^h/g	k^h/k
Labio-velar	k'^w	g^{wh}/g^w	k^{wh}/k^w

The distribution of these segments has been a matter of some debate, and indeed each Indo-European language seems to have generalised one allophone or another, or split allophones, according to differing circumstances.

This new system seems to provide phonetically natural solutions to the five problems posed above:

(a) The system with the three-way distinction above violates no naturalness condition or typological universal. In fact, it is a system found in modern Armenian dialects. Under this view, Indo-Iranian is an innovator, not a relic.
(b) The near absence of **b* now finds a simple solution. In systems employing glottalised stops, the labial member is the most marked. Thus this gap, unexplained by traditional views, is no longer anomalous.
(c) The complicated morpheme structure restrictions turn out to be fairly simple: two glottalised stops cannot occur in the same root; furthermore, non-glottalised root consonants must agree in voicing value.
(d) The absence of plain voiced stops in inflections turns out to be an absence of glottalics in the new reconstruction. Such a situation is typologically characteristic of highly marked phonemes such as glottalised sounds (Hopper 1981: 135).
(e) Under the new system the parallel Germanic and Armenian consonant 'shifts' turn out to reflect archaisms rather than innovations. All the other groups have undergone fairly regular phonological changes which can be efficiently derived from the system just outlined.

Proto-Indo-European generally, not just in phonology, is a 'rolling' concept. That is, this unwritten proto-language, which has been dated to anywhere between 2500 and 7000 BCE, certainly underwent many changes of its own during its evolution and eventual break-up into the attested descendant systems. What this means is that arriving at a uniform reconstruction that captures all the internal developments and external influences that must have affected the language over many millennia is a formidable task. In this context it is not surprising that differences exist among investigators on the structure of the phonology of PIE.

3.1.2 Ablaut

In the oldest stages of Proto-Indo-European, word roots were differentiated in their various grammatical functions by a modification of the root-vowel, a pattern which is recoverable throughout the attested languages. For example, in Latin there are three forms *te*gō/*to*ga/*tē*gula 'I cover/toga (garment)/roof tile', which represent three different forms of the root *teg-* 'cover', each with a different vowel (*e/o/*ē). The same pattern can be seen in various representations of the 'father' word in Greek: *patḗ*ra/

*eupát**o**ra*/*pat'**ḗ**r*. This type of vowel modification or alternation is known as 'ablaut' or 'vowel gradation'.

Vowel gradation patterns were based on the interplay of both vowel quality such as *e*/*o* (qualitative ablaut) and vowel quantity or length such as *e*/*ē* (quantitative ablaut). The main alternations were between the basic root-vowel, usually *e*, called the 'normal grade', alternating with *o* ('*o*-grade'), zero (∅) ('zero-grade'), lengthened grade (*ē*) and lengthening plus change (lengthened *ō*-grade). In what follows the quantitative and qualitative ablaut types will be presented separately, though it should be emphasised that this is one system, not two. The two are separated here because the daughter languages typically generalised either the qualitative or quantitative system, or eliminated ablaut altogether.

	e-grade		*o-grade*		*∅-grade*	
Gk.	pét-omai	'I fly'	pot-ḗ	'flight'	e-pt-ómēn	'I flew'
Gk.	ékh-ō	'I have'	ókh-os	'carriage'	é-skh-on	'I had'
Lat.	sed-eō	'I sit'	sol-ium (<*sodium)	'throne'		
Lat.	reg-ō	'I rule'	rog-us(?)	'funeral-pyre'		
Lat.	teg-ō	'I cover'	tog-a	'a covering'		
Gk.	leíp-ō	'I leave'	lé-loip-a	'I left'	é-lip-on	'I left'
Lat.	fīd-ō (<feidō)	'I trust'	foed-us	'agreement'	fid-ēs	'trust'
Gk.	peíth-ō	'I persuade'	pé-poith-a	'I trust'	é-pith-on	'I persuaded'
Gk.	dérk-omai	'I see'	dé-dork-a	'I saw'	é-drak-on	'I saw'
Gk.	pénth-os	'grief'	pé-ponth-a	'I suffered'	é-path-on	'I suffered'

Qualitative Ablaut

The primary qualitative relations were based on the vowels *e* ~ *o* ~ ∅ *(ei* ~ *oi* ~ *i; er* ~ *or* ~ *r̥; en* ~ *on* ~ *n̥*, *etc.)*. Different forms of a morpheme were represented by different ablaut grades. This system is rather well preserved in Greek, but is recoverable in nearly every Indo-European language to one degree or another. (Note: *e* ~ *o* ~ ∅ alternation is not the only series, nor does this account consider the many interactions between vowel length and quality.)

Quantitative Ablaut

Quantitative ablaut patterns are primarily based on the alternations of 'normal', 'lengthened', and 'reduced' varieties of a vowel, e.g. *o* : *ō* : ∅; *e* : *ē* : ∅; *a* : *ā* : ∅. While represented vestigially in a wide number of Indo-European languages, (cf. Lat. *pēs*, gen. *pedis* 'foot'; *vōx* 'voice', *vocō* 'I call'; Gk. *patḗr*, *patrós* (gen.), *patéra* (acc.) 'father'), the quantitative system is most systematically represented in Sanskrit. A few good examples are:

Normal grade	*Lengthened grade*	*Reduced grade*
pát-ati 'he/she flies'	**pāt**-áyati 'he/she causes to fall'	pa-**pt**-imá 'we flew'
kar-tŕ̥- 'doer'	**kār**-yá- 'business'	**kr̥**-tá- 'done'

Quantitative vowel alternation, in conjunction with the qualitative type, was an important means of morphological marking in Proto-Indo-European, providing a basis for distinguishing different grammatical representations of a morpheme.

3.1.3 Accent

Because of the widely different accentual patterns found in the daughter languages, reconstructing the accent of Proto-Indo-European has proven challenging. The most widely accepted accounts of Proto-Indo-European accent, which rely heavily on the evidence of Sanskrit, Greek, Baltic and Slavic, suggest that it was a dominantly pitch accent system. Every word (except clitics, which were unaccented) had one accented syllable which received high-pitch accent. The accent was 'free' in that it could fall on any syllable in a word, its specific position being conditioned by morphological considerations. Accent was one means of marking grammatical categories in Proto-Indo-European. (For a parallel, cf. Eng. *rébel* (n.): *rebél* (v.); *cónflict* (n.): *conflíct* (v.).)

For example, in certain nouns some cases are typically accented on the inflections, while others are accented on the root, as can be seen in some representatives of the word for 'foot'. The nominative and accusative cases, the so-called 'strong cases', have root accent, while the genitive and dative, locative (and instrumental) have inflectional accent, indicating that accent is interacting with case markers to indicate grammatical function.

Root/Inflectional Accent (Nouns)

	Gk.	*Skt.*
nom.	poús	pā́d
acc.	póda	pā́dam
gen.	podós	padás (gen./abl.)
dat.		padé
loc.	podí (actually dat.)	padí

Similarly, some verbal forms are accented on roots, some on inflections, as can be seen in the Sanskrit verb 'to turn': *vártāmi*, (pres.) *vavárta* (perf. sg.) *vavr̥timá* (perf. pl.) *vr̥tanáḥ* (part.).

3.2 Morphology

The unevenness of historical records and huge chronological gaps among many of the languages (e.g. ca. 3,000 years between Hittite and Lithuanian) pose special problems for the reconstruction of PIE. These problems surely exist in the reconstruction of morphology, perhaps even more dramatically than in phonology because of the much larger inventory of morphological elements. Many of the older, well-documented languages, especially Latin, Greek, Baltic, Slavonic and Sanskrit, have very complex morphologies: they have well-developed case systems in nouns, adjectives and pronouns; they have finely marked gender and number categories with fixed concord relations. In the verb they have elaborate systems of tense, voice, mood and aspect, as well as number markers and even gender concord in some forms, all marked with complex morphological formatives.

Many Indo-European languages reflect this complex morphology to one degree or another: Celtic, Armenian and, in part, Tocharian, in addition to the groups just mentioned. But many of the other languages of which we have adequate records show much less morphological complexity, with fewer formal categories and distinctions; and it is not only the modern ones. Hittite, Germanic, Tocharian (in part) and Albanian do not agree with the other groups in morphological complexity.

How does the analyst approach these problems in applying the comparative method? The answer is: cautiously. We must not think of Proto-Indo-European as a single monolithic entity, uniform and dialect-free, which existed at a certain time in a single place before it began to disintegrate. Rather, we must recognise that this language was itself the product of millennia of development. As Ivanov puts it (1965: 51):

> Within the limits of the case systems of the Indo-European languages it is possible to distinguish chronological layers of various epochs beginning with the preinflectional in certain forms of the locative and in compound words … right up to the historical period when the case systems were being formed … Between these two extreme points one must assume a whole series of intermediate points.
>
> (Quoted from Schmalstieg 1980: 46)

We will now proceed to a discussion of the traditional ('classical') system as reconstructed in the nineteenth and early twentieth centuries, and refined in more recent literature. This system represents one, surely very late, stage of Proto-Indo-European from which some, but not all of the daughter languages descended. In this context it has validity as the most probable system based on the comparative method.

3.2.1 Nominal and Pronominal Morphology

Traditionally, Proto-Indo-European is considered to be a fusional (inflectional) language which uses case markers to indicate grammatical relations between nominal elements and other words in a sentence, and to indicate gender and number agreement between words in phrases.

Based primarily on the evidence of Latin, Greek, Sanskrit, Baltic and Slavonic languages and to an extent Armenian, PIE is reconstructed with eight cases: nominative (subject of sentence), genitive (adnominal case), dative (beneficiary), accusative (direct object), ablative (source), locative (place where), instrumental (means and agent) and vocative (direct address). In addition, nouns and adjectives were inflected in three genders (m., f., n.), and three numbers (sg., pl., du.).

The structure of the noun was based on the following scheme: a *root*, which carried the basic lexical meaning, plus a *stem*, which marked morphological class, plus an *ending*, which carried grammatical information based on syntactic function. Thus a word like Lat. *lupus* (OLat. *lupos*) 'wolf' would be *lup+o+s.* Generally we recognise consonantal and vocalic stem nouns. Some examples of consonantal stems are **ped-* 'foot' (Skt. *pád-*, Gk. (gen.) *podós*, Lat. (gen.) *pedis*); **edont-/*dont-/*dent-* 'tooth' (Skt. *dánt-*, Gk. (gen.) *odóntos*, Lat. (gen.) *dentis*); **ĝʰom-* 'man' (Lat. *homo*, Go. *guma*); **māter* 'mother' (Skt. *mātár-*, Gk. *mḗtēr*, Lat. *māter*); **ĝonos-/*ĝenos-* 'race' (Skt. (gen.) *jánasas*, Gk. (gen.) *géneos* (< **génesos*), Lat. (gen.) *generis* (< **genesis*)).

To illustrate some of the vocalic stems we may cite the *i*-stem form **egnis/*ognis* 'fire' (Skt. *agní-*, Lat. *ignis*) or **potis* 'master' (Skt. *páti-*, Gk. *pósis*, Lat. *potis*); an *-eu-* diphthongal stem like **dyeu-* 'sky, light' (Skt. nom. *dyāús*, Gk. *Zeús*, Lat. *diēs*, *-diūs*); and finally the *o*-stem **wl̥kʷos* 'wolf' (Skt. *vŕ̥ka-*, Gk. *lúkos*, Lat. *lupus*).

Through a comparison of the various languages we arrive at the following reconstruction of case endings (Szemerényi 1999: 160; the order of cases as in the original). These endings represent a composite set of possibilities for the Proto-Indo-European noun; no single stem class reflects them all.

Reconstructed Case Endings

	Sg.	*Pl.*	*Du.*
Nom.	-s,-∅	-es	
Voc.	-∅	-es	-e, -ī/-i
Acc.	-m/-m̥	-ns/-n̥s	
Gen.	-es/-o̥s/-s	-om/-o̥m	-ous? -ōs?
Abl.	-es/-os/-s; -ed/-od	-bh(y)os, -mos	-bhyō, -mō
Dat.	-ei	-bh(y)os, -mos	-bhyō, -mō
Loc.	-i	-su	-ou
Inst.	-e/-o, -bhi/-mi	-bhis/-mis, -ōis	-bhyō, -mō

The Proto-Indo-European adjective followed the same declensional pattern as the noun. Adjectives were inflected for gender, number and case, in agreement with the nouns which they modified. Some adjectives are inflected in masculine, feminine and neuter according to masc. *-o* stem, fem. *-ā* stem and neut. *-om* patterns, as in **newos*, **newā*, **newom* 'new' (cf. Skt. *návas*, *návā*, *návam*, Gk. *né(w)os*, *né(w)ā*, *né(w)on*, Lat. *novus*, *nova*, *novum*). Other adjectival forms have identical masculine and feminine forms, but separate neuter (cf. Lat. *facilis*, *facile* 'easy'), and still others have all three identical in some cases (cf. Lat. *ferens* 'carrying' (< **ferentis*)).

Adjectives were compared in three degrees, as in English *tall*, *taller*, *tallest*. Comparative forms are typically derived from positive forms through the suffixation of **-yes*, **-yos* (cf. Lat. *senior* 'older' *(senex)*, Skt. *sánya-* 'older' (*sána-*) and with **-tero-* (cf. Gk. *ponērós* 'wicked', cmpr. *ponēróteros*). Superlatives are often found with the suffixes *-isto-* and *-samo-*, though there are others. Some examples: Gk. *béltistos*, Go. *batista* 'best', Skt. *náviṣṭha-* 'newest' *(náva-)*. For **-samo-*, cf. Lat. *proximus* 'nearest', *maximus* 'greatest', OIr. *nessam* 'next'. As with Gk. *béltistos*, Go. *batista*, adjectival comparison was occasionally carried out with suppletive forms, cf. Lat. *bonus*, *melior*, *optimus* 'good, better, best'.

Proto-Indo-European distinguished many different types of pronouns. A short sample of personal pronouns is given in Table 1.2. Pronouns followed the same general inflectional patterns as nouns, though they have their own set of endings for many of the case forms, except personal pronouns, which are almost entirely different from nouns and did not mark gender. In addition to the personal pronouns 'I/we', 'you/you' *(*eĝ(h)om*, *eĝō/*wei*, **n̥smés*; **tū*, **tu/*yūs*, **usmés*), Proto-Indo-European also had demonstrative pronouns with the form (m.) **so*, (f.) **sā*, (n.) **tod* and **is*, **ī*, **id.* These also served the function of third person pronouns in many of the Indo-European languages. The first of these is represented in Skt. *sa*, *sā*, *tad*, Go. *sa*, *so*, *þata* and Gk. *ho*, *hē*, *tó.* The latter Proto-Indo-European demonstrative forms are represented in Lat. *is*, *ea*, *id* and in various forms in Sanskrit and Germanic such as Skt. nom. sg. n. *id-ám*, acc. sg. m. *im-ám*, f. *im-ám* and Go. acc. sg. *in-a*, nom. pl. m. *eis*, acc. pl. *ins.*

Interrogative and relative pronouns are also well represented, though it is not possible to reconstruct a single relative. From a PIE (anim.) **kʷis*, (inanim.) **kʷid*, which had either interrogative or indefinite meaning, we find Lat. *quis*, *quis*, *quid*, Gk. *tís*, *tís*, *tí*, Hitt. *kwis*, *kwit*, Skt. *kás*, *kā́*, *kím*, and a number of variants of this stem with interrogative or indefinite meaning. In Italic, Tocharian, Hittite, Celtic and Germanic the root **kʷis*, **kʷid* also functioned as a relative pronoun (as does Eng. *who*). In Indo-Iranian, Greek and Slavonic a different form **yos*, **yā*, **yod* served the relative function (cf. Skt. *yás*, *yā́*, *yád*, Gk. *hós*, *hḗ*, *hó*). There is also a recoverable reflexive form **sew-*, **sw* (OCS *sę*, Lat. se, Go. *si-k*).

