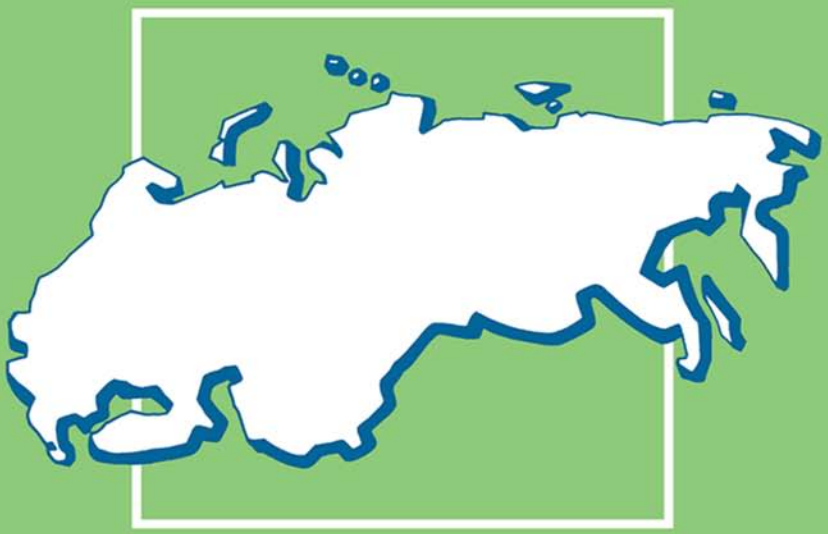
ROUTLEDGE
Reference

The MAJOR LANGUAGES of EASTERN EUROPE



Edited by
BERNARD COMRIE

THE MAJOR LANGUAGES OF EASTERN EUROPE

THE MAJOR LANGUAGES Edited by Bernard Comrie

The Major Languages of Western Europe

The Major Languages of Eastern Europe

The Major Languages of East and South-East Asia

The Major Languages of South Asia, The Middle East and Africa

THE MAJOR LANGUAGES OF EASTERN EUROPE

EDITED BY
BERNARD COMRIE

London

First published as part of
The World's Major Languages in 1987 by
Croom Helm Ltd

Reprinted with revisions and additional material in 1990 by
Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Transferred to Digital Printing 2006

British Library Cataloguing in Publication Data
Available on request

ISBN 0–415–05771–X

Typeset in 10 on 12pt Times by Computype, Middlesex

Publisher's Note
The publisher has gone to great lengths to ensure the quality of this reprint but points out that so imperfections in the original may be apparent

Contents

Preface

The text of this book has been extracted from that of *The World's Major Languages* (Routledge, 1987). The aim of that book was to make available information on some fifty of the world's major languages and language families, in a form that would be accessible and interesting both to the layman with a general interest in language and to the linguist eager to find out about languages outside his or her speciality. Not all of those interested in major languages of the world, however, have an interest that includes all parts of the world, and it therefore seemed advisable to publish portions of the original text in a series of paperbacks — *The Major Languages*. Readers interested in only one part of the world now have access to discussion of those languages without having to acquire the whole volume.

Perhaps the most controversial problem that I had to face in the original volume was the choice of languages to be included. My main criterion was admittedly, a very subjective one: what languages did I think the reader would expect to find included? In answering this question I was, of course, 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, was included because of its immense cultural importance — and I bear full responsibility, as editor, for the final choice.

The notion of 'major language' is obviously primarily a social characterisation, and the fact that a language was not included 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 Walbiri (Warlpiri) and Dyirbal (Jirrbal). Other editors might well have come up with different selections of languages, or have used somewhat different criteria. When linguists learned in 1970 that the last

speaker of Kamassian, a Uralic language originally spoken in Siberia, 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 this person, *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 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 lent strength to the original volume, since within the space limitations of what is quite a sizable 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 criterion for dividing the contents of the original volume among the four new books has been my assessment of likely common and divergent interests: if the reader is interested in language X, then which of the other major languages of the world is he or she likely to be most interested in? In part, my decisions have been governed by consideration of genetic relatedness (for instance, all Romance languages, including Rumanian, are included in *The Major Languages of Western Europe*), in part by consideration of areal interests (so that *The Major Languages of The Middle East, South Asia and Africa* includes the Indo-Iranian languages, along with other languages of the Middle East and South Asia). Inevitably, some difficulties arose in working out the division, especially given the desire not to have too much overlap among volumes, since a reader might want to acquire more than one of the paperback volumes. In fact, the only overlap among the volumes is in the Introduction, substantial parts of which are the same for all volumes, and in the fact that the chapter on Indo-European languages is included in both of the European volumes (given that most of the languages of both western and eastern Europe are Indo-European).

Editorial support in the preparation of my work on the original volume was provided by the Division of Humanities of the University of Southern California, through the research fund of the Andrew W. Mellon Professorship, which I held during 1983–4, and by the Max Planck Institute for Psycholinguistics (Nijmegen, The Netherlands), where I was a visiting research worker in the summer of 1984. I am particularly grateful to

Jonathan Price for his continuing willingness to consult with me on all details of the preparation of the text.

