Chapter 2

Guide to the Register: the Referential Framework

2.1 Background to the Register

Objectives

The Linguasphere Register spotlights the importance of languages and multilingual communication in the future construction of a planetary, transnational society. The present volumes are designed not only to present an outline view of the linguasphere at the turn of the millennium, but also – through the Observatoire Linguistique (Linguasphere Observatory) website – to stimulate the convergence and co-ordination of data and research on the world's languages and speech communities, from the most widespread to those under threat of imminent extinction. They are equally concerned to stimulate the sharing of ideas and initiatives leading towards the harmonious development of a multilingual global society, from which no community will be excluded.

The only practical way to classify the continuum of human society around the globe is in terms of the language or languages used within each community, and the Linguasphere Register is the first attempt to undertake the necessary systematic overview of contemporary *speech-communities*. It provides a first approximation on the road towards a detailed classification of modern humanity or *humankind*, presented not in terms of transient political frontiers, but of transnational linguistic relationships and of networks of communities within a continuous global system of communication. The compilation of the Register has involved the collection and piecing together of available data on the languages and speech-communities of the modern world, within a coherent and systematic framework of classification and coded reference. Rather than being constructed from the "top" downwards, or centred on the study of ancient or presently dominant languages, this classification is based on the living dialects and "inner languages" of speech-communities in all parts of the world, regardless of their relative size or historical importance.

The central objective of this framework edition of the Linguasphere Register has been to record all known variants of living and recorded languages in the world and to classify them in terms of their closest relationships (to a current total of over 20,000 *inner languages* and *dialects*).

In realising this objective, the present edition provides the only detailed classification of the world's languages and speech communities completed before the end of the 20th century, and constitutes the basis for the Observatoire Linguistique's global survey of languages and speech communities from the year 2000.

An underlying philosophical purpose of the Register is to proclaim the essential equality of every speech community in the world, however small or isolated or hitherto neglected or disregarded its language may be. It is hoped that speakers or observers of every such language will verify that its name is now correctly registered and indexed in this Register, alongside and on the same terms as the current *megalanguages* of Arabic, Bengali, Chinese, English, French, German, Hindi-Urdu, Japanese, Malay-Indonesian, Portuguese, Spanish and Russian (see pp. 291-92). A comprehensive Register of the world's languages and speech-communities now exists in these volumes, and the continuing input of correspondents around the world will ensure that every language is included accurately, together with adequate information on its nomenclature, location, written forms, current use and relative number of speakers (see section on Expanding the Register, below). The Index to the Register, published below with a total of over 70,000 entries, is already the longest list of ethnic and linguistic names yet published.

The Linguasphere Register provides a new source of reference and reflection, not only for students, teachers and researchers involved in the study of languages and linguistics, but also for all concerned with the global, regional and even local study of human diversity – in the fields of geography, the social sciences, politics and economics, development studies and humanitarian aid. The kaleidoscope of the world's speech communities may now be viewed as the footprints of a single but complex human family, extending across the frontiers of nationality, religion and ethnic culture, and across the divides of wealth, prestige and education. Humankind is revealed as even more varied than previously supposed, yet also as a continuous entity.

The Register is intended to serve not only as a source of reference but also as an educational resource, and parallel materials are being developed to encourage its use in schools and colleges. These will provide a new window on the global continuum of humankind, and of speech itself, encouraging young people everywhere to situate themselves and their local communities within that global continuum, and to collect further data on languages and language-use and on communities in contact. The design criteria for the Register have required it to be:

- comprehensive (applicable to all modern forms of speech, and of all modern and inherited forms of writing);
- straightforward (enabling each language to be readily located within the linguasphere and alongside its closest linguistic relatives); and
- elegant (based on a simple and regular method of coding and display of data).

This first edition of the Register has been compiled and authored by a single person, thus ensuring a unity of conception and application. Although errors and omissions are his sole responsibility, a very large number of individuals have contributed information and insights during the years of preparation, and the warmest appreciation and thanks of the Observatoire Linguistique are due to them all, as expressed in the Acknowledgements below. Based on a very wide range of published sources, and on numerous discussions with colleagues and grass-roots observers, the Register does not endeavour to replicate the uncertainties and disagreements of the professional literature, but rather to establish a coherent picture of the whole linguasphere.

It must be made clear that even an inventory as lengthy and complex as the Linguasphere Register can provide only a very rough preliminary outline of the linguasphere, and that its gradual improvement and expansion, and detailed documentation and annotation, could continue indefinitely – if life permitted - in the quest for a perfectly researched volume.

Since further delay is not an option, however, the present framework edition of the Linguasphere Register has been completed within a deadline of 31st December 1999, and is now published with the following aims:

- to provide the first planetary outline of humankind's linguistic environment or *linguasphere*, at the beginning of a new era of global communication;
- to launch a transnational "roll-call" of human speech communities at the turn of the millennium, in the form of a classified and indexed guide to the nomenclature, location, linguistic relationships and demography of the languages and peoples of the world.
- to establish a stable framework of worldwide reference for the documentation and mapping of the world's languages and speech communities, from the beginning of the 21st century;
- to set out a flexible scale of linguistic proximity for the expansion and adjustment of a worldwide corpus of data on individual languages, independently of their demographic importance;
- to begin the transnational monitoring of all human communities, however small, isolated or socially disadvantaged, within a perspective of global relationship and shared concern.
- to participate in the development of education about humankind, as an essential part of teachers' and students' understanding of the modern telecommunicational world.
- to encourage the application of information technology to the co-ordinated study of the world's languages, including the bridging of close linguistic relationships by automatic conversion, the audio-recognition of specific languages and dialects (via a codified 'language-bank'), and the data-based mapping of humankind.