3.2.2 Verb Morphology

The Proto-Indo-European verb presents the analyst with many of the same issues as the noun. The various daughter languages show wide variation in formal categories and inflectional complexity; some of the ancient classical languages, especially Greek, Latin and Sanskrit, have highly diversified formal structure characterised by intricate relations of tense, mood, voice and aspect. Others, like Hittite and Germanic, have fairly simple morphological systems with few formal distinctions. We can contrast formal complexity by the following simple chart.

Verbal Categories

	Voices	*Moods*	*Tenses*
Greek	3	4	7
Sanskrit	3	4	7
Hittite	2	2	2
Gothic	2	3	2

As with the noun, we may take several paths to a reconstructed system. We can propose a robust Proto-Indo-European system with losses and syncretisms in Hittite and Gothic, we may propose a simple Proto-Indo-European system with additions and splits in Greek and Sanskrit, or we may assume different periods of development and break-off from the parent language. Accepting this final alternative, realistic though it is, in effect prohibits us from reconstructing a single system which underlies the others. All we can do, then, is to present one version, surely quite late, of the Proto-Indo-European verbal system as traditionally reconstructed, recognising that many unanswered chronological questions remain which are outside the scope of this chapter.

The classical reconstruction of the Proto-Indo-European verbal system posits two voices, four moods and from three to six tenses. In addition, there were person and number suffixes and a large number of derivational formatives by which additional categories were fashioned. The verb categories are as follows:

Voice refers to the relationship of the subject to the activity defined by the verb, i.e. whether the subject is agent, patient or both. In Proto-Indo-European there were two voices, active and medio-passive. An active verb is one in which the subject is typically the agent, but is not directly affected by the action (e.g. ***John*** *called Bill*). Medio-passive is a mixed category which includes the function of middle (= reflexive) and passive. When the subject of the verb is both the agent and the patient, the verb is in the middle voice (e.g. Gk. *ho paĩs loúetai* 'the boy washes himself', Skt. *yájate* 'he makes a sacrifice for himself'). When the subject of the verb is the patient, but there is a different agent, the verb is in the passive voice (e.g. Gk. *ho paĩs loúetai hupò tẽs mētrós* 'the boy is washed by his mother'). In general, the various Indo-European languages generalised either the middle or the passive function from the Proto-Indo-European medio-passive. For example, in Sanskrit the middle function dominates, the passive being late and secondary. In Greek the middle and passive are morphologically identical in all but the future and aorist tenses, with the middle dominating. Italic and Celtic have mostly passive use, though there are ample relics of the middle in deponent verbs like Lat. *loquitur* 'he speaks', OIr. *-labrathar* 'who speaks', as well as Lat. *armor* 'I arm myself', Lat. *congregor* 'I gather myself', and others. Germanic has no traces of the middle, while Hittite shows a largely middle function.

Mood describes the manner in which a speaker makes the statement identified by the verb, i.e. whether he believes it is a fact, wishes it, doubts it or orders it. In Proto-Indo-European

there were probably four moods: indicative, optative, subjunctive and imperative. With the indicative mood the speaker expresses statements of fact. Indicative is sometimes marked by a vowel suffix (thematic class) and sometimes not (athematic class), e.g. Skt. *rud-á-ti*, Lat. *rud-e-t* 'he cries' (thematic); Skt. *ás-ti*, Lat. *es-t* 'he is' (athematic). The optative mood is used when the speaker expresses a wish or desire, and is also marked by a vowel which depends on the vowel in the indicative, e.g. OLat. *siet*, Gk. *eíē*, Skt. *syā́t* 'let him be'. The subjunctive is used when the speaker is expressing doubt, exhortation or futurity. Its theme vowel depends on the vowel of the verb in the indicative, though it is common with *e/o* ablaut. Some examples are Lat. *erō* 'I will be', *agam*, *agēs* 'I, you will/might drive', Gk. *íomen* 'let us go'. The final mood is the imperative, which is used when the speaker is issuing a command. The imperative was formed from the bare verbal stem, without a mood-marking vowel as with the other three. Imperatives are most common in the second person, though they are found in the first and third as well. Examples are (second person) Gk. *phére*, Skt. *bhára*, Lat. *fer* 'carry' (sg.) and *phérete*, *bhárata*, *ferte* (pl.). There were other imperative suffixes as well which need not concern us here.

Tense refers to the time of the action identified by the verb. The original Proto-Indo-European verb was probably based on aspectual rather than temporal relations (aspect refers to the type of activity, e.g. momentary, continuous, iterative, etc.), though tense comes to dominate in most IE languages, Slavonic being a prominent exception. We usually identify three tense stems, the present, the aorist and the perfect. The present stem identifies repeated and continuing actions or actions going on in the present (= imperfective aspect): Lat. *sum*, Gk. *eimí*, Skt. *ásmi* 'I am', or Lat. *fert*, Gk. *phérei*, Skt. *bhárati* 'he carries'. The aorist stem (= perfective aspect) marks actions that did or will take place only once, e.g. Gk. *égnōn* 'I recognised', Skt. *ádāt* 'he gave', Gk. *édeikse* 'he showed', Skt. *ánāiṣam* 'I led'. The final stem is the perfect stem (= stative aspect), which describes some state pertaining to the subject of the verb. Examples are Skt. *vḗda*, Gk. *oīda*, Go. *wáit* 'I know'.

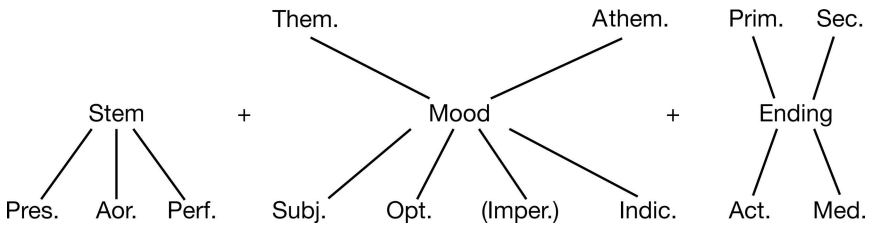
The exact internal structure of the various tense systems is extremely complicated. A number of formal types exist, including stems characterised by ablaut, reduplication, prefixation (augment), infixation and a wide variety of derivational suffixes. An interesting fact is that though tense was not directly and explicitly marked in Proto-Indo-European, most of the daughter languages generalised tense as the defining characteristic of their respective verbal systems.

In addition to the tense, voice and mood categories, the Proto-Indo-European verb carried at the end of the verbal structure a set of endings which indexed first, second or third person and singular, plural or dual number, and also carried much of the information on voice and tense in the daughter languages. There were different sets of endings for different voices, tense stems and moods. Here we list only the principal 'primary' and 'secondary' endings; they are identical except for the final *-i*, an earlier particle which marks the primary endings. These endings were originally used with specific tenses and moods, but have been largely generalised in the daughter languages.

Verbal Endings

	Primary	*Secondary*
1st sg.	-mi (Skt. bhárāmi)	-m (Lat. sum)
2nd sg.	-si (Skt. bhárasi)	-s (OLat. ess)
3rd sg.	-ti (Skt. bhárati)	-t (Lat. est)
3rd pl.	-nti (Skt. bháranti)	-nt (Lat. sunt)

We can schematise the overall structure of the Proto-Indo-European verb as follows:



The Structure of the Indo-European Verb

A few examples:

Lat.		am	ā	s		'you love'	(Pres. indic. 2nd pers. sg. act.)
		am	ē	s		'you might love'	(Pres. subj. 2nd pers. sg. act.)
		am	ā	ris		'you are loved'	(Pres. indic. 2nd pers. sg. pass.)
		am	ē	ris		'you might be loved'	(Pres. subj. 2nd pers. sg. pass.)
Gk.		paideú		ei	s	'you teach'	(Pres. indic. 2nd pers. sg. act.)
		paideú		ē	s	'you might teach'	(Pres. subj. 2nd pers. sg. act.)
		paideú		oi	s	'may you teach'	(Pres. opt. 2nd pers. sg. act.)
Skt.		bhár	a	ti		'he carries'	(Pres. indic. 3rd pers. sg. act.)
	á	neṣ		vahi		'we two led ourselves'	(Aor. indic. 1st pers. du. mid.)
		sunu		yắma		'we might press'	(Pres. opt. 1st pers. pl. act.)

Besides the finite verb forms which we have been discussing, Proto-Indo-European also made use of a number of derivative forms which were nonfinite, i.e. they did not stand as independent tensed predications. We include here a number of infinitive forms, which were originally noun forms in various oblique cases (mostly accusative and dative) and became re-analysed as part of the verbal system: cf. Skt. *dātum* (acc.), *dātavē* (dat.) 'to give'. There were also participial formations represented in most of the languages from Proto-Indo-European formations in **-nt-* (e.g. Go. *bairands*, Skt. *bhára*n-, Lat. *ferens* 'carrying'), as well as others in **-wes-* (cf. Skt. *vidvás-* 'knowing'), **meno-* (cf. Gk. *hepómenos* 'following') and **-to-* (cf. Lat. *amātus* 'loved'). These secondary formations, as well as a number of others such as gerunds, gerundives, supines and other verbal nouns, are widely represented and used throughout the Indo-European family.

3.3 Syntax

The reconstruction of syntax has lagged behind the reconstruction of the phonological, morphological and lexical structures of Proto-Indo-European. This is initially surprising given the central role played by syntax and syntactic theory in modern linguistics. There are many reasons for this lag; among them are the following:

(a) The lack of native speakers. Modern linguistics draws its data from the speech and intuitions of native speakers, but of course a reconstructed language has no such data source.
(b) The abstractness of syntax. Phonological, morphological and lexical units are more concrete units than rules or patterns of syntax. Fewer theoretical notions are

required in order to isolate concrete units, whereas in syntax, nothing exists pre-theoretically. Syntax is a set of principles, requiring abstract theories before even data organisation can begin.

(c) The structure of the descendant languages. The Indo-European daughter languages are of a highly inflecting type, and carry out a great deal of their 'syntax' in morphological expressions. Consider the difference between (1) and (2) in English:

(1) The boy sees the girl.
(2) The girl sees the boy.

The Latin equivalents to these sentences can have the words arranged in any order without affecting the agent/patient relations:

(1′)	i	Puer	puellam	videt.
		Boy	girl	sees
	ii	Puellam	puer	videt.
		Girl	boy	sees
	iii	Videt	puer	puellam.
		Sees	boy	girl
			etc.	
(2′)	i	Puella	puerum	videt.
		Girl	boy	sees
	ii	Puerum	puella	videt.
		Boy	girl	sees
	iii	Videt	puella	puerum.
		Sees	girl	boy
			etc.	

From these few examples we can readily see that the morphology/syntax division in fusional languages is quite a different matter from the same division in a language like English.

(d) The data. The Indo-European languages on which the reconstruction of Proto-Indo-European is based are simply not uniform enough to allow a straightforward account of syntactic patterns. The problem is no greater, and no less, than that found in phonology and morphology.

We will move now to a brief and highly selective review of some major features of Proto-Indo-European syntax. Because the citation of examples is extremely complicated, data will be cited sparingly.

3.3.1 Word Order

Late Proto-Indo-European was most likely a subject–object–verb (SOV) language with attendant adjective–noun (*good boy*), genitive–noun (*John's hat*), standard–marker–adjective (*John than bigger)* order, postpositions (*the world over*) and the preposing of relative clauses (*the who I saw man*). The reconstruction of these structural patterns is based on principles of typological inference developed largely by W. Lehmann (e.g. 1974), who extended the concepts of word order harmony formulated by J. Greenberg

(1963) to historical syntax. According to these principles, there are major structural configurations in languages which are harmonious or compatible with each other. They take the form of statements like the following: if a language has some property P, then it will also have some property Q. For example, if a language is SOV in its basic sentence pattern, it will also have postpositions; if it is SVO, it will have prepositions.

Lehmann has put such 'implicational universals' to work in the reconstruction of Proto-Indo-European word order patterns. For example, Hittite, Vedic Sanskrit and Tocharian are SOV; Latin is predominantly SOV (Homeric Greek is apparently alternately SVO/SOV). Concentrating on Hittite, we find that it has postpositions and adjective–noun order, and dominant genitive–noun and relative clause–noun order. This seems to be ample evidence for an SOV Proto-Indo-European, a conclusion which is strengthened by the existence of SOV-harmonic forms in otherwise SVO languages. For example, we find relic postpositions in Slavonic and Baltic, as well as large numbers of postpositions in the Italic languages which, though they are mostly SOV, are moving towards SVO. The archaic-like nature of the frozen postpositions in Latin *mēcum* 'with me', *tēcum* 'with you' (not **cum mē*, **cum tē*, as expected; cf. *cum puellā* 'with the girl') can be taken as an indication of early SOV structure, even in languages like Latin which show a move towards SVO structures. Discovering such patterns and drawing inferences for reconstruction depends crucially on the assumption that such marked structures as the Latin postpositions are indeed archaisms and not innovations.

There has been much criticism of this approach to syntax. For one thing, it has been noted that inflected languages have much freer word order possibilities than do languages like English, which rely on word order for marking grammatical function. According to this view, the word order issue is a false one, since word order serves mainly secondary functions like marking topic or focus relations

Another problem with the approach is the fact that the pure types are very rare (in the Indo-European family only Celtic is consistent, and it is VSO!). But the method has a built-in escape: languages which are internally inconsistent, like English with its SVO but adjective–noun structures, are said to be in transition from one type to another; the process is not yet complete. This begs the issue, because languages are always in such a transitional state. In other words, Greenberg's observations should be regarded as frequently occurring and interesting tendencies which should not be elevated to the status of explanatory devices. Furthermore, there is ample evidence that such implicational universals do not serve as reliable predictors of future syntactic change in a language.

Finally there are matters of method. Typological inferences often are based on data being used in a circular manner, viz. if a language is SOV, one expects postpositions. And if a language is SVO but a stray postposition is found, one assumes that it must have been SOV at one time. There is also the issue of 'marked' vs 'unmarked' structures. Determining that a language is SOV or SVO when both are present in the data requires a judgement that one of the structures is more natural, more basic, more regular than the other. The problem with ancient languages with no native speakers is that judgements about marked/unmarked structures often reduce to simple frequency counts, and this is not adequate.