Bernard Comrie
Los Angeles

Abbreviations

abilit.	abilitative
abl.	ablative
abstr.	abstract
acc.	accusative
acr.	actor
act.	active
act.n.	action nominal
adj.	adjective
adv.	adverb
Alb.	Albanian
Am.	American
anim.	animate
aor.	aorist
Ar.	Arabic
Arm.	Armenian
art.	article
Ashk.	Ashkenazi(c)
asp.	aspirated
AT	actor-trigger
athem.	athematic
aux.	auxiliary
Av.	Avestan
ben.	beneficiary
BH	Biblical Hebrew
BN	B-Norwegian
Boh.	Bohemian
BP	Brazilian Portuguese
Br.	British
BT	beneficiary-trigger
c.	common
Cast.	Castilian
Cat.	Catalan
caus.	causative
cc	class concord
Cent.	Central
cl.	class(ifier)
clit.	clitic
comp.	comparative
conj.	conjunction
conjug.	conjugation
conjv.	conjunctive
cont.	contemplated
cop.	copula
cp	class prefix
crs.	currently relevant state
Cz.	Czech
Da.	Danish
dat.	dative
dbl.	double
decl.	declension
def.	definite
dent.	dental
deriv. morph.	derivational morpheme
de-v.	deverbal
dir.	direct
disj.	disjunctive
Dor.	Doric
drc.	directional
DT	dative-trigger
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
foc.	focus
Fr.	French
fut.	future
g.	gender
gen.	genitive

ger.	gerund(ive)
Gk.	Greek
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
intr.	intransitive
inv.	inversion particle
irr.	irrational
It.	Italian
IT	instrument-trigger
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
NN	N-Norwegian
nom.	nominative
noms.	nominalisation
NP	New Persian
nt.	neuter
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
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
pos.	position
poss.	possessive
prep.	preposition
prepl.	prepositional
pres.	present
pret.	preterit

prim.	primary	st.	standard
prog.	progressive	su.	subject
pron.	pronoun	subj.	subjunctive
PT	patient-trigger	sup.	superlative
Ptg.	Portuguese	s.v.	stative verb
Q	question	SVO	subject–verb–object
rat.	rational	Sw.	Swedish
recip.	reciprocal	tap.	tense/aspect pronoun
refl. pron.	reflexive pronoun	tg.	trigger
rel.	relative	them.	thematic
rep.	reported	Tk.	Turkish
res.	result	Toch.	Tocharian
Ru.	Runic	top.	topic
Rum.	Rumanian	tr.	transitive
Rus.	Russian	transg.	transgressive
Sard.	Sardinian	t.v.	transitive verb
SCr.	Serbo-Croat	U.	Umbrian
sec.	secondary	v.	verb
Seph.	Sephardi(c)	v.n.	verbal noun
sg.	singular	vd.	voiced
S-J	Sino-Japanese	Ved.	Vedic
Skt.	Sanskrit	VL	Vulgar Latin
Slk.	Slovak	vls.	voiceless
SOV	subject–object–verb	VO	verb–object
Sp.	Spanish	voc.	vocative
spec.	species	VSO	verb–subject–object

* 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.

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 California 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 this section, 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 any hope of not being contradicted, is that at a very conservative estimate some 4,000 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.

One is that many parts of the world are insufficiently studied from a linguistic viewpoint, so that we simply do not know precisely what languages are spoken there. 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 to the stage where most (though still not all) of the 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. But there are still some areas where uncertainty remains, so that even the most detailed recent index of the world's languages, Voegelin and Voegelin (1977), lists several languages with accompanying question marks, or queries whether one listed language might in fact be the same as some other language but under a different name.

A second 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. With the languages of Europe, 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 grounds rather than strictly linguistic grounds.

One criterion that is often advanced as a purely linguistic criterion is mutal 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 often 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.

While 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 — the native speaker of English would be hard put to interpret a sentence in Tok Pisin, the English-based pidgin of much of Papua New Guinea, like *sapos ol i karamapim bokis bilong yumi, orait bai yumi paitim as bilong ol* 'if they cover our box, then we'll spank them', although each word, except perhaps *i*, is of English origin ('suppose all ?he cover-up-him box belong you-me, all-right by you-me fight-him arse belong all').

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 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 clearcut 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 minimal 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: those who remember the first introduction of American films into Britain often recall that 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 to have died out, 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 language. On the other hand, Greek is referred to in the same way 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 probably been the case throughout mankind's history that languages have died out, the historically recent expansion of European population to the Americas and Australia has resulted in a greatly accelerated rate of language death among the indigenous languages of these areas. Perusal of Voegelin and Voegelin (1977) will show a number of languages as 'possibly extinct' or 'possibly still spoken', plus an even greater number of languages with only a handful of speakers — usually of advanced age — so that a language may well be dying out somewhere in the world as I am writing these words. When a language dies, this is sometimes an abrupt process, such as the death of a fluent speaker who happened to have outlived all other speakers of the language; more typically, however, the community's facility with the language decreases, as more and more functions are taken over by some other language, so that what they speak, in terms of the original language of the community, is only a part of that language. Many linguists working on Australian Aboriginal languages have been forced, in some cases, to do what has come to be called 'salvage linguistics', i.e. to elicit portions of a language from someone who has neither spoken nor heard the language for decades and has perhaps only a vague recollection of what the language was like.