An interest in that task of cataloguing the languages of the world – regardless of their literary importance - is not new, and extends from the pioneering interest of Catherine the Great of Russia in the late 18th century to the achievements of Barbara Grimes in Hawaii two hundred years later. Barbara Grimes' work has sometimes been criticised – as this Register may also be - because of its inevitable imperfections and lacunæ. But her regularly revised compilation, the Ethnologue (13th edition, Grimes 1996), has proved to be the most valuable single source of information hitherto available on the languages of the world, and has made an important contribution to the compilation of this Register.

In addition to her global documentation of spoken languages, Barbara Grimes is also to be congratulated for her attention to the remarkable wealth of deaf sign languages in use throughout the world. This is an important aspect of the linguasphere, which has not been included in the present Register of spoken (and written) languages, but which will need to be covered in future editions.

Philosophy of the Register

Most aspects of planet Earth - from its geological structure to its flora and fauna – have been studied and classified. Yet perhaps the most remarkable element on its surface – the family of humankind - has never been systematically classified. The artificial boundaries of nation-states and the old-fashioned division of humankind into 'races', based on differing degrees of pigmentation, are of no help in studying human beings as members of a planetary society. The only measurable framework for the detailed classification of human communities in today's world is that provided by the languages they speak.

There have been attempts in the past to classify the world's languages historically, and to glean scraps of linguistic evidence in the search for clues about human prehistoric origins. This commendable endeavour should, however, be clearly distinguished from the otherwise neglected task of observing and classifying the web of human languages - in living speech, in writing and in electronic form - as they exist and function around the globe today.

Hitherto, classifiers of languages have been concerned primarily with looking back towards the reconstruction of distant and often hypothetical prehistoric relationships. The Linguasphere Register, on the other hand, is concerned primarily with recording all known variants of living and recorded speech in the world and with classifying them in terms of their closest relationships. The purpose of this approach is to clarify the bonds of speech which unite communities across national frontiers and to provide a framework of reference which makes it easier to find a path through the maze of languages and communities around the world, and through their vast array of names.

Although primarily concerned with the observation of the modern linguasphere, the Register is in no way anti-historical, and it is intended that it may provide a service also to comparative linguists, as a neutral framework for the referencing of new or revised historical classifications.

Rather than being presented as a dry academic or statistical exercise, the Linguasphere Register has been designed to observe and present humankind as a single but highly complex planetary society. The *linguasphere* – or communicational environment of the world - is the shared inheritance of all, and the Register is the first classified catalogue of that unique global heritage, compiled to create a better understanding of each person's place within it. In presenting a structured overview of the world's languages, the Register seeks to draw attention to these two neglected end-points of communication in the contemporary world, which are the *voice* of the individual person and the totality of the *linguasphere* itself, as already discussed in the previous chapter. A referential system of linguistic classification and coding, as set out below, has been developed to span the range of *layers* of communication between these two end-points.

The present revolution in telecommunications is the most rapid and global in history, and obliges humankind to look back at the mixed benefits of its collective cultural heritage. It is vital for the survival of future generations that traditions encouraging the freedom and diversity of language and expression, and of the creative arts, be treasured and preserved, but that all traditions of hatred and prejudice, and of social, ethnic and doctrinal rivalry, be left behind with the debris of the last millennium.

From the average person's viewpoint of their primary language, as spoken during a single lifetime, that language may appear relatively static, apart from established variations associated with different localities or income groups. The changes observed between childhood and old-age are not overwhelming, even if an elderly person chooses to interpret them as departures from a desired or former standard of supposed correctness or purity.

Even the traditional 19th and 20th century view of languages did not undermine this impression of regularity and stability. Each 'family' of languages was considered to descend from a single common 'ancestor', sounds changing according to regular rules. Some languages were frozen in a standardised form, to serve as 'cultivated' vehicles for polite conversation and literature, and for governing and educating their respective nations. Languages not chosen for this status were allowed to languish as unwritten local 'patois', or be dissolved in an apparent welter of localised variants.

At first sight, the Linguasphere Register appears to confirm this orderly view, with every form of language neatly labelled and classified. And yet its underlying lesson is very different. By classifying <u>all</u> forms of language in terms of *layers* of *immediate* and *close relationship*, the Register emphasises the need to respect the language of every speech community, regardless of geographical location and economic circumstances, and regardless of the degree of written development or standardisation.

The introduction of a form of coding in the Register has been essential, in order to create signposts around the remarkable maze of human languages and speech-communities which encircle the globe. Within this transnational frame of reference, their continuity and inter-relationships can be viewed for the first time across the boundaries of nation-states and state-languages, which during recent centuries have been pasted over the map of the world.