3.3.2 Ergative vs Nominative/Accusative Structure

It is clear from the daughter languages that late (classical) Proto-Indo-European was of the nominative/accusative type. That is, the agent of the verb was inflected in the

nominative case, and the patient or goal was inflected in the accusative: cf. Lat. *Marc-us amat puell-am* 'Marcus loves the girl'. But there is evidence that Proto-Indo-European was at one time of the ergative type, i.e. a language in which the subject of a transitive verb is in a different case from the subject of an intransitive verb. There are many instances throughout the early Indo-European languages of agents in the genitive case: cf. OArm. *ēr nora* (gen.) *hraman aṙeal* 'he (of him) had received a promise', Lat. *attonitus serpentis* 'astonished by (of) the serpent'. There are other cases where the real object of a verb of perception is in the accusative while the producer of the perceived act is in the genitive: cf. Skt. *vā́cam* (acc.) *śr̥ṇóti* 'he hears a voice' vs *devásya* (gen.) *śr̥ṇóti* 'he hears a god'. These agentive genitives may at one time have been the subjects of intransitive verbs with genitive agents, as would be found with ergative languages. As Proto-Indo-European developed its complex nominal and verbal morphology, these genitives were reinterpreted as objects of transitive verbs and are now considered simply irregular formations.

3.3.3 Some Further Syntactic Characteristics of Proto-Indo-European

Proto-Indo-European made use of a simple phrase structure principle by which the verb was the only obligatory constituent of a sentence. The subject of the verb was in the nominative, the object in the accusative and a number of other grammatical functions were served by the remaining cases. Verb structures could be expanded with case expressions of time, place-to, place-in, place-from, goal, possession and a number of other qualifiers. Conjunction of both noun phrases and other constituents was possible, including sentence conjunction. Simple sentences could be extended by the use of cases, adverbs and particles to indicate circumstance, purpose, result or manner. Particles were used to introduce different types of clauses (e.g. subordinate, interrogative, relative, co-ordinate). The modality of a sentence, as well as tense and aspect, were expressed inflectionally, though they may have been originally marked only by particles. Finally, there is evidence of a well-developed. noun-compounding system, represented chiefly by Sanskrit.

As a final note to the structure of Proto-Indo-European, it may be useful to take a brief look at a version of a reconstructed Proto-Indo-European sentence. This sentence is from Lehmann and Zgusta's (1979: 462) reinterpretation of Schleicher's famous Indo-European fairy tale, which was written in 1868.

	Owis		ek̂wōskʷe		
	Sheep		horses-and		
Gʷərēi	owis,	kʷesyo	wl̥hnā	ne	ēst
Hill-on	sheep,	of whom	wool	not	is
ek̂wōns	espek̂et	oinom	ghe	gʷr̥um	
horses	he-saw,	one	emph. prt.	heavy	
woĝhom	weĝhontm̥	oinomkʷe	meĝam		
load	pulling,	one-and	great		
bhorom	oinomkʷe	ĝhm̥enm̥	ōk̂u	bherontm̥	
burden	one-and	man	swiftly	carrying.	

The sheep and the horses

On a hill, a sheep which had no wool saw horses, one pulling a heavy load, one carrying a great burden and one (carrying) swiftly a man.

4 Aspects of Proto-Indo-European Culture and Civilisation

When we reconstruct a proto-language, we are by implication also reconstructing (parts of) a proto-culture and civilisation. But linguistic evidence alone is not sufficient to provide a complete picture of a proto-culture; it must be supplemented by information from archaeology, history, folklore, institutions and other sources. The question 'Who were the Indo-Europeans?' has been studied ever since the Indo-European family was established. Where was their homeland, when (if ever) were they a unit, and what was the nature of their culture?

Many different areas of the world have been suggested for the Proto-Indo-European homeland. Central Asia was an early favourite because of the strong Biblical tradition that this was the home of mankind; the Baltic region, Scandinavia, the Finnic area, Western Europe, the Babylonian Empire, southern Russia, the Mediterranean region and a number of other places have been advanced as possibilities. The reason such a wide variety of views exists lies not only in the complexity and ambiguity of the issues, but also in the trends of the times and the prejudices of individual investigators, many of whom were motivated by racial or ethnic considerations rather than scientific investigation. For example, many of the early researchers, lacking the insights of modern anthropology, believed that the obviously strong and warlike Indo-European people could only have been blond, blue-eyed Aryans who must have originated in Northern Europe, and not Asia or the Mediterranean region, for example. Such a confusion of the matters of race, culture and language, fuelled by religious prejudice and scientific immaturity, produced much speculation on the homeland issue.

A famous argument about the homeland was made by Thieme (1953, summarised in 1958). Using the word for 'salmon' *$la\hat{k}s$ (Eng. *lox* < Yiddish *laks*), Thieme argued that these fish fed only in the streams of northern Europe in the Germano-Baltic region during Indo-European times. Since *$la\hat{k}s$ is recoverable with the meaning 'salmon' in Germanic and Baltic and 'fish' in Tocharian, this distribution suggests a northern homeland. In Indo-Iranian a form Skt. *lakṣá* 'one hundred thousand' is interpreted by Thieme as an extension of the uncountable nature of a school of salmon. Thieme concludes that the existence of this root in Indo-Iranian and Tocharian, spoken in areas where salmon are unknown, confirms the Germano-Baltic region as the original homeland.

Thieme uses similar argumentation with the reconstructed words for 'turtle' and 'beech tree'. There is a botanical line where the beech flourished about five thousand years ago, as well as an area which defines the limits of the turtle at the time. Finding these roots in a number of Indo-European languages where the physical objects are unknown suggests the north European region again.

Of course, the problem with such argumentation is that the botanical evidence for the beech line of five thousand years ago is not conclusive. Also, it is well known that speakers frequently transfer old names to new objects in a new environment, as American speakers of English have done with the word *robin.* Thus the root **bhāgo-* may have been used to designate trees other than the beech in some dialects.

This brief review provides us with some background to consider current thinking on the Indo-European homeland. A widely held view is that of M. Gimbutas, who argued in a number of research articles (largely collected in Gimbutas 1997) that the Proto-Indo-European people were the bearers of the so-called Kurgan or Barrow culture found in the Pontic and Volga steppes of southern Russia, east of the Dnieper river, north of the Caucasus, and west of the Ural mountains. The Kurgan culture (from Russian *kurgan*

'burial mound') is typified by the tumuli, round barrows or 'kurgans', which are raised grave structures from the Calcolithic and Early Bronze Age periods. Evidence from the Kurgan archaeological excavations gives clear evidence of animal breeding, and even the physical organisation of houses accords with the reconstructed Proto-Indo-European material. For example, Go. *waddjus* 'wall' is cognate with Skt. *vāya-* 'weaving', which reflects the wattled construction of walls excavated from the Kurgan sites.

Kurgan culture is divided into three periods, beginning in the fifth millennium BCE. The Indo-Europeanisation of the Kurgan culture took place during the Kurgan II period, roughly 4000–3500 BCE. Kurgan sites from this period have been found in the north Pontic region, west of the Black Sea in the Ukraine, Rumania, the former Yugoslavia and Eastern Hungary. During the Kurgan III period (c. 3500–3000 BCE), Kurgan culture spread out across Central Europe, the entire Balkan area and into Transcaucasia, Anatolia and northern Iran. Eventually, it also spread into northern Europe and the upper Danube region. During the final period, Kurgan IV, waves of expansion carried the culture into Greece, West Anatolia and the eastern Mediterranean.

According to Gimbutas, the archaeological evidence attesting to the domesticated horse, the vehicle, habitation patterns, social structure and religion of the Kurgans is in accord with the reconstruction of Proto-Indo-European, which reflects a linguistic community from about 3000 BCE.

For additional discussion of this topic see Mallory and Adams (1997), Fortson (2004). For a proposal to locate the PIE homeland east, in Anatolia at a much earlier date (ca. 7000 BCE), see initially Renfrew (1987) and many subsequent publications.

Salient lexical items which give insight into Proto-Indo-European culture can be cited. In the remaining space we will note those items which are particularly useful in developing a view of Proto-Indo-European culture. Mallory and Adams (1997) provides an encyclopaedic account of PIE cultural vocabulary.

Physical Environment. One or more words for *day*, *night*, the *seasons*, *dawn*, *stars*, *sun*, *moon*, *earth*, *sky*, *snow* and *rain* are plainly recoverable. A number of arboreal units have been identified and successfully reconstructed. Words for *horse*, *mouse*, *bear*, *wolf*, *eagle*, *salmon*, *beaver*, *otter*, *dog*, *cattle*, *sheep*, *pig*, *goat*, *wasp*, *bee* and *louse* can also be reliably postulated. It is interesting that no single word for *river* or *ocean* can be established.

Family Organisation and Social Structure. The Proto-Indo-European family was probably patriarchal and patrilocal, living in small houses and adjacent huts. Villages were small, distant and presumably exogamous. There is good evidence for patriliny, and cross-cousin marriage was probably not permitted. Kinship terms are reconstructible for *father*, *mother*, *brother*, *sister*, *son*, *daughter*, *husband's in-laws* and *probably grand-relatives*. The word for husband means 'master' and the wife was probably 'a woman who learns through marriage'. Evidence for Proto-Indo-European patriarchal kinship comes not only from the lexicon, but also from epic songs, legal tracts and ethnological sources from the various ancient Indo-European languages.

There is widespread evidence of a word for *tribal king*, giving some indication that government was established.

Technology. The Indo-European languages confirm the technological advancements of the proto-culture. Evidence from farming and agricultural terms indicates small-scale farmers and husbandmen who raised pigs, knew barley, and had words for *grain*, *sowing*,

ploughing, *grinding*, *settlement* and *field* or *pasture*. We can also safely reconstruct words for *arrow*, *axe*, *boat*, *gold*, *wagon*, *axle*, *hub* and *yoke*, showing a rather advanced people with knowledge of worked metals and agriculture.

Religion and Law. From lexical, legal and other sources we find clear indications of a religious system among the Proto-Indo-European people. There is a word for *god*, and a designation for a *priest*; words for *worship*, *prayer*, *praise* and *sacred* give clear indications of organised religion. There is lexical evidence and evidence from ancient institutions for legal concepts such as *religious law*, *pledge*, *justice* and *compensation*.

Bibliography

General overviews of the Proto-Indo-European and the Indo-European languages include Ramat and Ramat (1998) and Fortson (2004). Meillet (1937) remains a lucid exposition of the principles of Indo-European linguistics, while Szemerényi (1999) is currently the most authoritative handbook. For a general text on historical linguistics and its methods, see Campbell (2004).

Pokorny (1951–59) is the standard in Indo-European etymology and lexicography, while Buck (1951) is a resource of synonyms arranged by semantic class. More recent accounts are Mallory and Adams (1997), Watkins (2000), Rix (2001), and a number of on-line resources.

References

Brugmann, K. 1903. *Kurze vergleichende Grammatik der indogermanischen Sprachen* (Trübner, Strassburg)

Buck, C.D. 1951. *A Dictionary of Selected Synonyms of the Principal Indo-European Languages* (University of Chicago Press, Chicago)

Campbell, L. 2004. *Historical Linguistics: an Introduction*, 2nd edn (MIT Press, Cambridge)

Fortson, B.W., IV. 2004. *Indo-European Language and Culture*: *An Introduction* (Blackwell, Oxford)

Gamkrelidze, T.V. 1981. 'Language Typology and Language Universals and Their Implications for the Reconstruction of the Indo-European Stop System', in Y.L. Arbeitman and A.R. Bomhard (eds) *Bono Homini Donum: Essays in Historical Linguistics in Memory of J. Alexander Kerns* (John Benjamins, Amsterdam), pp. 571–609

Gamkrelidze, T.V. and Ivanov, V.V. 1995. *Indo-European and the Indo-Europeans*. (Part 1, The text; part 2, Bibliography, indices.), Trends in Linguistics: Studies and Monographs, 80 (Mouton de Gruyter, Berlin and New York). [English version by J. Nichols of *Indoevropejskij jazyk i indoevropejcy*. 1984 (Tbilisi University Press, Tbilisi)]

Gimbutas, M. 1997. *The Kurgan Culture and the Indo-Europeanization of Europe: Selected Papers from 1952 to 1993* (Institute for the Study of Man, Washington)

Hopper, P. 1981. '"Decem" and "Taihun" Languages: An Indo-European Isogloss', in Y.L. Arbeitman and A.R. Bomhard (eds) *Bono Homini Donum: Essays in Historical Linguistics in Memory of J. Alexander Kerns* (John Benjamins, Amsterdam), pp. 133–42

Ivanov, V.V. 1965. *Obăceidoevropejskaja praslavjanskaja i anatolijskaja jazykovye sistemy* (Nauka, Moscow)

Kuryłowicz, J. 1927. 'ə indo-européen et h hittite', in *Symbolae Grammaticae in Honorem Ioannis Rozwadowski* (Drukarnia Uniwersytetu Jagiellońskiego, Cracow), pp. 95–104

Lehmann, W.P. 1952. *Proto-Indo-European Phonology* (University of Texas Press, Austin)

Lehmann, W.P. and Zgusta, L. 1979. 'Schleicher's Tale After a Century', in Bela Brogyanyi (ed.), *Festschrift for Oswald Szemerényi on the Occasion of his 65th Birthday* (John Benjamins, Amsterdam), pp. 455–66

Mallory, J.P. and Adams, D. (eds) 1997. *The Encyclopedia of Indo-European Culture* (Fitzroy Dearborn, London)

Meillet, A. 1937. *Introduction à l'étude comparative des langues indo-européennes*, 8th edn (reprinted by University of Alabama Press, University, Alabama, 1964)

Pokorny, J. 1951–9. *Indogermanisches etymologisches Wörterbuch* (Francke, Bern and Munich)

Ramat, P. and Ramat, A. 1998. *The Indo-European Languages* (Routledge, London)

Renfrew, C. 1987. *Archaeology and Language: The Puzzle of Indo-European Origins* (Cape, London)

Rix, H. 2001. *Lexicon der indogermanischen Verben (LIV)* (Reichert, Wiesbaden)

Saussure, F. de. 1879. *Mémoire sur le système primitif des langues indo-européennes* (Teubner, Leipzig)

Schleicher, A. 1876. *Compendium der vergleichenden Grammatik der indogermanischen Sprachen* (Böhlau, Weimar)

Schmalstieg, W.R. 1980. *Indo-European Linguistics: A New Synthesis* (Pennsylvania State University Press, University Park)

Szemerényi, O. 1999. *Introduction to Indo-European Linguistics* (Oxford University Press, Oxford) [English translation based on 4th German edn, 1990]

Thieme, P. 1953. 'Die Heimat der indogermanischen Gemeinsprache', in *Abhandlungen der geistes- und sozialwissenschaftlichen Klasse* (Akademie der Wissenschaften und Literatur, Wiesbaden), pp. 535–610

—— 1958. 'The Indo-European Language', *Scientific American*, vol. 199, no. 4, pp. 63–74

Watkins, C. 2000. *The American Heritage Dictionary of Indo-European Roots*, 2nd edn (Houghton Mifflin, Boston)

2

Germanic Languages

John A. Hawkins

The Germanic languages currently spoken fall into two major groups: North Germanic (or Scandinavian) and West Germanic. The former group comprises: Danish, Norwegian (i.e. both the Dano-Norwegian Bokmål and Nynorsk), Swedish, Icelandic, and Faroese. The latter: English (in all its varieties), German (in all its varieties, including Yiddish and Pennsylvania German), Dutch (including Afrikaans and Flemish) and Frisian. The varieties of English are particularly extensive and include not just the dialectal and regional variants of the British Isles, North America, Australasia, India and Africa, but also numerous English-based pidgins and creoles of the Atlantic (e.g. Jamaican Creole and Pidgin Krio) and the Pacific (e.g. Hawaiian Pidgin and Tok Pisin). When one adds to this list the regions of the globe in which Scandinavian, German and Dutch are spoken, the geographical distribution of the Germanic languages is more extensive than that of any other group of languages. In every continent there are countries in which a modern Germanic language (primarily English) is extensively used or has some official status (as a national or regional language). Demographically there are at least 450 million speakers of Germanic languages in the world today, divided as follows: North Germanic, over 18 million (Danish over 5 million, Norwegian over 4 million, Swedish approximately 8.8 million, Icelandic 260,000 and Faroese 47,000); West Germanic apart from English, approximately 125 million (90 million for German in European countries in which it has official status, German worldwide perhaps 100 million, Dutch and Afrikaans 25 million, Frisian over 400,000); English worldwide, 320–80 million first language users, plus 300–500 million users in countries like India and Singapore in which English has official status (cf. Crystal 2003).