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 organisation 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 below, 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 the name Germanic, and the ancestor language is Proto-Germanic. It should be emphasised 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 or 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 contain 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 Uralic, Mongolian 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, Mongolian, and Tungusic, is rather widely accepted, and some scholars would advocate increasing the size of this family by adding some or all of Uralic, Korean and 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 are somewhat more widely accepted, such as the following list of sixteen stocks, where a stock is simply the highest hierarchical level of genetic relatedness (just as a language family has branches, so families would group together to form stocks): Dravidian, Eurasiatic (including, inter alia, Uralic and Altaic), Indo-European, Nilo-Saharan, Niger-Kordofanian, Afroasiatic, Khoisan, Amerind (all indigenous languages of the Americas except Eskimo-Aleut and Na-Dene), Na-Dene, Austric (including Austro-Asiatic, Tai and Austronesian), Indo-Pacific (including all Papuan languages and Tasmanian), Australian, Sino-Tibetan, Ibero-Caucasian (including Basque and Caucasian), Ket, Burushaski – this schema still operates, incidentally, with two language isolates (Ket and Burushaski), i.e. languages not related to any other language, and retains a number of established language families as distinct (Dravidian, Indo-European, Nilo-Saharan, Niger-Kordofanian, Afro-asiatic, Khoisan, Australian, and Sino-Tibetan). 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 South America) 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 have 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 a Proto-Australian form something like **gudaga* (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* 'daugher'), it clearly can 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 particularly 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' which has given English *mother*, Spanish *madre*, Russian *mat'*, Sanskrit *mātā*. In Mandarin Chinese, the equivalent word is *mā*, while in Wiyaw (Harui) (Papua New Guinea) it is *mam*. Once again, examination of other features of Indo-European languages, Chinese and Wiyaw would soon dispel any possibility of assigning Chinese or Wiyaw 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 cooccur 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 order universals by Greenberg (1963) 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 cooccurrence 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, Mongolian 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 — the dominant and official language of Ethiopia — 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 (with 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 inflection 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 a 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, as members of the Sino-Tibetan family. 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 incontrovertibly 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 other Germanic languages; this is ultimately a loan from the Latin agentive suffix *-ārius*, 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. 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 artifact) 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 change occurred under the influence of Papuan languages, almost all of which are verb-final. Basic vocabulary comes next. And last of all one finds bound morphology. 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 Languages of Eastern 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 Soviet Central Asia (Tadzhikistan), 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).

Much of eastern Europe is covered by one branch of Indo-European, namely Slavonic, though a number of other Indo-European branches and languages occupy more limited territories. The Baltic languages, forming a branch of Indo-European closely related to Slavic, are spoken in Lithuania and Latvia. Rumanian is an outlier of the predominantly southwestern European Romance group of languages. Finally Greek, in its various

historical and geographical variants, forms a separate branch of Indo-European (the branch sometimes being referred to as Hellenic).

Some other languages of Europe belong to the Uralic family. These include Hungarian, Finnish, Estonian and Lappish, to which can be added a number of smaller languages closely related to Finnish or 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 straddles eastern Europe and Asia, being spoken both in the Balkans and in Asia Minor. It belongs to the Turkic family, which is spoken in Turkey, parts of the Caucasus, some areas on the Volga, most of Soviet Central Asia (and stretching down into northwestern Iran), and large parts of southern Siberia. Turkic is perhaps to be joined in a single family (Altaic) with the Mongolian and Tungusic families, spoken in Mongolia, northern China and the eastern USSR.

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, 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 Malay. Swahili is the native language of a relatively small population, 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 even stretches into parts of Zaire. Malay too is the native language of relatively few people in western Malaysia and an even smaller number in Indonesia, but its adoption as the lingua franca and official language of both countries has raised the combined first and second language speakers to well over a hundred million. 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, Cebuano has more native speakers than Tagalog, but I selected Tagalog because it is both the national language of the Philippines and used as a linga franca across much of the country. Among the Indonesian languages, Javanese has more native speakers than Malay and is also the bearer of an old culture, but in terms of the current social situation Malay is clearly the dominant language of this branch of Austronesian. A number of other Indo-Aryan languages would surely have qualified for inclusion in terms of number of speakers, such as Marathi, Rajasthani, Panjabi, Gujarati, 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, where Telugu actually has more speakers than Tamil; Tamil, however, is the more ancient 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 dangerous: as I emphasised 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.

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. Even those languages