But most importantly, the panorama of the world's languages provided by the Register enables them to be seen as constituent parts of the *linguasphere*, the artificial but essential operating system which humankind has evolved for itself. This unconscious and cumulative creation is not an independent dimension of the planet but exists as a collective dimension of every participating *voice* and communicating brain, within the *cerebral network* of an increasingly globalised society.

Languages are traditionally associated with the geographical areas where they are spoken, and the Register records a large proportion of these territorial associations.[28] Linking a language with specific places on the globe's surface should not, however, obscure its real location within the human brains of its speakers, the majority of whom may indeed reside within a given geographical area. Since each act of human communication has its origin in the human mind and is addressed to or through the human mind, so the linguasphere itself, including the underlying system of every language – memorised, spoken and heard, written and read - can be considered to reside essentially within the cerebral network of individual human brains. The medium of communication among the constituent parts of that cerebral network is provided by the tactile, spoken, written, printed, electronic and telepathic space which surrounds and unites humankind.

In the 1960's, Marshall McLuhan introduced the idea of the world becoming a *global village*, but even this relatively modern cliché is now too static and too comfortable. A more appropriate image for the 2000's can be found by comparing a now globalised human community to travellers on a leaking ocean-liner, in which some passengers live in conspicuous luxury on the top decks and others in grinding poverty in the steerage, with open violence and arson on the decks in between. In this context, the Linguasphere Register provides a first roll-call of the groups and families of passengers who in the 20th century have been voyaging through space on S/S Earth.

As fellow passengers on a leaking vessel, with no lifeboats, all human beings are now "us", and there is no more "them".

Framework of the Register

Central to the aims set out above has been the establishment of a stable *framework* of worldwide reference for the documentation and mapping of the world's languages and speech communities, alongside a flexible scale of linguistic proximity for the progressive expansion and refinement of a worldwide corpus of data on individual languages.

Since the linguasphere is a fluid continuum of human communication through both time and space, its classification into distinctly separated parts is in contradiction to its nature. On the other hand, a form of superimposed classification is essential in order that the linguasphere may be studied in detail, and in order that large amounts of data may be stored, retrieved and manipulated electronically. The absence of any appropriate system of classification has therefore been a principal reason for the absence until now of any co-ordinated study of all the world's languages. This is the gap which the Register has sought to fill, by producing a classificational structure whose outer layer is stable enough to provide a framework for the sorting, storing and retrieval of data, but whose internal layers are flexible enough to allow the data to be amended and expanded as necessary, including inevitable cases of resorting.

A requirement of such a classification is that it be rigid at the time of publication, with every language and dialect assigned a fixed and coded position.

[28] As indicated by the symbol ⊕ in column 3.

Based on a wide range of published sources, and on discussions with specialists and grass-roots observers, the present Register does not endeavour to incorporate the uncertainties and disagreements and lacunae of the professional literature, but to arrive at a first coherent picture of the whole linguasphere. Areas of uncertainty are indicated, by a star rather than a question-mark, but every language and dialect is assigned a coded place within the overall classificatory scheme, as explained below. In a small minority of cases, the details of classification are stated specifically to be *notional*, but the coding of all such languages remains firm until the appearance of a subsequent edition. A simplified system of reference – the *Linguasphere Key* – enables individual languages and dialects to be identified by use of a stable and unambiguous two-digit code, even in cases where the its detailed classification may be subsequently adjusted or amended (see end of this section).

For the present framework edition, the unified design and consistent implementation of the classification have been developed as the work of a single person. From the year 2000, however, the Linguasphere Register has become a collaborative venture, and the Observatoire Linguistique website <www.linguasphere.info> is operating as a focal point for the collection and co-ordination of additional and improved data and documentation.

All information received and incorporated will be acknowledged in subsequent editions. Please see 2.6 Expanding the Register, below.

History and Compilation of the Register

Research leading to the subsequent Register began with the study of European languages from the 1950's [29] and with the comparative study and mapping of the languages of Africa from the 1960's and 1970's [30], based at the School of Oriental and African Studies in London and at the International African Institute in London and Paris.

Africa is the most multilingual continent in a highly multilingual world, and there was a clear need for a continental map which would present the overall linguistic complexity of Africa as an entity in itself. This first task was completed in 1977 - cataloguing and presenting over 2000 languages - and it provided two fundamental lessons for future research and classification, which drew the compiler from Africa towards the global study of the linguasphere.

The first lesson was that Africa is not an island, and that the complex linguistic structure of that continent can only be satisfactorily viewed within the wider terrestrial context. From the end of the 1970s, it was possible to embark on the gradual gathering of data for the present Register.

The second lesson was that a new referential system of classification needed to be designed which would permit the comprehensive organisation and further collection of data on the languages and dialects of the world, but which would not need to be fundamentally reorganised to account for subsequent reconstructions of prehistoric linguistic relationships.