There is a third group of languages within the Germanic family that needs to be recognised: East Germanic, all of whose members are now extinct. These were the languages of the Goths, the Burgundians, the Vandals, the Gepids and other tribes originating in Scandinavia that migrated south occupying numerous regions in western and eastern Europe (and even North Africa) in the early centuries of the present era. The only extensive records we have are from a fourth-century Bible translation into Gothic. The Goths had migrated from southern Sweden around the year nought into the area around what is now Gdańsk (originally Gothiscandza). After AD 200 they moved

south into what is now Bulgaria, and later split up into two groups, Visigoths and Ostrogoths. The Visigoths established new kingdoms in southern France and Spain (AD 419–711), and the Ostrogoths in Italy (up till AD 555). These tribes were subsequently to become absorbed in the local populations, but in addition to the Bible translation they have left behind numerous linguistic relics in the form of place names (e.g. *Catalonia*, originally 'Gothislandia'), personal names (e.g. *Rodrigo* and *Fernando*, compare Modern German *Roderich* and *Ferdinand*), numerous loanwords (e.g. Italian-Spanish *guerra* 'war'), and also more structural features (such as the Germanic stress system, see below). In addition, a form of Gothic was still spoken on the Crimean peninsula as late as the eighteenth century. Eighty-six words of Crimean Gothic were recorded by a Flemish diplomat in 1562, who recognised the correspondence between these words and his own West Germanic cognates.

The earliest records that we have for all three groups of Germanic languages are illustrated in Figure 2.1. These are runic inscriptions dating back to the third century AD and written (or rather carved in stone, bone or wood) in a special runic alphabet referred to as the Futhark. This stage of the language is sometimes called Late Common Germanic since it exhibits minimal dialect differentiation throughout the Germanic-speaking area. Further evidence of early Germanic comes from words cited by the classical writers such as Tacitus (e.g. *rūna* 'rune') and from some extremely early Germanic loanwords borrowed by the neighbouring Baltic languages and Finnish (e.g. Finnish *kuningas* 'king'). The runic inscriptions, these early citations and loans, the Gothic evidence and the method of comparative reconstruction applied to both Germanic and Indo-European as a whole provide us with such knowledge as we have of the Germanic parent language, Proto-Germanic.

There is much uncertainty surrounding the origin and nature of the speakers of Proto-Germanic, and even more uncertainty about the speakers of Proto-Indo-European. It seems to be agreed, however, that a Germanic-speaking people occupied an area comprising what is now southern Sweden, southern Norway, Denmark and the lower Elbe at some point prior to 1000 BC, and that an expansion then took place both to the north

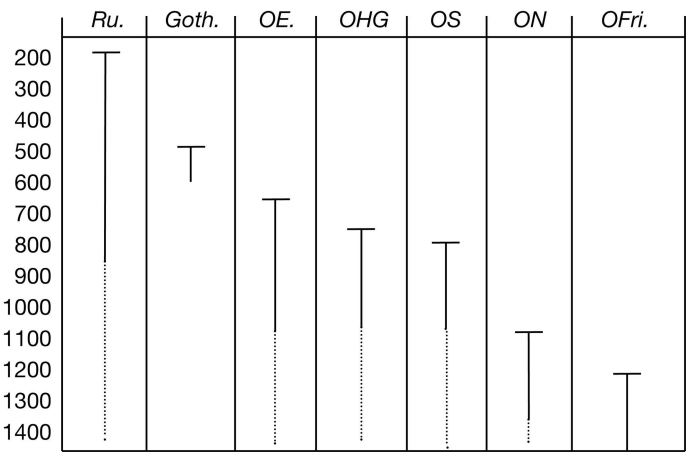


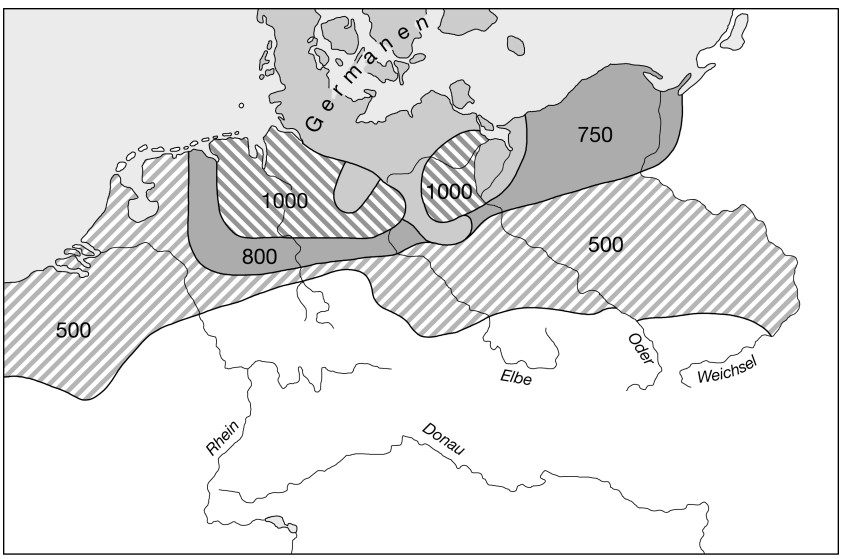
Figure 2.1 The Earliest Written Records in the Germanic Languages.
Source: Kufner 1972.

and to the south. Map 2.1 illustrates the southward expansion of the Germanic peoples in the period 1000 to 500 BC. A reconstruction of the events before 1000 BC is rather speculative and depends on one's theory of the 'Urheimat' (or original homeland) of the Indo-European speakers themselves (see pages 25–26). The pre-Germanic speakers must have migrated to their southern Scandinavian location sometime before 1000 BC and according to one theory (cf. Hutterer 1975) they encountered there a non-Indo-European-speaking people from whom linguistic features were borrowed that were to have a substantial impact on the development of Proto-Germanic from Proto-Indo-European. According to Hutterer as much as one-third of the vocabulary of the Germanic languages is not of Indo-European origin (see page 56).

The major changes that set off Proto-Germanic from Proto-Indo-European are generally considered to have been completed by at least 500 BC. In the phonology these were the following: the First (or Germanic) Sound Shift; several vowel shifts; changes in word-level stress patterns; and reductions and losses in unstressed syllables.

The First Sound Shift affected *all* the non-nasal stops of Proto-Indo-European and is illustrated in Figure 2.2.

The reconstructed Proto-Indo-European consonants of Figure 2.2 are those of Brugmann (1903) (see Baldi, this volume, page 11). According to this reconstruction Proto-Indo-European had a voiceless and a voiced series of consonants, each of which could be unaspirated or aspirated, and within each series there was a bilabial, a dental, a palatal, a velar and a labio-velar (labialised velar) stop, as shown. Proto-Germanic abandoned the palatal/velar distinction throughout, and collapsed the unaspirated and aspirated series of voiceless stops. Unaspirated voiced stops shifted to their voiceless counterparts (see, for example, Lat. ***d****ecem*, Eng. ***t****en*), voiceless stops shifted to voiceless fricatives (e.g. Lat. ***t****res*, Eng. ***th****ree*), and aspirated voiced stops shifted to voiced fricatives (most of which subsequently became voiced stops). The dotted lines in Figure 2.2 indicate the operation of what



Map 2.1 Expansion of the Germanic People 1000–500 BC.
Source: Adapted from Hutterer 1975.

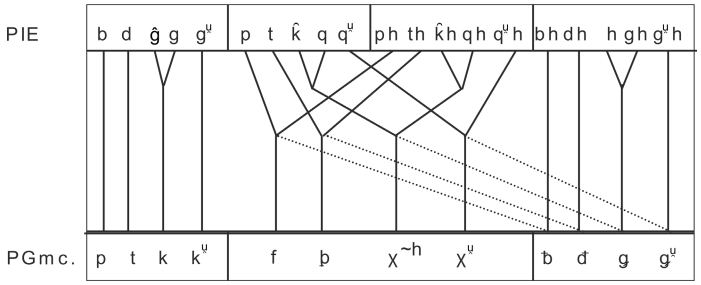


Figure 2.2 The First (Germanic) Sound Shift.
Source: Adapted from Krahe 1948.

is called 'Verner's Law'. Depending on the syllable that received primary word stress, the voiceless fricatives of Germanic would either remain voiceless or become voiced. For example, an immediately following stressed syllable would induce voicing, cf. Go. *fa**d**ar* 'father' pronounced with [ð] rather than [θ], from PIE **pətér*, cf. Skt. *pi**t**ár-*, Gk. *pat**ḗ**r*.

According to the more recent Proto-Indo-European consonantal reconstruction of Gamkrelidze (1981) (see Baldi, this volume, page 14) the unaspirated voiced stops of Figure 2.2 were actually glottalised stops, which lost their glottalic feature in Proto-Germanic, resulting in the voiceless stops shown. For further details, and also a critique, of this reconstruction see Voyles (1992).

The vowel shifts are illustrated in Figure 2.3. Short *a*, *o* and ə in Proto-Indo-European were collapsed into Germanic *a* (compare Lat. ***a**ger*, Go. ***a**krs* 'field, acre'; Lat. ***o**cto* (PIE ***ok̂tō***), Go. ***a**htau* 'eight'; PIE *pəter*, Go. *fa**d**ar* 'father'). The syllabic liquids and nasals of Proto-Indo-European became *u* plus a liquid or nasal consonant. Long *ā* and *ō* collapsed into *ō* (Lat. *fr**ā**ter*, Go. *br**ō**þ*ar* 'brother'; Lat. *fl**ō**s* (PIE **bhl**ō**men*), Go. *bl**ō**ma*, 'flower, bloom'), and the number of diphthongs was reduced as shown.

The changes in word stress resulted in the many word-initial primary stress patterns of the Germanic languages where in Proto-Indo-European the stress had fallen on a variety of syllable types (the root, word- and stem-forming affixes, even inflectional endings). This shift (from a Proto-Indo-European accentual system that has been argued to be based on pitch originally, i.e. high versus low tones) is commonly assumed to have occurred after the First Sound Shift, since the operation of Verner's Law presupposes variable accentual patterns of the Indo-European type that were subsequently neutralised by the reassignment of primary stress. Thus, both PIE **bhrā́ter* 'brother' and **pətér* 'father' end up with primary stress on the initial syllable in Go. *brṓþar* and *fádar*, and yet the alternation between voiceless [θ] in the former case and voiced [ð] in the latter bears testimony to earlier accentual patterns. Had the stress shifted first, both words should have changed *t* in the same way. A major and lasting consequence of initial stress was the corresponding reduction and loss of unstressed syllables. This process was well underway in predialectal Germanic and was to continue after the separation of the dialects. Indo-European final *-t* was regularly dropped (Lat. *velit*, Go. *wili* 'he will/wants'), and final *-m* was either dropped or reduced to *-n* (OLat. *quom*, Eng. *when*). Final short vowels were dropped (Gk. *oĩda* 'I see', Go. *wait* 'I know'), and final long vowels were reduced in length.

The extremely rich morphology of Proto-Indo-European was reduced in Proto-Germanic. The Proto-Indo-European noun distinguished three genders (masculine, feminine, neuter),

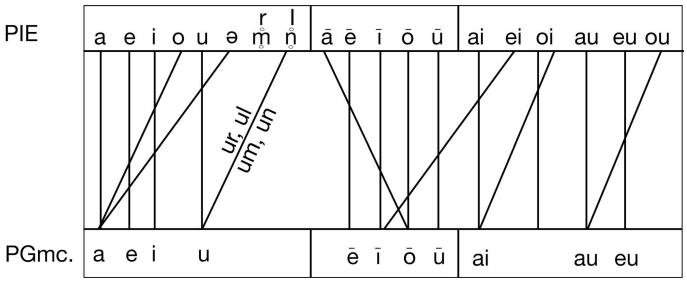


Figure 2.3 Germanic Vowel Shifts.
Source: Krahe and Meid 1969.

three numbers (singular, plural, dual) and eight cases (nominative, vocative, accusative, genitive, dative, ablative, instrumental and locative). The three genders were preserved in Germanic, but special dual inflections disappeared (though residual dual forms survive in the pronominal system of the early dialects). The eight cases were reduced to four: the original nominative, accusative, and genitive preserved their forms and functions; the vocative was collapsed with the nominative; the dative, instrumental and locative (and to some extent the ablative) were united in a single case, the Germanic dative, though occasional instrumental forms are attested; and some uses of the ablative were taken over by the genitive.

Proto-Indo-European nouns were also divided into numerous declensional classes depending on the final vowel or consonant of the stem syllable, each with partially different inflectional paradigms. These paradigms survive in Germanic, though some gained, and were to continue to gain, members at the expense of others (particularly the PIE *o*-class (Gmc. *a*-class) for masculine and neuter nouns, and the PIE *ā*-class (Gmc. *ō*-class) for feminine nouns). The inflectional paradigm for masculine *a*-stems in the earliest Germanic languages is illustrated in Table 2.1.

The syncretism of the case system was accompanied by an expansion in the use of prepositions in order to disambiguate semantic distinctions that had been carried more clearly by the morphology hitherto.

The pronouns of Germanic correspond by and large to those of Indo-European, except for the reduction in the number of dual forms.

As regards the adjective, Germanic innovated a functionally productive distinction between 'strong' and 'weak' inflections, which is still found in Modern German (cf. pages 97–98 for illustration). Proto-Indo-European adjectival morphology was fundamentally similar to that for nouns. The Germanic strong adjective inflections were formed from a fusion of pronominal inflections with the declensional paradigm for nouns and adjectives ending in a stem vowel, while the weak adjective inflections were those of nouns and adjectives with *n*-stems. Strong and weak adjectives in the early dialects carried a meaning difference similar to that of the indefinite versus definite articles of the modern Germanic languages, and it is no accident that adjectives within indefinite versus definite noun phrases are typically strong and weak respectively in German today.

Proto-Indo-European verbal morphology was considerably reduced in Germanic. The Proto-Indo-European medio-passive voice was lost (except for a few relics in Gothic and Old English), and only the active survives. Distinct subjunctive and optative forms were

Table 2.1 The Inflectional Paradigm for Germanic Masculine *a*-Stems

		Go.	*ON*	*OE*	*OS*	*OHG*
Sg.	Nom.	dags	dagr	dæg	dag	tag
	Gen.	dagis	dags	dæges	dages	tages
	Dat.	daga	dege	dæge	dage	tage
	Acc.	dag	dag	dæg	dag	tag
	Voc.	dag	(= Nom.)	(= Nom.)	(= Nom.)	(= Nom.)
	Inst.	—	—	dæge	dagu	tagu
Pl.	Nom.	dagōs	dagar	dagas	dagos	taga
	Gen.	dagē	daga	daga	dago	tago
	Dat.	dagam	dǫgom	dagum	dagum	tagum
	Acc.	dagans	daga	dagas	dagos	taga

Notes: Germanic *a*-stems exemplified by Gothic *dags* 'day' and cognates in the other Germanic dialects derive from Indo-European *o*-stems (cf. Latin *lupus,* earlier *lupos* 'wolf').

collapsed, and only two of several tense and aspect distinctions were maintained in the Germanic present versus past tenses. Separate verb agreement inflections for dual subjects survive only (partially) in Gothic and Old Norse. A special innovation of Germanic involved the development of a systematic distinction between strong and weak verbs. The former (exemplified by Eng. *sing/sang/sung*) exploit vowel alternations, or 'ablaut' (see pages 14–15), in distinguishing, for example, past from present tense forms, the latter use a suffix containing a dental element without any vowel alternation (e.g. Eng. *love/loved*). The verbal morphology of Proto-Germanic has been maintained in all the modern Germanic languages (though the number of strong verbs has been reduced in favour of weak ones), and in addition new periphrastic forms have evolved for the tenses (e.g. perfect and pluperfect) and voices (the passive) that were lost in the transmission from Proto-Indo-European to Proto-Germanic.

The Germanic lexicon, like the phonology and morphology, reveals clearly the Indo-European origin of Germanic. Yet, as pointed out earlier, Hutterer (1975) argues that as much as a third of Germanic lexical items cannot be derived from Proto-Indo-European. These items, far from being peripheral, belong to the core of the basic vocabulary of Common Germanic. They predominate in the following semantic fields: seafaring terms; terms for warfare and weaponry; animal names (particularly fish) and terms for hunting and farming; communal activities and social institutions and titles; and miscellaneous terms. Some examples (taken from English) are: *sea*, *keel*, *boat*, *rudder*, *mast*, *steer*, *sail*; *sword*, *bow*; *carp*, *eel*, *calf*, *lamb*; *thing* (originally a communal meeting), *king*, *knight*; and *leap*, *bone*. In the absence of independent evidence for the Germanic substrate language, arguments for lexical borrowing, or for other distinctive features of Germanic from the substrate must be considered speculative. More recent studies of early Germanic such as Voyles (1992) and Robinson (1992) do not refer to it. On the other hand, the Dutch dictionary of Marlies Philippa et al. (2003) gives systematic attention to the substrate idea, and unless Indo-European etymologies can be found for these basic vocabulary items in Germanic it must be considered a serious possibility worthy of further research.

Common Germanic also took numerous loanwords from neighbouring Indo-European peoples, especially from Latin, though also from Celtic. The Latin loans reveal the strong influence of Roman culture on the early Germanic peoples in areas such as agriculture

(cf. Eng. *cherry*/Lat. *ceresia*, *plum/pluma*, *plant/planta*, *cheese/caseus*), building and construction (*street/strata*, *wall/vallum*, *chamber/camera*), trade (*pound/pondo*, *fishmonger/mango* (= slave-trader), *mint/moneta*), warfare (*camp/campus*). Most of the days of the week are loan translations from the Latin (e.g. *Sunday/solis dies*, etc.).

There is much less certainty about the syntax of Proto-Germanic, though the word order of the earliest inscriptions (Late Common Germanic) has been quite extensively documented by Smith (1971). He establishes that the basic position of the verb was clause-final (62 per cent of the clauses he investigated were verb-final, with 19 per cent verb-second and 16 per cent verb-first). Within the noun phrase, however, the predominant order of adjectival modifiers and of possessive and demonstrative determiners is after the noun, and not before it, as in many OV languages. In the earliest West Germanic dialects, by contrast, the verb is correspondingly less verb-final, and modifiers of the noun are predominantly preposed.

The precise manner in which the proto-language split up into the three groups (North, East and West) is a question of long-standing dispute. With the exception of the earliest runic inscriptions, the tripartite division is already very clearly established in the earliest records of Figure 2.1: each of the groups has undergone enough characteristic innovations to justify both the existence of the group itself and the assumption of a period of separate linguistic development for the languages involved following migration from the homeland. But whether these innovations point to the existence of, for instance, a West Germanic parent language which split off from Proto-Germanic and from which all the later West Germanic dialects are descended, or whether the innovations are the result of contact and borrowing between geographically proximate tribes speaking increasingly distinct dialects whose common point of departure was the Germanic parent language, is almost impossible to tell. Some scholars argue against the assumption of a West Germanic parent language on the grounds that a threefold dialect grouping within West Germanic (into North Sea Germanic, Rhine-Weser Germanic, and Elbe Germanic – also called respectively Istveonic, Ingveonic and Erminonic) can be reconstructed back as early as the second century AD. The runic inscriptions of this early period do not lend credence to such an early dialect split, however.

Bibliography

For the Indo-European background, see Baldi (Chapter 1, this volume), Brugmann (1903), Krahe (1948), and Gamkrelidze (1981). Van Coetsem and Kufner (1972) contains many papers (in English) on the phonology, morphology and syntax of Proto-Germanic, on the position of Germanic within Indo-European as a whole and on the reconstruction of developments within Germanic prior to the first records. It includes Kufner's (1972) summary and synthesis of the different theories concerning subgroupings within Germanic.

For the phonology and morphology of early Germanic languages, see Krahe and Meid (1969), Voyles (1992) and Robinson (1992). Robinson also includes discussion of syntax. Hutterer (1975) is a general compendium of the grammars and histories of all the Germanic languages and of the cultures of their speakers. Smith (1971) gives a summary of word order in early Germanic.

The chapters in König and van der Auwera (1994) give grammatical summaries (in English) of all the modern Germanic languages, including Germanic creoles (Romaine 1994). This volume also includes an overview chapter on the Germanic languages (Henriksen and van der Auwera 1994), a chapter on Gothic and the reconstruction of Proto-Germanic (W.P. Lehmann 1994), plus chapters on the historical stages of North and West Germanic languages.

References

Baldi, P., this volume. 'Indo-European Languages'

Brugmann, K. 1903. *Kurze vergleichende Grammatik der indogermanischen Sprachen* (Trübner, Strassburg)

Crystal, D. 2003. *English as a Global Language*, 2nd edn (Cambridge University Press, Cambridge)

Gamkrelidze, T.V. 1981. 'Language Typology and Language Universals and Their Implications for the Reconstruction of the Indo-European Stop System', in Y.L. Arbeitman and A.R. Bomhard (eds), *Bono Homini Donum: Essays in Historical Linguistics in Memory of J. Alexander Kerns* (John Benjamins, Amsterdam), pp. 571–609

Henriksen, C. and van der Auwera, J. 1994. 'The Germanic Languages', in E. König and J. van der Auwera (eds) *The Germanic Languages* (Routledge, London), pp.1–18

Hutterer, C.J. 1975. *Die germanischen Sprachen: ihre Geschichte in Grundzügen* (Akadémiai Kiadó, Budapest)

König, E. and van der Auwera, J. (eds) 1994. *The Germanic Languages* (Routledge, London)

Krahe, H. 1948. *Indogermanische Sprachwissenschaft* (Walter de Gruyter, Berlin)

—— and Meid, W. 1969. *Germanische Sprachwissenschaft*, 2 vols (Walter de Gruyter, Berlin)

Kufner, H.L. 1972. 'The Grouping and Separation of the Germanic Languages', in F. Van Coetsem and H.L. Kufner (eds) *Toward a Grammar of Proto-Germanic* (Max Niemeyer, Tübingen)

Lehmann, W.P. 1994. 'Gothic and the Reconstruction of Proto-Germanic', in E. König and J. van der Auwera (eds) *The Germanic Languages* (Routledge, London), pp. 19–37

Philippa, M., Debrabandere, F. and Quak, A. (eds) 2003. *Etymologisch Woordenboek van het Nederlands*, Vol. 1 (Amsterdam University Press, Amsterdam)

Robinson, O.W. 1992. *Old English and its Closest Relatives: A Survey of the Earliest Germanic Languages* (Stanford University Press, Stanford)

Romaine, S. 1994. 'Germanic Creoles', in E. König and J. van der Auwera (eds) *The Germanic Languages* (Routledge, London), pp. 566–603

Smith, J.R. 1971. 'Word Order in the Older Germanic Dialects' (PhD dissertation, University of Illinois, available from University Microfilms, Ann Arbor, Mich.)

Van Coetsem, F. and Kufner, H.L. (eds) 1972. *Toward a Grammar of Proto-Germanic* (Max Niemeyer, Tübingen)

Voyles, J.B. 1992. *Early Germanic Grammar: Pre-, Proto-, and Post-Germanic Languages* (Academic Press, San Diego)

3

English

Edward Finegan

1 Introduction

Readers of *The World's Major Languages* may be assumed to have more than a nodding acquaintance with the English language. English is, moreover, widely studied and has received significant attention from distinguished grammarians since the nineteenth century. It thus seems appropriate here to discuss English in terms not entirely parallel to those in which other languages, perhaps less familiar, are described in this book. In somewhat more detail than is possible at present for most other languages, this chapter will describe the structural variation that characterises English functionally and socially, as well as some salient historical and regional variation.

Section 2 describes the status of English throughout the world, along with its social history and past contact with other languages. Section 3 offers a historical sketch of its lexicon, phonology, morphology and syntax, followed by a brief account of orthographic practices. Section 4 treats regional, social and functional variation in present-day English.

2 Status of English

2.1 Current Status of English

Although Chinese is spoken by a greater number of people, English is spoken around the globe with a wider dispersion than any other language. From its earlier home in Britain (now with 60 million speakers), it has spread to nearby Ireland (4 million), across the Atlantic to North America (where some 215 million residents over the age of five speak it in the United States and as many as 20 million in Canada) and across the world to Australia and New Zealand (with more than 20 million speakers).

English is the sole official language in more than a score of other countries: Ghana, Liberia, Nigeria, Uganda and Zimbabwe in Africa; Jamaica, the Bahamas, Dominica and Barbados in the Caribbean; and the Solomon Islands in the Pacific, to name a sample. Elsewhere it shares official status with one or more languages in a score of

nations, including Tonga, Tanzania, Cameroon, Kenya, Nigeria, South Africa, Singapore, the Philippines, Western Samoa, Kiribati, Pakistan and India (where it is an associate official language alongside Hindi). In still other nations, English holds no official status only because its widespread use in government (often alongside an indigenous tongue) and in trade is taken for granted. The United States, with no designated official language, is the most prominent of such cases. In still other nations, English is of such commercial and scientific advantage that consideration is repeatedly given to making it an official language.

Substantial portions of the populations of the United States and Canada speak English as a second language, many of them immigrants, but others born there and raised in families and neighbourhoods struggling to preserve the language and culture of ancestral lands. One well known example is that of French speakers in Canada, who constitute a majority only in Quebec province but whose influence is so strong nationally that Canada is officially a bilingual nation. Less well known is how many speakers of languages other than English reside in the United States. The United States Census reports that in 2003 more than 48 million residents over the age of five spoke a language other than English at home. This suggests that a good many residents of the United States speak English as a second language. Los Angeles is a sufficiently bilingual city that balloting materials for all elections are available in Spanish, Chinese, Japanese, Vietnamese, Korean and Tagalog, and similar multilingualism characterises voting materials in other American cities. Elsewhere, Nigeria, Ghana and Uganda have almost two million speakers of English as a second language each and the Philippines more than 11 million. Likewise, the millions who speak English in Pakistan and India have learned it, for the most part, not in their infancy but as a second language, a lingua franca for governmental and educational functions.

Beyond its uses as a first and second language in ordinary intercourse, English is the lingua franca of much scholarship, particularly of a scientific and technical nature. In addition, throughout the world are English-speaking universities in which instruction and textbooks use English as a principal medium. Reflective of its widespread dissemination and perhaps its adaptability is the fact that Nobel Prizes in literature have been awarded to more writers using English than any other language and, in particular, that these laureates have been citizens of Australia, Ireland, India, Trinidad, St. Lucia, Nigeria and South Africa, as well as the United States and Britain. At the United Nations, English is an official language, alongside Arabic, Chinese, French, Russian and Spanish. A quarter of the world's population is said to be competent in English—that would be 1.5 billion people (Crystal 2003).

2.2 Possible Reasons for Widespread Use of English

The widespread use of English around the globe is often attributed to social prestige and the need for English in technological advancement. Some also credit the simplicity of its inflections and the cosmopolitan character of its vocabulary. These latter features are influential only when coupled with social, historical and economic factors, for other languages and other peoples share them, though with different outcomes.

Among reasons sometimes suggested for the extension of English is the spread of technology, for the diffusion of American technologies during the twentieth and twenty-first centuries likewise spread English words for them, as with *television* and *HDTV* (high definition television). Likewise in other arenas, cultural artifacts have spread their English

names in their travels, from concrete objects like *jeans* and *discos* (both of which are originally French) to the intangible and ubiquitous *OK*.

English words have not always been welcomed abroad. Troubling such watchdog institutions as the *Académie française*, Anglo-Americanisms like *weekend* and *drug-store* have been banned in France, while guardians of German balked at the introduction of words like *Telefon* for the native *Fernsprecher*. Elsewhere, people are more open to English loanwords. The Japanese, for example, have drafted the words *beesubooru* 'baseball', *booringu* 'bowling' and *futtobooru* 'football', along with the games they name, trading them for *judo*, *jujitsu* and *karate*, which have joined the English team.

Further contributing to the popularity of English may be its inflectional structure, for it exhibits notable inflectional simplicity compared to languages like German and Russian. Assuming, as many linguists would, that a language simple at one level will be compensatorily complex elsewhere in order to carry out equivalent communicative tasks, it is difficult to assess the impetus of grammatical simplicity on the spread of a language. To be sure, English inflections are few and relatively easy to learn compared to heavily inflected languages. English nouns, for example, generally have only two variants in speech, a marked variant for possessive singulars and all plurals, an unmarked one for all other functions. Aside from a few exceptions like *teeth* and *oxen*, plurals are formed by adding /-s/ or /-z/ or /-əz/ to the singular, according to straightforward conditions detailed below (Section 3.1.1). As for possessives, the rules are identical to those for the plural, but without exceptions. Further, English exhibits no variation in adjectives for number, gender, or case, there being but one form each in the positive degree (*tall*, *beautiful*, *old*) although comparative (*-er*) and superlative (*-est*) degrees are marked by inflection or, under specified circumstances, analytical forms with *more* and *most.* Verbs are only minimally inflected, with suffixes for third person singular concord, present participle (*-ing*), past tense (/-t/, /-d/, /-əd/) and past participle. In all, there are but eight productive inflectional suffixes in present-day English: two on nouns, four on verbs and two for adjectives. There are no inflectional prefixes or infixes.

Breadth of vocabulary is an oft-cited reason for the acceptance of English around the globe. *Webster's Third New International Dictionary* (1961) boasts that it contains some 450,000 words. A four-volume supplement to the *Oxford English Dictionary* was completed in 1986, updating the original with words and senses that had arisen or been recognised during the decades of publication between 1884 and 1928 and afterward, and the entries in these supplements are incorporated into the twenty-volume second edition; a third edition in progress can be engaged online. At the *Oxford English Dictionary* website, lists of new entries, subentries and senses appear regularly. A list posted in 2007 displays the nouns *wiki*, *technopreneur*, *tighty-whities*, *tweener*, *lightstick*, *irritainment*, *HDTV*, *edamame*, *bad girl* and *asswipe*; the verbs *cannonball*, *dog-pile* and *virtualize*; and the adjectives *caramelized*, *cardiothoracic*, *fricking* and *trepidatious*. Not all these words are new to English, but even this small selection illustrates the vitality of shortening, compounding, borrowing, affixation and functional shift as processes continuing to expand the word stock.