The design and compilation of the Register was subsequently developed as a major activity of the Observatoire Linguistique in France during the 1980's and 1990's. The first experimental part of the Register was published in French,[31] with the generous assistance of the Agence de la Francophonie[32] and the Centre international des industries de la langue (Université de Paris-Nanterre). French was also the language of the first published presentations of the Register, which appeared in Belgium in 1992 and (with translations into Spanish and Italian) in 1994,[33] and it was originally intended to publish the present framework edition either in French alone, or bilingually in French and English.

[29] See Dalby 1965b
[30] See Dalby 1962, Dalby 1977, etc.
[31] Dalby 1993
[32] At that time known as the Agence de coopération culturelle et technique.
[33] See Dalby 1992a & 1994b.

In the event, financial and practical considerations have prevented the present edition from appearing simultaneously in both languages, but support is currently being sought for the early finalisation of a French version of the Register, entitled « Répertoire de la Linguasphère ». Its publication will prepare the way towards a system of multilingual access to the Register, and to the projected *Linguasphere Mapbase*.

After the first presentations of the Register in 1992-94 (see above), a grant was awarded by the Leverhulme Trust Fund in 1994 to the Department of Geography at the School of Oriental and African Studies (SOAS) in London for the development of the cartographic applications of the Register, including the design of new techniques of computerised language mapping. This programme led to the undertaking of three separate projects:

- the preparation of the Language Map of Africa in computerised format, as the first continental sheet of a projected Map of the Linguasphere;[34]
- the drafting of a series of experimental maps of Mauritius, based on linguistic questions in the 1991 general census,[35]
- and the undertaking of a statistical and cartographic survey of immigrant language communities in greater London.[36]

In 1996, an important statement was made at Bilbao on the subject of Unesco's priorities in the field of language, by the then Director-General of Unesco, Sr. Federico Mayor Zaragoza, in 1996. He called for the preparation of a global language map - "un mapa lingüístico mundial", and at a subsequent meeting held at the Unesco centre in Catalonia in June 1997, it was agreed that the Unesco Linguapax Programme (devoted to the promotion of multilingualism in the interests of peace) and the Observatoire Linguistique would collaborate in the eventual production of a computerized and data-based language map of the world. In Paris, on 25th September 1997, the first copy of the preview edition of the Linguasphere Register[37] was formally presented to the Director-General of Unesco, and it was subsequently proposed by Unesco that the Register should constitute an introductory part of a first World Linguistic Report.[38] The finalisation of the Register has been a necessary forerunner to the development of a *mapbase* of the world, which – if adequate funding is forthcoming - will involve the (Indian) Centre for Development of Advanced Computing in Pune and the (French) Centre du traitement informatique et géographique in Aix-en-Provence, as well as the Department of Geography at the London School of Oriental and African Studies.

This first edition of the Linguasphere Register has now been prepared for publication in Wales, and it is hoped that Wales will continue to be one of the focal points for its subsequent development.

At a time when speech-communities throughout the world have to take account of the expanding use of global English, it is appropriate that the Linguasphere Register should now be closely associated with two countries which have had such intensive experience of that language as a colonial intrusion on their own linguistic scene. Originally an instrument of external power, the English language has become an integral and permanent feature of both countries, but in each case within the context of their own rich and vigorous linguistic and literary traditions:

---

[34] for which the principal work was undertaken by Yasir Mohieldeen, based on Dalby 1977

[35] undertaken by Philip Baker of Westminster University and Professor Vinesh Hookoomsing of the University of Mauritius, with cartographic support from Mike Farmer.

[36] undertaken by Philip Baker in collaboration with John Eversley of Queen Mary Westfield College, with cartography by Yasir Mohieldeen, and subsequently published with the support of the Lord Mayor and Corporation of London: see Baker & Eversley 2000.

[37] *The Global Language* Register*: a transnational key to the world's languages and peoples*, pre-published with the support of the School of Oriental and African Studies (University of London): see Dalby 1997.

[38] Letter from the Unesco Assistant Director-General for Education, dated 18th July 1997.

- in the small European country of Cymru (Wales), whose rural communities have preserved their Celtic language throughout this millennium, on the doorstep of the English language;
- and in India, one of the two most populous countries in the world, which while preserving its own languages is today second only to the United States in terms of the numbers of voices speaking English.

Although the Observatoire Linguistique is independent of all national and international bodies, and has prepared and produced this first complete edition of the Register from the resources of its own members, it welcomes this opportunity of collaboration with all like-minded organisations committed to the study and development of the world's linguistic heritage.

The preview editions of the Register in 1998 and 1999 were directly associated with Unesco and with a number of universities and research institutions in Europe and India. However, in order to emphasise the neutrality and independence of the Observatoire Linguistique, it has been decided to publish the present framework edition in the name of the Linguasphere Press only, which has been created to manage the publications interests of the Observatory, and thereby to provide support for its research and information activities.

Layout of the Register

The 742 pages of the tabulated Register, as published in the accompanying Volume Two, are ordered and coded according to a consistent classification, with a corresponding index in the present volume, on pp. 121-286 below. Every language has its place within this 1999/2000 edition of the Linguasphere classification, regardless of the relative state of knowledge or ignorance in the area concerned. Provisional or uncertain elements are indicated by a raised star (*) but a firm decision on current classification and coding has been made for every language, based on probabilities or assumptions in the minority of cases where this is necessary.