Still further evidence of the abundance of English words can be seen in the fact that the number of synonyms or near synonyms for many words is large, each reflecting some variation on a semantic core or use in different situations. A thesaurus can provide scores of synonyms for the adjective *inebriated* and more than a dozen for the noun *courtesy*, to offer only two examples from different parts of speech (and without intending to suggest the relative richness of these notions among English speakers).

English also boasts a distinctively cosmopolitan vocabulary, having borrowed extensively from its Germanic cousins and Latin and French, but absorbing tens of thousands of words from scores of languages over the centuries. From earliest times English has exerted a remarkable magnetism for loanwords, not only in foods and toponyms but also in virtually every other arena of human activity. Some indication of this cosmopolitan nature is suggested by words like *alcove*, *alcohol* and *harem* (from Arabic), *tycoon* and *ikebana* (Japanese), *taboo* (Tongan), some 10,000 words of French origin added during Middle English and an even larger influx from Latin during the Renaissance. Recent borrowings reveal an extraordinary range of donor languages, more than seventy-five in number. French provides most items, followed by Japanese, Spanish, Italian, Latin, Greek, German, Yiddish, Russian, Chinese, Arabic, more than two dozen African and three dozen other languages around the globe.

Maps of the English-speaking parts of the world are dotted with borrowings from many sources. Hundreds of streets in Los Angeles exhibit names of Spanish origin (from *La Cienega* to *Los Feliz*) as does the city itself. Elsewhere in the USA, place-names like *Mississippi* and *Minnesota* come from Amerindian languages, while *Kinderhook*, *Schuylerville* and *Watervliet*, all in New York State, come from Dutch. In England, common place name designations come from the Scandinavian languages (as discussed in Section 2.3) and after the Norman invasion from French.

Names for popular foods such as *taco*, *burrito*, *chili* and *guacamole* (from Mexican Spanish), *hamburger*, *frankfurter* and *wiener schnitzel* (German), *teriyaki* and *sukiyaki* (Japanese), *chow mein* and *foo yong* (Cantonese), *kimchi* (Korean), *pilaf* (Persian and Turkish), *falafel* (Arabic) and a thousand others indicate the catholic tastes of the English tongue both gustatorily and linguistically. Playing a special role, French culinary words have leavened the English lexicon in kitchens around the world: *hors d'œuvre*, *quiche*, *pâté*, *fondue*, *flambé*, *soufflé*, *sauté*, *carrot*, *mayonnaise*, *bouillon*, *flan*, *casserole*, a whole series with *crème* including *crème brûlée*, *crème caramel*, *crème de menthe* and *crème de cacao*, and now stretching beyond the dining room are such mainstays as *à la* 'in the manner of' (*à la mode* and *à la carte*) and *crème de la crème* 'the best (of anything)'. A wide stripe of languages is represented by other familiar culinary words: *semolina* (Italian), *chocolate* (Nahuatl, via Spanish and French), *coleslaw* (Dutch), *chutney* (Hindi), *moussaka* (Greek), *bamboo* (Malay), *gazpacho* (Spanish), *yoghurt* (Turkish), *kebab* (Arabic), *caviar* (Persian, via Turkish, Italian and French), *pepper* (Latin), *whiskey* (Irish), *maize* (Taino—an Arawakan language—via Spanish) and *blintz* and *knish* (both Ukrainian, via Yiddish).

Another suggested reason for the spread of English is the simplicity of its common words. In the 100-million-word British National Corpus all of the fifty most frequent words in both speech and writing are monosyllabic. Of the fifty next most frequent written words, none contains more than two syllables and only eleven are disyllabic (Leech et al. 2001). Similar information for a wide range of languages might make it clear that, in accordance with Zipf's law, languages generally abbreviate words of frequent use. English has had an additional historical impetus in that most disyllabic words ending in an unstressed syllable became monosyllabic in early Modern English, as described below (Section 3.2).

One final explanation offered by some scholars for the diffusion of English lies in the supposed nature of the relationships between grammatical structures and the processing mechanisms for comprehension. Though not universally accepted, this explanation relies on the claim that SVO languages like English are perceptually simpler than

languages whose basic orders are SOV or VSO. Proponents of this view point out that, even granted their sociological and political statuses, it is noteworthy that Chinese, French, Russian and Spanish are SVO and of wide diffusion, as is the spoken form of Arabic that is spreading. The putative perceptual advantage of SVO languages over SOV or VSO languages is the ready identification of subjects and objects by virtue of their being separated by the verb. It might also be mentioned that English tends to have topics in sentence-initial position (though to a lesser degree than many other languages); given its preference for SVO word order, subject and topic often coincide, a coincidence that may enhance processability, especially when the subject is the semantic agent.

2.3 English and its Social History

English did not always hold so prominent a position among the world's languages as it holds today. Even in England it has faced competitors at times. Nor has it always been clear that the United States and Canada would be English-speaking countries, and encroachments by Spanish and French on the status of English in North America remain vigorous.

English derives from the West Germanic branch of the Indo-European family of languages. It is most closely related to the Low German dialects in northern Germany and to Dutch and Frisian, sharing with them the characteristic absence of the Second, or High, German Sound Shift, occurring around AD 600 and markedly differentiating the phonology of the West Germanic varieties of the highland south from those of the lowland north. Geographically separated from the Continent since the middle of the fifth century, English would not have been subject to this shift, but its origins in the northernmost part of the Germanic-speaking area would also have spared it.

It was in AD 449, according to Bede's *Ecclesiastical History of the English People*, that bands from the three Germanic tribes of Angles (after whom England and its language were named), Saxons and Jutes began leaving the area known today as northern Holland and Germany and southern Denmark. These Teutons sailed to Britain, which had been deserted by the Romans four decades earlier, to assist the Celtic leader Vortigern, who had called upon them to help repulse the invading Picts and Scots from the north of Britain. Preferring Britain to their continental homelands, the Teutons settled, driving the hapless Celts into remote corners, where their descendants remain to this day.

Surviving the Roman occupation of the British Isles there remain but few linguistic relics of Latin origin, including the second element of such place names as *Lancaster*, *Manchester* and *Rochester* (from Latin *castra* 'camp'). This influence of Latin through Celtic transmission was the slightest of several Latin influences on the English lexicon. As for direct Celtic influence on the early Germanic settlers, it is noticeable only in place-names like *Dover*, *Kent*, *York*, possibly *London*, and a few other toponymics like the river names *Avon*, *Thames* and *Trent*. In 563 St Columba established an Irish monastery on the island of Iona off the coast of Scotland, and his missionary activities introduced a few Celtic words like *cross* and perhaps *curse* into the English word stock.

It is not until the end of the seventh century that we have written records of a Germanic language spoken in England and not until the reign of King Alfred (871–99) that we have 'Englisc' recorded in quantity. In 597, St Augustine (not the bishop of Hippo known for his fifth-century *Confessions* and *City of God*) christianised the English people, giving them scores of Latin words like *abbot*, *altar*, *angel*, *cleric*, *priest* and *psalm* in the religious sphere and others like *grammatical*, *master*, *meter*, *school* and *verse* in learned arenas.

In the eighth and ninth centuries, a series of invasions by the Scandinavians brought a secondary Germanic influence into the Anglo-Saxon lexicon, though it does not vigorously manifest itself in the written record until after the eleventh century. Sporadic raids started in 787, with monasteries sacked and pillaged at Lindisfarne and Jarrow (Bede's monastery). In the year 850, as many as 350 ships carried Danish invaders up the Thames. At length King Alfred defeated these Vikings in 878 and signed the Treaty of Wedmore with Guthrum, who agreed to become Christian. There followed a period of integration during which bilingualism prevailed in the Danelaw, an area governed by Danish practices and including Northumbria, East Anglia and half of central England.

The intermingling of these groups brought an influx of more than 900 everyday words from the Scandinavian tongues, including such homely nouns as *gift*, *egg*, *skirt*, *skill*, *skin* and *sky* and the verbs *take*, *give* and *get*. In addition, about 1,400 Scandinavian placenames pepper English maps, besides some 600 ending in *-by* (as in *Derby*, *Rugby*), 600 in *-thorp* or *-thwaite* (*Kettlethorpe*) and another hundred or so in *-toft* (*Lowestoft*), all Scandinavian. Besides this toponymic evidence, the close relationship between the Scandinavians and the English is suggested by the possibility that both pronoun and verb in the phrase *they are* derive not from OE *hīe sindon* but from Scandinavian sources.

In the development of English, the most significant historical event is the invasion by the Normans in 1066. In that year William, Duke of Normandy, crossed the Channel and with his French-speaking retinues established an Anglo-Norman kingdom in England. During the following century and a half, one could not have confidently predicted the reemergence of English and its eventual triumph over French in all domains. Only a series of extraordinary social events contributed definitively to reestablishing a Germanic tongue emblematic of England.

After 1066 the Normans established themselves in the court, in the church and her monasteries, throughout the legal system and the military and in all other arenas of wealth and power. The upper class spoke only French, while English remained chiefly on peasant tongues. Naturally, between those social extremes a significant number of bilinguals eventually used English and French, but for generations England was ruled by French-speaking monarchs, unable to understand the language of many subjects and unable to be understood. Only when King John lost Normandy to King Philip of France in 1204 did the knot between England and the Anglo-Norman language start to come undone. Following other political and military antagonisms, the linguistic tide turned.

Finally, the Black Death struck England in 1348, wiping out perhaps 30 per cent of the population and increasing the value of every peasant life. Ironically, this plague lifted the English-speaking lower classes to positions of greater appreciation and enhanced the value of their work. Along with a rise in their stature came increased stature for their language. In 1362 Parliament passed the Statute of Pleading, which mandated that all court proceedings, conducted solely in French since the Norman conquest, should thenceforth use English. By about 1300 all the inhabitants of England knew English, and French began to fall into disuse. During the fourteenth century, English again became the language of England and her literature. (Details of this story are well told in Baugh and Cable (2002).)

Literature in English is known since Old English times. *Beowulf*, a 3,000-line heroic poem, is still studied even in secondary schools. The surviving manuscript dates probably from the late tenth century, but the poem was likely composed in the eighth. Other texts also survive: poetry (starting at the end of the seventh century), translations of the Bible, chronicles and religious writings particularly from the time of Alfred. Besides

known translations, including Boethius's *Consolation of Philosophy*, Alfred is thought to have translated Bede's *History* from Latin and is credited with establishing the practice of maintaining the Anglo-Saxon Chronicles. Reigning from Wessex, his kingdom lay within the West Saxon dialect area, making West Saxon the basis for the study of Old English even though it is not the ancestor of the London dialect of Chaucer that is the basis of modern standard English.

Less was written between 1066 and the thirteenth century, but English language traditions remained vital enough for the fourteenth century to produce Chaucer (1340–1400) and his *Canterbury Tales*, an extraordinary work still enjoyed for its earthy, humorous narrative and poetic achievement. From quite early times English has been robust in its literary manifestation, except for the period of Anglo-Norman dominance from which it nevertheless emerged a great literary language, lexically enriched and inflectionally simplified.

3 English Structure and Its History

English is usually divided into three major periods: Old English, dating from either the arrival of the Germanic tribes in 449 or the earliest documents, about 700, to about 1100 (shortly after the Norman conquest); Middle English from about 1100 to 1500; and, from 1500, Modern English, including an early Modern English period between 1500 and 1700. These dates are somewhat arbitrary in that English did not develop at the same rate in all regions nor at all levels of the grammar. The dates are in fact more appropriate to a phonological than a grammatical history because Modern English morphology and syntax displayed essentially their current form by about 1400, the year of Chaucer's death.

Old English had four principal dialect areas: Northumbrian, Mercian, Kentish and, representing most extant texts, West Saxon. In Middle English, Mercian is divided into West Midland and East Midland dialects, and East Midland, which incorporated features of other dialects, gave rise to standard Modern English. In the discussion to follow, little detail is provided for Middle English because it represents a transitional period whose general nature can be inferred from knowledge of Old English and Modern English; because while spoken English remained vital, written documents are relatively scarce; and because Middle English is far more diverse in its regional dialects than is susceptible to a brief exposition. (Details about Middle English can be traced in Mossé (1968).)

3.1 Lexicon

Enriched by compounding of native elements and by borrowing from other languages, the English word stock has grown continuously although the chief mechanisms for enriching it have shifted in the course of time. The Old English lexicon shows traces of Latin and Celtic influence but is almost purely Germanic. To a great extent it shares etymons with the other Germanic languages and like them developed its word stock chiefly by compounding, as well as by prefixing and suffixing. Compounds were especially frequent and imaginative in Old English poetry, and the resulting kennings enhanced poetic resources as in these examples from *Beowulf*: *seġlrād* 'sail road' and *hrōnrād* 'whale road' for sea and *bānhūs* 'bone house' for body. Old English nouns productively suffixed *-dōm*, *-hād*, *-ere* and *-scipe* (all with reflexes in Modern English), as in *wīsdōm* 'wisdom', *cildhād* 'childhood', *wrītere* 'writer' and *frēondscipe* 'friendship'.

Verbs commonly prefixed *ā-*, *be-*, *for-*, *fore-*, *ge-*, *mis-*, *of-*, *ofer-*, *on-*, *tō-*, *un-*, *under-* and *wiþ-*. From *settan* 'to set' Old English could create: *āsettan* 'place', *besettan* 'appoint', *forsettan* 'obstruct', *foresettan* 'place before', *gesettan* 'people, garrison', *ofsettan* 'afflict', *onsettan* 'oppress', *tōsettan* 'dispose', *unsettan* 'put down' and *wiþsettan* 'resist' (Baugh and Cable 2002). It prefixed *wiþ-* to 50 verbs, only one of which (*withstand*) survives in Modern English (*withdraw* and *withhold* originated in Middle English).

The Norman invasion gave new impetus to linguistic borrowing and when English reemerged in the thirteenth century it did so in a context in which anybody who was anybody spoke French and many of the elite spoke little or no English. From that period on, besides smithing with native elements, English has energetically imported words from the languages with which its speakers came into contact. Forty per cent of all French words in English were borrowed between 1250 and 1400 (according to Baugh and Cable (2002)), a period during which English came again to be used for official and learned purposes. From this flood of 10,000 French words inundating Middle English, 75 per cent remain in use. English had earlier borrowed from the Celtic tongues and Latin and, during the ninth and tenth centuries, from its Viking cousins, as we saw. Still, one effect of the invasion was to promote borrowing above the more characteristic English word-smithing practices of affixing and compounding, which formerly had been the most productive springs of new words and would become so again in the twentieth century.

Until recently, it would have been difficult to describe the size and character of the English lexicon accurately, but the availability of standard computerised corpora has changed that. The data presented here rely on three corpora. The Standard Corpus of Present-day Edited American English (the Brown Corpus) comprises 500 text samples, of about 2,000 words each, representing 15 genres of informational and imaginative prose that appeared in print in 1961—about a million words all told. One hundred times that size, the British National Corpus (BNC) contains about 94 million words of British writing and 6 million of transcribed British speech. The Longman Spoken and Written English Corpus comprises about 40 million words in four genres. (See for the Brown Corpus Kučera and Francis (1967) and Francis and Kučera (1982), on which, along with Kučera (1982), we rely in this chapter; for the BNC, Leech et al. (2001) and the references there; for the Longman Corpus, Biber et al. (1999).)