This heuristic approach has been catered for by the basic structure of the Register, as explained in detail below, a structure which not only allows for classification by affinity and/or geographical location, but which is also readily adaptable to future improvements and expansion of knowledge. A rigorously consistent system of coding enables each element to have its place, even if provisionally, and opens the door to the computerised organisation, retrieval, cross-referencing and quantification of data on any form of spoken or written language.

From this publication onwards, work on the Register has become collective, and the Observatoire Linguistique is already establishing a network of linguistic observers and commentators in different parts of the world, who will participate in the preparation of future editions of the Register. At each stage in its future development, the Linguasphere Register and its projected *Mapbase* will aim, as this edition now does, to present the best possible coherent model of the current linguasphere.

The structure of the Register has been designed to shift the emphasis of linguistic classification away from distant *levels* of historical relationship, often hypothetical, towards observable and verifiable *layers* of close and immediate relationship among modern languages and dialects. In discussing and presenting hierarchies of linguistic relationships, the Linguasphere Register uses the term "layer" throughout (rather than "level") in order to avoid possible undesirable implications of "higher" and "lower" varieties of language.

To address the fundamental problem of classifying a complex set of data when the collection and analysis of those data are still in progress, the Register has been constructed around a two-tier system of coding:

- an outer system of numerical codes, designed to provide a stable framework of worldwide reference;
- and an inner system of alphabetical codes, designed to serve as a sliding-scale of linguistic proximity and to be adaptable to all additional data and refinements of analysis.

The linguasphere is thus first divided for reference purposes into ten sectors and one hundred zones, coded numerically. *Sets* of individual languages are classified and letter-coded within those zones in terms of their close and immediate relationships with other languages.

This system of classification is reflected in the vertical structure of the Linguasphere Register (and in the codes of column 1), while its horizontal structure conveys key information on successive groups of languages, and on individual languages and dialects (row by row).

In a continuous table, within which each row is subdivided into five columns, the Register provides an overview of the modern linguasphere, covering all languages spoken or written during at least part of the 20th century.

Also included are certain forms of written languages recorded from earlier centuries, wherever those records have remained part of the cultural environment of the 20th century, especially in literary or liturgical contexts (and which thus still form part of the modern linguasphere).

Entries have also been included for some languages known to have become extinct during the previous four centuries (from the late 15th to the end of the 19th), since these are directly relevant to the consideration of the impact of European languages on the state of the linguasphere. For the present edition, this important extension of the Register's coverage is still largely confined to North American and Tasmanian languages (see 6=North America and 29=Trans-Australia).

A raised star ***** is suffixed to items of data which are unreliable or which require corroboration.

The Five Columns of the Register

The five columns of the Register (compressed to three in the Synopsis, see Vol.Two pp. 16-35) are organised as follows. For the framework nature of the data provided in columns 3 and 4, see the introductory paragraphs to section 2.6 below.

Column 1 presents a complete coded classification of the world's *language-groups* (sets, chains and nets) and *idioms* (outer languages, inner languages and dialects). This classification is constructed around the numerical and alphabetical codes presented in the table on p.297.

Column 2 presents a list of selected reference names for all language-groups and idioms, their classificational hierarchy being visually apparent from sequences of typography (ranging from bold capitals to normal lower-case). The reference names of idioms represent wherever possible speakers' own-names or *autoglossonyms* for their primary forms of speech. Most reference names of language-groups in the Register are constructed from a combination of the names of two of their component elements, rather than utilising existing, often artificial or foreign "cover names". Names of languages which are today read rather than spoken are prefixed by the icon of a book 🕮, whereas spoken languages modelled at least partially on the written word are preceded by the icon of a writing hand –. For this first framework edition of the Register, names have been recorded only in the Latin script.

A series of capitalised suffixes has been introduced, for the layers of immediate relationship, to distinguish reference names which would otherwise be identical within the same zone. This includes the distinction of languages and dialects by their directional location or by an aspect their usage.

The suffixes are as follows (followed by a point in the original printed version, e.g. –N. –E. –S. etc) :

> **Directional suffixes**, in columns 2 & 3: -N north(ern); -E east(ern); -S south(ern); -W west(ern); -C central; plus combinations, e.g. -NW northwest; -CW west central, etc.
>
> **Other suffixes**, in column 2: -A "proper" (name); -F formal or standard; -G generalized; -L liturgical / pre-modern literary; -M middle; -U urban; -V vehicular

The suffix -A is used in the sense of "proper" («proprement dit» in French), in cases where a linguistic name is repeated in successive non-identical rows, e.g. [53=] Slovensko-A (Slovenian "proper" or Common Slovenian) as one of the components of [53=] Slovensko (Wider Slovenian).

Column 3 presents the alternative names recorded for many language-groups and idioms, including alternative reference names in bold type. Other names applied to languages and communities are distinguished by the use of lower case initials, as opposed to initial capitals for geographical names. (This typographical convention does not apply to textual notes, printed in italics.) Notes are categorised by a series of icons:

⊕	notes on locations or epicentres	🕮	notes on scripts or written models
¶	notes on speech communities		
➢	notes on languages	#	notes on nomenclature
X	notes on contacts and relationships among languages		

Cross-references to other languages are preceded by the *linguasphere key* in square brackets, e.g. [49=] *Telugu* (i.e. classified in zone 49=, see table of sectors and zones on p.300).