The Brown Corpus contains 61,805 different word forms belonging to 37,851 lemmas. A lemma is a set of word forms, all of which are inflectional or spelling variants of the same base word; thus, the lemma GET comprises the word forms *get* (and *git*), *gets*, *got* (and *gotta*), *gotten, getting* (and *gettin'*). Extrapolating these figures to an infinite sample would yield about 170,000 lemmas in English, excluding proper nouns and highly specialised and technical terms. Remarkably, just 2,124 lemmas (comprising 2,854 word forms) constitute 80 per cent of all tokens in the Brown Corpus. Approximately 22,000 other word forms occur just once each; such *hapax legomena* thus account for 58 per cent of all lemmas. This fact gives some hint as to the range of the lexicon, for the most frequently occurring words are grammatical (i.e., function) words, not lexical (or content) words (cf. Sections 4.3 and 4.4). Because content words are the least predictable textual elements, knowing the 2,124 lemmas that account for 80 per cent of the corpus tokens would fall far short of sustaining comprehension that approximated 80 per cent (Kučera 1982).

The British National Corpus contains 757,087 different word forms, 52 per cent of which (397,041 word forms, according to Leech et al. 2001) occur just once. In any

large body of English texts, then, it appears that most words will be used only once or very few times, while relatively few words will be repeated numerous times. Only about 124,000 word forms occur 10 times or more. For example, the BNC's most common lexical verbs are *say* (3,344 occurrences per million words of running text), *get*, *make*, *go*, *see*, *know*, *take*, *think*, *come*, *give*, *look*, *use*, *find*, *want* and *tell* (775); the most common nouns, *time* (1,833), *year*, *people*, *way*, *man* (1,003), *day*, *thing*, *child* (710), *government*, *work*, *life* and *woman* (631); the most common adjectives, *other* (1,336), *good*, *new*, *old*, *great*, *high*, *small*, *different*, *large*, *local*, *social*, *important*, *long*, *young* and *national* (375); and the most common adverbs, *so* (1893), *up*, *then*, *out*, *now*, *only*, *just*, *more*, *also*, *very*, *well*, *how*, *down*, *back*, *on*, *there*, *still*, *even* and *too* (701). Frequency lists for prepositions, pronouns, determiners, conjunctions, interjections and discourse markers are also provided in Leech et al. (2001). Strictly speaking, while such quantitative findings are valid only for the corpus on which they are based, the broad outlines and most details are likely to be much the same for comparable corpora. (The distribution of word classes across genres is described in Sections 4.3 and 4.4.)

3.2 Phonology

Throughout its history, English has exhibited striking instability in its system of vowels, while its consonants have remained relatively stable since the fourteenth century. Old English, Middle English and Modern English all exhibit considerable vocalic variation across dialects, while consonants show negligible variation from region to region. Socially significant variation, on the other hand, affects both consonants and vowels, as described in Section 4.2.

The evolution of unstressed vowels has played a pivotal role in the development of English morphology and grammar. The most pregnant phonological feature of the earliest stages of English is the characteristic Germanic stress placement on the first or root syllable. From before the settlement of England, the language of the Angles, Saxons and Jutes suffered certain phonological reductions that differentiate it from High German (e.g. loss of nasals preceding /f/, /θ/, /s/, with compensatory lengthening of the preceding vowel; compare German *Mund* and *Gans* with English *mouth* and *goose*). Such correspondences between the stressed vowels of High and Low German only begin to suggest the wholesale reductions that were to affect English unstressed vowels and consequently the entire inflectional system.

While Gothic (known to us from several centuries earlier than Old English) apparently preserves both long and short vowels in its inflections, Old English exhibits only short vowels there, and syncretism among these inflections is apparent in late Old English, especially in the Northumbrian dialect. While early Old English had a relatively elaborate inflectional system, the characteristic Germanic stress placement began to effect reductions of such magnitude in unstressed vowels that inflectional suffixes in late Old English and Middle English were reduced essentially to the bare system of Modern English. In particular, unstressed /u/, /a/, /e/ and /o/ fell together into *e* [ə]. Coupled with the merging of final /-m/ and /-n/ in /-n/, the collapse of unstressed vowels and subsequent loss of final inflectional /-n/ and of final [ə] led to the virtual elimination of inflectional suffixes except those with final *-s* or *-þ*. This sequence of phonological levellings explains the plural and genitive forms of Modern English nouns, as well as third person singular verbs in orthographic *-s* and past tenses in *-d*.

As to stressed vowels, their history is complicated by the substantial dialectal variation of Old English and the shifting locus of literary standards until the fifteenth century. Still, the extensive diphthongisation and monophthongisation that characterise Old English recur throughout the history of English. When American southerners pronounce *ride* as [ra:d], they evidence the same kind of monophthongisation that took place in late Old English when *sēon* became *seen* 'to see' and *heorte* became *herte* 'heart'.

Today, some fourteen to sixteen phonemic vowels and diphthongs exist in the regional varieties of standard English, including /aj/, /aw/ and /ɔj/, the last of which was borrowed from Anglo-Norman. (A more detailed treatment of stressed vowels is available by period in Algeo and Pyles (2004) and by sound in Kurath (1964).) No discussion of English historical phonology can ignore the dramatic shifting of long vowels that occurred mostly between Chaucer's death in 1400 and the birth of Shakespeare in 1564. This so-called Great Vowel Shift altered the pronunciation of the long vowels and diphthongised the high vowels /i:/ and /u:/ to their Modern English reflexes /aj/ and /aw/. Charted in Figure 3.1, this shift is responsible for the discrepancy between English and the Romance languages in the pronunciation of orthographic vowels. Traditional English spellings were propagated with Caxton's introduction of printing into England in 1476, preceding the completion of the shift.

Subsequent to this vowel shift, early ModE /e:/ (< ME /ɛ:/) came to be pronounced /i/, thus merging with earlier raised /e:/ and producing two sets of ModE /i/ words, those like *sweet* and *see* from OE /e:/ and those like *sheaf*, *beacon* and *sea* from OE /ɛ:/. The raising tendency exhibited in the Great Vowel Shift continues today, where it is sometimes regionally distinctive and sometimes socially marked, as discussed in Section 4.1.

As to consonants, the system has remained relatively stable throughout history, and the inventory of phonemes has changed only slightly since about 1400, although certain allophones have been lost and phonotactic constraints have altered somewhat. The Modern English spelling of *know* and *knife* is indicative of earlier pronunciations in that Old English allowed initial clusters of /hl-/ as in *hlāf* 'loaf', /hr-/ as in *hring* 'ring' and /kn-/ as in *cniht* 'knight', all of which are now prohibited.

Table 3.1 contains a list of Modern English consonant phonemes, followed by exemplars illustrating word-initial, word-medial and word-final occurrence.

Several differences between the consonant systems of Old English and Modern English can be mentioned. The members of the ModE voiced and voiceless fricative

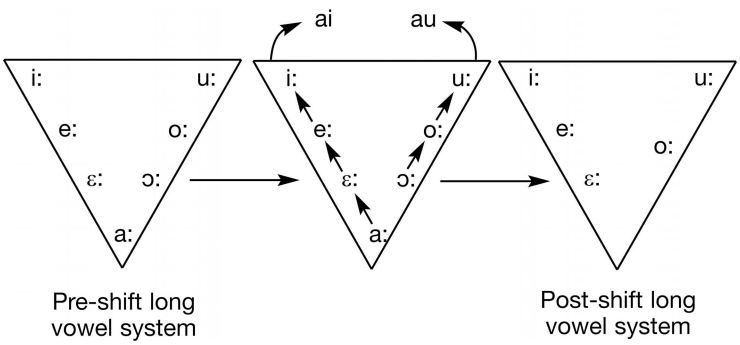


Figure 3.1 The English Vowel Shift (also called Great Vowel Shift).
Source: Bynon 1977.

pairs (/f/-/v/, /θ/-/ð/, /s/-/z/) were allophones of single phonemes in OE, the voiced sounds occurring between other voiced segments, the voiceless sounds occurring initially, finally and in clusters with voiceless obstruents. Relics of the OE allophonic distribution remain in the morphophonemic alternants *wife/wives*, *breath/breathe* and *house/houses*, where the second word in each pair, disyllabic in OE, voiced the intervocalic fricative. Significantly, initial /ð/ is limited in Modern English to the function words *the*, *this*, *that*, *these*, *those*, *they* and *them*, *there* and *then*, *thus*, *thence*, *though* and *thither*, with initial voiceless /θ/ in Old English later becoming voiced by assimilation when unstressed, as these words often are. Similarly, /θ/ does not occur medially in any native words, though it can be found in borrowings. During the Middle English period, with the baring of the voiced phones word-finally once the syncreted inflections disappeared, the allophones achieved phonemic status, contrasting in most environments; there may also have been some Anglo-Norman influence.

Modern English /n/ and /ŋ/, non-distinctive variants in Old English, became phonemic during late Middle English or early Modern English. Of the OE allophones of /h/, both [x] and [ç] have disappeared, leaving only [h]. In addition, /g/ had two OE allophones, ModE [g] and a fricative [ɣ] occurring intervocalically but now lost.

Finally, the gap existing in Old English and Middle English where one might expect a voiced palatal fricative /ʒ/ to parallel native /ʃ/ was filled about 1600 when /ʒ/ arose by assimilation of /zj/ from earlier /zi/ (as in *glazier*, *lesion* and *vision*) and /ziu/ (as in *measure* and *usual*). More recently, word-final /ʒ/ has been borrowed directly from French in words like *mirage*, *prestige* and *rouge* (though /ʤ/ is also often heard). /ʒ/ is the only

Table 3.1 Modern English Consonants

Phoneme	*Initial*	*Medial*	*Final*
p	pat	caper	tap
b	bat	labour	tab
t	tap	button	bat
d	dad	ladder	pad
k	cad	sicker	talk
g	gab	dagger	gag
f	file	beefy	thief
v	vile	saving	crave
θ	thin	author	breath
ð	then	weather	breathe
s	sin	mason	kiss
z	zebra	posit	pose
ʃ(š)	shame	lashes	push
ʒ(ž)		measure	rouge
ʧ(č)	chin	kitchen	pitch
ʤ(j)	jury	bludgeon	fudge
m	moon	dummy	room
n	noon	sunny	spoon
ŋ		singer	sing
h	hen	ahoy	
j(y)	year	beyond	
r	red	berry	deer
l	lot	silly	mill
w	wind	away	

Modern English consonant not fully native to the English inventory, for the /zj/ cluster from which it arose entered the language mainly in French and Latin loanwords. (This alien sound also developed in American English in words like *Asia(n)*, *emersion* and *version*, where British English has /ʃ/; /ʒ/ sometimes also occurs in American English in words like *transients* and phrases like *as yet* and *all these years*.) The uneven distribution in the pattern of Modern English consonants apparent in the list above reflects the historical development of these sounds.

3.3 Morphology

Old English morphology is considerably more complex than that of Middle English and Modern English. As a consequence of the phonological reductions and mergers described in Section 3.2, extensive syncretism of the OE distinctive inflections occurred, and the inflectional morphology of ModE exhibits only eight inflections. Only pronouns preserve anything resembling their OE complexity; adjectives and the definite article preserve the least. Here we describe the OE and ModE pronominal and adjectival systems, with briefer discussions of nouns, verbs and the definite article.

3.3.1 Nouns

	a-stems		*o-stems*	*athematics*
Singular	*M.*	*Nt.*	*F.*	*M.*
Nominative	stān	dēor	lār	fōt
Accusative	stān	dēor	lāre	fōt
Genitive	stānes	dēores	lāre	fōtes
Dative	stāne	dēore	lāre	fēt
Plural				
Nom./Acc.	stānas	dēor	lāra	fēt
Genitive	stāna	dēora	lāra	fōta
Dative	stānum	dēorum	lārum	fōtum

Old English had several noun declensions, both strong (from Indo-European vowel stems) and weak (from Indo-European consonant, i.e. *-n*, stems). Each noun carried a grammatical gender irrespective of its natural gender (although nouns with human referents show a notable fit between grammatical and natural gender). Nouns were inflected generally for three or four cases in the singular, three in the plural; the nominative and accusative are identical in the plural and also in most singular nouns; paradigms for *stān* 'stone', *dēor* 'animal', *lār* 'learning' and *fōt* 'foot' are illustrative.

From the *stān* declension come the productive Modern English genitive singular in *-s* and all the productive plurals, while the *fōt* declension has yielded the few nouns (like *foot*, *goose* and *tooth*; *louse* and *mouse*; and *man*) whose plurals, generalised from the nominative and accusative, exploit a functional vowel alternation instead of the common suffix in *-s*. This palatal mutation was caused by earlier assimilation of the stem vowels to suffixes. ModE relic phrases like 'a ten-foot pole' derive from the OE genitive plural (translated roughly 'a pole of ten feet'), whose form *fōta* has the reflex *foot*. From the *dēor* declension come ModE uninflected plurals like *deer* and *sheep*.

Other noun declensions in Old English showed variations according to the phonological characteristics of the stems at various periods in their development (and there was considerable dialectal variation). In Modern English the only productive forms of the

genitive singular and of the plural are the reduced reflexes of the masculine *a*-stems, like which many older nouns have been analogically reformed and all new nouns are inflected. The plural and genitive morphemes have the same phonologically conditioned allomorphs, which by dissimilation have /-əz/ after stems ending in /s, z, ʃ, ʒ, ʧ, ʤ/, and by assimilation /-s/ after stems ending in other voiceless consonants and /-z/ after voiced segments. The plural and genitive morphemes exhibit syncretism (as in *boys'*) except when the plural noun is marked by a stem change, as in *women's, children's, geese's.* The genitive marker can be attached to such complex nouns as *his father in law's criticism* and *the president of the university's right to consultation* and even to more complex phrases such as *the geek I told you about's blog, one of the buddies I went with's coworkers* and *the person you sent it to's computer screen*.

3.3.2 Verbs

Like the other Germanic languages, Old English and its reflexes exhibit two types of verbs, called strong and weak by Jakob Grimm. While weak verbs (characteristically Germanic) exhibit a dental suffix (/d/ or /t/) in the preterit, strong verbs show an internal vowel change (characteristically Indo-European ablaut). OE had seven classes of strong verbs, with scattered reflexes surviving today, though starting even in OE many strong verbs became weak, while others were reformed analogically. Many OE strong verbs have developed regular ModE forms, with past tense and past participle suffixes in /-t/ or /-d/. Listed here are the principal parts (infinitive, past singular, past plural and past participle) for each OE verb class.

I	rīdan 'ride'	rād/ridon	geriden
II	sēoðan 'boil'	sēað/sudon	gesoden
III	bindan 'bind'	band/bundon	gebunden
IV	beran 'bear'	bær/bǣron	geboren
V	giefan 'give'	geaf/gēafon	gegiefen
VI	standan 'stand'	stōd/stōdon	gestanden
VII	feallan 'fall'	fēoll/fēollon	gefeallen

From these principal parts can be formed the two tenses (present and preterit) in the indicative and subjunctive moods. A typical weak verb conjugation is provided here, making it apparent that while the present indicative exhibits three singular forms and one plural, the subjunctive contrasts only singular and plural forms. The twelve distinct forms of an OE weak verb have been reduced to four in ModE.