Column 4 lists the nation-state or states in which an idiom is spoken (with provinces in brackets), with any official status indicated by the icon of a flag υ before the name of that state (or province). A separately bracketed flag (υ) indicates that the idiom no longer has the official status which it once had in the relevant nation-state. ~~Use in two or more states is marked by the icon of crossed flags.~~ A technical problem led to omission of the flag symbol υ at some places in Volume Two, but an alternative source of information on the official languages of all nation-states is available from the National Index on pages 287 ff. below.

Column 5 presents a single-digit *scale of voices* (i.e. speakers) for individual languages, and for the combined languages of each zone. This digit records the order of magnitude of the number of primary and alternate speakers of every outer language in the Register (and of some inner languages), as known or estimated at the end of the 20th century. This estimate is expressed on a scale from 0 (extinct since 1900) through 1 (less than 100), 2 (100+), 3 (1000+), 4 (10,000+), 5 (100,000+), 6 (1,000,000+), 7 (10,000,000+), 8 (100,000,000+) to 9 (over one billion). The icon ι marks complete nets, chains or sets of idioms which were extinct before the end of the 20th century, while the icon ● marks idioms known or believed to be extinct before the end of the 19th.

The Linguasphere Key

The numerical framework of worldwide reference, composed of one hundred referential *zones* within ten referential *sectors* as tabulated on p.300, makes it possible for any idiom (*language* or *dialect*) or any defined *group* of languages in the world, or any *speech community*, to be simply and unambiguously identified by means of its *linguasphere key*.

> ●*This linguasphere key consists of the two digits (plus equal sign) of the relevant zone, placed in square brackets before the reference name (or any other name which is unique within the relevant zone) of the idiom or group or community to be identified.*

Throughout the five-column Register, the first two digits in column 1 record the relevant zone (and linguasphere key) of each idiom or group, while column 2 records the corresponding reference name of that idiom or group (always unique within its own zone). Alternative reference names (also unique within each zone, wherever they occur) are printed in bold as the first items in column 3.

The reference name (or other appropriate name) of any language or group of languages in the world may thus be unambiguously identified for referential purposes by the prefixing to that name of the relevant zonal digits in square brackets, for example:

- [59=] Bangla or [59=] Bengali or [59=] Bengali (Bangla)

- [99=] Kiswahili or [59=] Swahili, etc.

This device is used in discussing or referring to specific languages throughout this volume, and is available as a free option in references to languages in other published sources, wherever this appears useful. The Linguasphere Key is designed to provide as simple as possible an indicator of the identity of any language or speech community in the world, and may in fact be prefixed to any form or variant of any relevant name, provided that name is unque within its zone (as may be confirmed from the Index to the Register). In order to distinguish variations in the application of the same or identical name within the same zone, a series of suffixed letters has been introduced in the Register, as set out in section 2.5 Linguistic Nomenclature, for example:

- [50=] Cymraeg-F or [50=] formal "book Welsh"
  cf. [50=] Cymraeg-N. or [50=] Cymraeg y Gogledd or [50=] Northwalian Welsh, etc.

The citation of each relevant linguasphere key (zonal code plus reference name) provides a unique and unambiguous reference for any language or dialect in the world, or for any group of languages, and enables the full classificational code to be retrieved from the Linguasphere Index below, even in cases where an identical reference name occurs in two or more zones or where the following alphabetical code may be subsequently modified.

The typographical rules applied to the citation of language names in columns 2 and 3 of the Register (e.g. lower case initials for names of dialects in column 2) or in the Index (see page 121 below) do not apply to names cited outside the Register and Index themselves and do not normally apply to cross-references to other language names within italicised notes in column 3.

The name of a specific sector or zone is always prefixed by the relevant digit or digits plus a double hyphen (equal sign) without square brackets, e.g. 9=Transafrican or 99=Bantuic, thus distinguishing it from any dialect, language or group of languages within that referential sector or zone, e.g. [99=] Zulu.

The Linguasphere key may be freely used as a standard form of reference in any other printed or digital source. The simplified table of sectors and zones in section 2.2 below (with or without totals of component *sets*) and the corresponding map (facing p.300) may also be freely reproduced in connection with any referential use of the linguasphere key.

2.2 Sectors and Zones

Sectors

Although the Linguasphere Register is not concerned with the classification of distant relationships, it makes use of the fact that the primary (mother-tongue) languages of around 85% of humanity have been classified, with widespread agreement, within only five major linguistic 'families' or *affinities*. These have been known in English as: Austronesian, Afro-Asiatic (or 'Hamitic-Semitic'), Indo-European, Sino-Tibetan and Atlantic-Congo (or 'Old Niger-Congo' less Mande).

Such linguistic 'families' (as described in the genetic terminology of historical linguistics) are referred to in the Register as continental or intercontinental affinities, reflecting the Register's primary concern with synchronic rather than diachronic relationships.