		Indicative	*Subjunctive*
Pres.	Sg. 1	dēme	
	2	dēmst/dēmest	dēme
	3	dēmþ/dēmeþ	
	Pl.	dēmaþ	dēmen
Pret.	Sg. 1	dēmde	
	2	dēmdest	dēmde
	3	dēmde	
	Pl.	dēmdon	dēmden
Gerund		tō dēmenne/dēmanne	
Pres. Part.		dēmende	
Past Part.		dēmed	

The OE verbal system is inflectionally a mere shadow of its Indo-European predecessors. By way of contrast, recall that Latin is inflected for active and passive voice, perfective and imperfective aspects and present, preterit and future tenses, as well as for several moods. On the other hand, the OE verb is a far cry from the periphrastic complexity of its modern counterpart in tense and aspect.

3.3.3 Articles

Though its pragmatically determined use is complicated, the Modern English definite article has only the single orthographic shape *the*, with standard pronunciations [ði] before vowels and [ðə] elsewhere. As shown here, the Old English demonstrative (forerunner of today's definite article) was formally complex, inflected in the singular for five cases (including the instrumental) in the masculine but with fewer distinctive case forms in the feminine and neuter. Neither the neuter singular nor the gender-neutral plural forms differentiated between nominative and accusative case forms.

	Singular			*Plural*
	M.	*F.*	*Nt.*	*M./F./Nt.*
Nom.	sē	sēo	þæt	þā
Acc.	þone	þā	þæt	þā
Gen.	þæs	þǣre	þæs	þāra
Dat.	þǣm	þǣre	þǣm	þǣm
Inst.	þȳ	þǣre	þȳ	

The initial consonant of the nominative masculine singular *sē* and feminine singular *sēo* differed from all other forms, which begin with /θ/, orthographic *þ-*. Thus, ModE *the* has no direct OE etymon but arose analogically from forms with initial /θ/, influenced by parallel Scandinavian forms introduced in the late eighth and the ninth centuries. By the Middle English period *þe* had become the invariant definite article in the north, whence it soon spread to all dialects. Chaucer uses only *the.* (Customary lack of stress on *the* fostered assimilation of the initial voiceless segment to the vocalic nucleus.)

Such a history is somewhat surprising for what is by far the most common word in Modern English. *The* occurs twice as often as its nearest competitor (*of*) in the British National Corpus and about half again as often as the verb *be* in all its forms combined. Remarkable also is the history of the indefinite article *a/an*, which likewise did not exist as such in Old English. Like ModE indefinite plurals, OE indefinites were frequently unmarked, except that *sum* 'a certain' and *ān* 'one' appear sometimes for emphasis and are declined like adjectives.

3.3.4 Adjectives

Like the definite article, Old English adjectives were formally more complex than those of Modern English. Inflectionally, they agreed with their head noun in gender, number and case in 'strong' and 'weak' declensions. The weak declension (with few distinctions) occurred with the highly inflected demonstratives or possessive pronouns. All other environments, including predicative usage, required the more varied strong declension. Thus, adjective inflections partly compensated for the relative lack of inflections in other parts of the noun phrase. Both strong and weak adjectival declensions are given here.

Strong Declension

	Singular			*Plural*		
	M.	*F.*	*Nt.*	*M.*	*F.*	*Nt.*
Nom.	gōd	gōd	gōd	gōde	gōda	gōd
Acc.	gōdne	gōde	gōd	gōde	gōda	gōd
Gen.	gōdes	gōdre	gōdes	gōdra	gōdra	gōdra
Dat.	gōdum	gōdre	gōdum	gōdum	gōdum	gōdum
Inst.	gōde	gōdre	gōde			

Weak Declension

	Singular			*Plural*
	M.	*F.*	*Nt.*	*M./F./Nt.*
Nom.	gōda	gōde	gōde	gōdan
Acc.	gōdan	gōdan	gōde	gōdan
Gen.	gōdan	gōdan	gōdan	gōdra (-ena)
Dat.	gōdan	gōdan	gōdan	gōdum

Nothing of the inflectional system of OE adjectives remains in ModE.

3.3.5 Pronouns

Personal Pronouns

The Modern English pronominal paradigm maintains more of its earlier complexity than does any other form class, as can be seen here.

	Old English					*Modern English*				
	1st	*2nd*	*3rd Person*			*1st*	*2nd*	*3rd Person*		
Singular			*M.*	*F.*	*Nt.*					
Nom.	ic	þū	hē	hēo	hit	I	you	he	she	it
Acc.	mē	þē	hine	hīe	hit	me	you	him	her	it
Gen.	mīn	þīn	his	hiere	his	mine	yours	his	hers	its
Dat.	mē	þē	him	hiere	him	me	you	him	her	it
Dual										
Nom.	wit	git								
Acc.	unc	inc								
Gen.	uncer	incer								
Dat.	unc	inc								
Plural										
Nom.	wē	gē		hīe		we	you		they	
Acc.	ūs	ēow		hīe		us	you		them	
Gen.	ūre	ēower		hiera		ours	yours		theirs	
Dat.	ūs	ēow		him		us	you		them	

A striking difference between Old English and Modern English personal pronouns is the disappearance of the dual number. Further attrition appears in the loss of distinctive singular and plural second person pronouns, where the historical plural forms now serve ambiguously for singular and plural referents; in Middle and early Modern English, *thou*, *thee* and *thine*—reflexes of the OE singulars—developed specialized uses and then disappeared, occurring today only as relics.

With distinct singular and plural forms for first and third person pronouns, the utility of distinct singular and plural second person pronouns is highlighted in vernacular

varieties today by several patently plural forms: *yous* in parts of England, Ireland and Australia; also in metropolitan New York City, parts of the northeast, and elsewhere in the United States, though by no means generally; *y'uns* in Western Pennsylvania and the Ohio Valley; and *y'all* in the American South. Recent decades have seen growing use of *you guys*, initially in the United States and increasingly in Britain. While the *OED* recognizes *yous* and *y'all*, it does not yet enter *you guys*, despite a citation from as early as 1896 (albeit with male addressees) and more than a score of subsequent ones pointing to increasingly gender-neutral usage. (Today such possessive forms as *you guys's* and even *your guys's*, though marginal, can be heard.)

Third person pronouns also show interesting developments. *Its*, the Modern English singular genitive neuter form, is not a reflex of Old English *his* but appeared toward the end of the sixteenth century, extended analogically from the nominative and accusative *hit*, thereby leaving the previously ambiguous *his* unambiguously masculine. The source of *she*, not a reflex of its OE counterpart *hēo*, remains puzzling although a probable origin lies in the feminine demonstrative *sēo* (or its variants *sīo*, *sīe*). Strikingly, too, present day plural *th-* forms are not reflexes of their OE counterparts; rather, *they*, *their* and *them* were borrowed from Scandinavian languages in Britain's north and gradually spread south in the course of Middle English.

At the present time, third person singular pronouns are again undergoing adjustment and accommodation. English has shown continuing tension between gender and number in pronoun use for antecedent nominals. Historically the use of inclusive *he* as an epicene pronoun for male and female referents is common, as in Tennyson's *He makes no friend who never made a foe* (from *OED*), as is the use of *he* and *she* in contrast, as in this *OED* 1509 citation: *He or she that mariage doth breke May fere of deth eternall whan they dye.* Commonly, too, a plural anaphoric *they* is used for indefinite singular, as well as plural, antecedents, as in *Evidently someone knows, don't they?* and *Anyone interested in taking up quilling should contact their local library* (from BNC). Prescriptivists as far back as the eighteenth century have objected to plural anaphoric pronouns after formally singular nominals and have privileged number agreement and generic *he*, discounting the socially more salient gender mismatch. Proposals for a gender-neutral third person singular pronoun such as *thon* and *na* have not succeeded (Baron 1986). During the 1970s and 1980s feminist scholars and others recognized that the use of ostensibly gender-neutral *he* often had the effect of erasing women's roles and contributions, and recent decades have seen increased use of *he or she*, as in *It is of little value to a child if he or she can express him or herself in words but is unable to write in a legible hand* (from BNC). Some commentators judge *he or she* to be awkward and instead recommend reformulations such as *It is of little value to children to express themselves in words if they are unable to write in a legible hand.* It is difficult to say whether singular *they* or a new epicene form will eventually succeed; it is probably safe to predict that generic *he* will continue to decrease in use.

Relative Pronouns

In Old English, an invariant particle *þe* served to mark relative clauses; it was often compounded with a form of the demonstrative *sē, sēo, þæt*, as in masculine *sē þe* and feminine *sēo þe* 'who'. The forms of *sē* also occur alone as relatives, as in *ānne æðeling sē wæs Cyneheard hāten* 'a prince who was called Cyneheard'. OE relatives are also sometimes marked by *þe* and a form of the personal pronoun, as in:

nis	nū	cwicra	nān	þe	ic	him	modsefan	minne	durre	āsecgan
there isn't	now	alive	no one	REL	I	him	mind	my	dare	speak

'There is no one alive now to whom I dare speak my mind.'

Middle English favoured solitary *that* as a relative pronoun, the OE indeclinable *þe* surviving only into early Middle English. In the fifteenth century, *which* (from the OE interrogative *hwylc* 'which') appears as a relative, alternating with *that.* Modern English relative *that*, a functionally adapted reflex of the OE demonstrative, is the relative with broadest pronominal application, anaphoric for noun phrases in nominative and oblique cases other than the possessive, though its use is now limited to restrictive clauses. The ModE relatives *who/whom/whose* and *which* derive from OE interrogative pronouns and can be used with restrictive and nonrestrictive clauses. *Whose* (< ME *whōs* by analogy to *whō* < OE *hwā* and to *whōm* < OE *hwām*) ultimately derives from the OE interrogative pronoun *hwæs*. *Who*, *whose* and *whom* are late developments; while Chaucer occasionally used relative *whose* and *whom*, relative *who* did not come into widespread use until the sixteenth century.

3.4 Syntax

Old English is a synthetic language, relying partly on inflectional morphology to indicate grammatical relations of noun phrases and, to a lesser extent, their semantic roles. Its noun phrases exhibit concord in gender, number and case among the demonstrative/definite article, adjective and noun, with gender a grammatical rather than semantic category. Verbs are inflected for person, number and tense (present and preterit) in indicative and subjunctive moods, the subjunctive occurring far more frequently than in Modern English. Passive voice is signalled periphrastically with *wesan* 'to be' or *weorþan* 'to become' and a past participle; infinitives are sometimes employed passively, and the verb *hātan* 'be called' is generally used with passive force.

As to its word order, late Old English exhibits patterns similar in many respects to those of Modern English. Both show a strong preference for SVO, which ModE exploits in both independent and subordinate clauses, whereas OE, like Modern German, prefers verb-final subordinate clauses. While SOV patterns occur in almost 30 per cent of OE sentences, the twelfth century witnessed the development of an almost exclusively SVO pattern (according to J. Smith, as reported in Hawkins 1983). OE negative sentences introduced by the particle *ne* favour verb-second position, producing a VS order as in the first clause of *ne geseah ic næfre þā burg, ne ic þone sēaþ nāt* 'I have never seen that city, nor do I know the well'. Also apparent in that example is the characteristic negative concord (*ne/næfre* in the first clause; *ne/nāt* in the second, where *nāt* is a contraction of *ne wāt* from *witan* 'to know'). Clauses introduced by *þā* 'then' or *hēr* 'here, in this year' also commonly exhibit verb-second order, as in *þā andswarode Satanas and cwæþ* … 'then Satan answered and said … '; *þā gegaderode Ælfred cyning his fierd* 'then King Alfred gathered his army'; *hēr gefeaht Ecgbryht cyning wiþ fīf and þritig sciphlæsta æt Carrum* 'in this year King Ecgbryht fought against thirty-five shiploads at Charmouth'.

Within Old English noun phrases, the order of elements is usually determiner-adjective-noun, as in Modern English: *sē gōda mann* 'the good man'. Genitives usually precede nouns (far more frequently than in ModE), as in *folces weard* 'people's protector', *mæres līfes man* 'a man of splendid life' and *fōtes trym* 'the space of a foot'. It has

been calculated that the percentage of postnominal genitives increased from about 13 per cent in the year 1035 to 85 per cent in 1300 (J. Smith, as reported in Hawkins 1983). Though prepositions usually precede nominals, with pronouns they often follow, as in *sē hālga Andreas him tō cwæþ* ... 'St Andrew said to him … '. Adjectives too are almost uniformly prenominal (*sē foresprecena here* 'the aforementioned army'), but modifiers can be postnominal, as in these isolated examples cited by Quirk and Wrenn (1955:88–89): *wadu weallendu* 'surging waters'; *ēþel þysne* 'this country'; *wine mīn Unferð* 'my friend Unferð'. Relative clauses generally follow their head nouns.

To a greater degree than Modern English, Old English exhibits a preference for parataxis over hypotaxis. Much OE prose and poetry was written as a series of loosely associated independent clauses, often linked solely by *and*, leaving relationships among succeeding clauses unspecified. While certain genres of informal speech exhibit considerable parataxis in ModE, writing and most spoken genres exploit a high degree of hypotaxis, with logical relationships among clauses marked explicitly by subordinators (*that*, *as*, *if*, *than* and *like* being frequent exemplars).

Denied its earlier inflectional signposts, Modern English developed into an analytical language, more like Chinese than Latin and other early reflexes of Indo-European. With nouns inflected only for possessive case, the chief signal of grammatical relations is now word order, displacing inflectional morphology to such an extent that even the comparatively fuller pronominal inflections are subordinate to the grammatical relations signalled by word order; consequently, an utterance like **her kicked he* may be understood as *she kicked him*.

Why English should have advanced farther along the path to analyticity than other Germanic languages is not altogether clear. A likely basis for explanation lies in the thoroughgoing contact between the Danes and the English after the ninth century, in French ascendance over English for many secular and religious purposes in early Middle English, and in the preservation of the vernacular chiefly in folk speech during the eleventh and twelfth centuries. Decades before the Norman conquest (almost a century earlier in Northumbria), those inflectional reductions started that are everywhere apparent when written English reemerges, and they were doubtless more advanced in speech than extant texts indicate. The syncretism spread as word-order patterns became fixed. Thus, phonological reductions undermined the inflectional morphology and, as flexion grew less able to signal grammatical relations and semantic roles, word order and prepositions (which had somewhat redundantly borne certain aspects of meaning) came to bear those communicative tasks less redundantly. Gradually the freer word order of Old English yielded to the relatively fixed orders of Modern English, whose linear arrangements are the chief carrier of grammatical relations.

Spurred possibly by the virtual absence of inflectional differentiation in its nouns, English syntax has evolved to permit unusually free interplay among grammatical relations and semantic roles. With nouns marked only for genitive case and most pronouns additionally for objective case, Modern English subjects represent a wide range of participant roles. Besides being agents (as in sentence (a)), they may be patients (b), instruments (c), benefactives (d), experiencers (e), locatives (f), temporals (g) and so on; dummy subjects, empty of any semantic content, also occur as in (h):

(a) The janitor (agent) opened the door.
(b) The door (patient) opened.
(c) His first record (instrument) expanded his audiences to thousands of strangers.