The term *affinity* has the further advantage that it does not seek to exclude the effects of the historical convergence of languages over long periods of time, or of elements copied (or 'loaned') among languages. It should be noted that proposed prehistoric relationships which have been disputed, or which have not yet received general acceptance, are referred to in the Register as *hypotheses* rather than *affinities* (as for example in the case of the "Altaic" hypothesis, under zones 44= and 45=).

The first major referential division which can be established within the linguasphere is the division between (i) languages classified outside the five major 'families' or affinities, and (ii) all those languages which have been classified within them:

Languages in the former category (i) have been classified by cautious historical linguists into more than two hundred separate entities (treated in the Register as isolated *sets* of one or more languages, or as *groupings* or *affinities* of two or more *sets* each), and are therefore classified initially within the Register according to purely geographical criteria, within five *geosectors* corresponding to the continent where they are spoken.

Languages in the latter category (ii) (including, as it happens, all major languages with an "intercontinental" distribution) are classified within five linguistic *phylosectors* corresponding to the continental or intercontinental affinity to which each of them belongs.

Each of the five *geosectors* bears the name of the relevant continental area (ending in English always in –a), while each of the *phylosectors* bears the name of, or a name corresponding to, the relevant linguistic affinity (ending in English always in –an).

The ten sectors are then ordered (both alphabetically and numerically) in such a way that:

***geosectors* are indicated by even digits:**

0=AFRICA

2=AUSTRALASIA

4=EURASIA

6=NORTH-AMERICA

8=SOUTH-AMERICA

and *phylosectors* are indicated by odd digits:

1=AFRO-ASIAN (containing languages of the Afro-Asiatic or Hamito-Semitic affinity)

3=AUSTRONESIAN (containing languages of the Austronesian affinity)

5=INDO-EUROPEAN (containing languages of the Indo-European affinity)

7=SINO-INDIAN (containing languages of the Sino-Tibetan affinity)

9=TRANSAFRICAN (containing languages of the Atlantic-Congo affinity)

Numerical framework of worldwide reference > SEE FULL TABLE on p.297

Each *set* of languages is classified and coded within one of 100 referential *zones* within one of 10 referential *sectors* (one of 5 *phylosectors* or 5 *geosectors*).

linguasphere key : a fixed two-digit **numerical code** (99= as an example)	marking ➢ two **layers of worldwide reference**	for an inventory of sectors and zones, see table on p.300	TOTALS
(uncoded)	(LINGUASPHERE)	the totality of the world's languages	1
9=	**SECTOR**	**phylosector** (odd digit 1, 3, 5, 7, 9) *or* **geosector** (even digit 0, 2, 4, 6, 8)	10
99=	ZONE	**phylozone** *or* **geozone**	100

Comparative implications of the Geosectors and Phylosectors

When the five *geosectors* and *phylosectors* were established and developed as a framework for the referential classification of the world's languages, it was not anticipated that their analysis as entities would be more than of statistical interest – a way of organizing data on the contemporary linguasphere. In the event, however, their comparison as categories of language, and especially as phylosectors contrasted with geosectors, has stimulated reflection on the global dynamics of the linguasphere.

Attention has already been drawn to the fact that the present linguasphere has most probably – and certainly most logically – developed from a *paleolinguasphere* which was even more complex and diverse than the *neolinguasphere* recorded in this Register (rather than the reverse picture of diversifying "family-trees", as some historical linguists have tended to imply).

The announcement, two centuries ago in India, that the languages now known as Indo-European had diverged from a common source, has inspired subsequent generations of linguists to seek to emulate that discovery. Many successful and even more unsuccessful attempts have been made to try to recreate other linguistic family-trees wherever similarities among languages, from regular correspondences to scattered items of vocabulary, have seemed to point in that direction. The much vaunted regularity of phonological relationships which can sometimes be found among languages which are quite closely related, like those of the Indo-European *affinity* itself, has nurtured a tradition whereby the fluidity and complexity of linguistic change has sometimes been understated in the quest for supposed 'genetic' origins.

The genetic and family-tree analogies are inappropriate for a continuum which is basically a history of mental and social conventions rather than of biological varieties.

The general and of course much simplified picture which emerges from a review of the languages of all continents is that linguistic traditions – ways of speaking – associated with the spread of agriculture and larger social entities have progressively overwhelmed, replaced, and/or absorbed the much more

fragmented and diversified languages of small communities living very often from hunting, gathering and fishing. Recent historical events in Australia, Amazonia, North America, Southwest Africa, Siberia and in scattered hills, forests and islands of Asia and Oceania have been among the last chapters of this sad but inevitable story. Of those who lament the human tragedy involved (and who does not?), how many would be prepared to turn the clock back, or to return lands to their original use which once belonged to hunter-gatherer communities?

In this context, it is tempting to picture the languages classified within the phylosectors as representing a "secondary" spread of language, and of widely distributed linguistic *affinities*, in contrast to those classified in so many small and fragmentary *sets* within the geosectors. Presented with this temptation, it is interesting to consider the first comparative statistics which have resulted from the completion of the *Register*, as summarised briefly on the last page of this volume (p. 300: table of sectors and zones).

From these statistics, it becomes apparent that the languages of four of the five phylosectors (1=Afro-Asian, 5=Indo-European, 7=Sino-Indian and 9=Transafrican) have a very different collective profile, in terms of their apparent vitality, from those classified within four of the five geosectors (0=Africa, 2=Australasia, 6=North-America and 8=South-America). If phylosector 3=Austronesian and geosector 4=Eurasia are thus (for the moment) excluded from the comparison, then the contrast between the profiles of the four geosectors and four phylosectors is overwhelming:

On a first provisional count, the four phylosectors account for a total of 161 outer languages with over one million voices each (scale 6 or above), whereas the four geosectors account for only 19 languages of this same demographic importance. On the other hand, the four geosectors account for 337 outer languages which have become extinct during the 20th century, against a total of only 13 for the four phylosectors.

A further striking fact is that the most "dynamic" and demographically important of all the intercontinental *affinities*, presented as the 5=Indo-European phylosector, does not appear to have lost a single *outer language* through extinction of a speech community during the entire 20th century[39].

Much further comparative work, of historical as well as contemporary interest, remains to be completed on the detailed statistical analysis of the Linguasphere Register, in both its present and future updatable forms.

FOR A SYNOPSIS OF THE TEN SECTORS SEE VOLUME TWO, pp.16-35

[39] The extinction in the 20th century of [50=] Gaelg (Manx) on the Isle of Man represents the loss of an *inner language* within the surviving [50=] Gaeilge+Gàidhlig *outer language*, which is still spoken by indigenous speech communities extending from the Northern Hebrides to the south coast of Ireland.

framework of linguasphere sectors and zones (1999-2000)

0=AFRICA *geosector*	sets
00=MANDIC	4
01=SONGHAIC	1
02=SAHARIC	3
03=SUDANIC	2
04=NILOTIC	3
05=EAST-SAHEL *geozone*	16
06=KORDOFANIC	4
07=RIFT-VALLEY *geozone*	4
08=KHOISANIC	2
09=KALAHARI *geozone*	5
2=AUSTRALASIA *geosector*	sets
20=ARAFURA *geozone*	26
21=MAMBERAMO *geozone*	22
22=MADANGIC	23
23=OWALAMIC	11
24=TRANSIRIANIC	22
25=CENDRAWASIH *geozone*	25
26=SEPIK-VALLEY *geozone*	22
27=BISMARCK-SEA *geozone*	26
28=NORTH-AUSTRALIA *geozone*	21
29=TRANSAUSTRALIA *geozone*	25
4=EURASIA *geosector*	sets
40=EUSKARIC	1
41=URALIC	3
42=CAUCASUS *geozone*	3
43=SIBERIA *geozone*	4
44=TRANSASIA *geozone*	3
45=EAST-ASIA *geozone*	3
46=SOUTH-ASIA *geozone*	11
47=DAIC	1
48=MIENIC	1
49=DRAVIDIC	5
6=NORTH-AMERICA *geosector*	sets
60=ARCTIC	1
61=NADENIC	3
62=ALGIC	3
63=SAINT-LAWRENCE *geozone*	2
64=MISSISSIPPI *geozone*	3
65=AZTECIC	1
66=FARWEST *geozone*	26
67=DESERT *geozone*	5
68=GULF *geozone*	8
69=MESO-AMERICA *geozone*	11
8=SOUTH-AMERICA *geosector*	sets
80=CARIBIC	1
81=INTER-OCEAN *geozone*	16
82=ARAWAKIC	2
83=PRE-ANDES *geozone*	20
84=ANDES *geozone*	13
85=CHACO-CONE *geozone*	10
86=MATO-GROSSO *geozone*	16
87=AMAZON *geozone*	23
88=TUPIC	10
89=BAHIA *geozone*	11

1=AFRO-ASIAN *phylosector*	sets
10=TAMAZIC	1
11=COPTIC	1
12=SEMITIC	1
13=BEJIC	1
14=CUSHITIC	7
15=EYASIC	2
16=OMOTIC	6
17=CHARIC	7
18=MANDARIC	9
19=BAUCHIC	8
3=AUSTRONESIAN *phylosector*	sets
30=TAIWANIC	11
31=HESPERONESIC	18
32=MESONESIC	5
33=HALMAYAPENIC	1
34=NEOGUINEIC	7
35=MANUSIC	9
36=SOLOMONIC	6
37=KANAKIC	4
38=WEST-PACIFIC	8
39=TRANSPACIFIC	3
5=INDO-EUROPEAN *phylosector*	sets
50=CELTIC	1
51=ROMANIC	1
52=GERMANIC	1
53=SLAVIC	1
54=BALTIC	1
55=ALBANIC	1
56=HELLENIC	1
57=ARMENIC	1
58=IRANIC	1
59=INDIC	1
7=SINO-INDIAN *phylosector*	sets
70=TIBETIC	1
71=HIMALAYIC	3
72=GARIC	2
73=KUKIC	4
74=MIRIC	1
75=KACHINIC	2
76=RUNGIC	4
77=IRRAWADDIC	3
78=KARENIC	1
79=SINITIC	1
9=TRANSAFRICAN *phylosector*	sets
90=ATLANTIC	16
91=VOLTAIC	9
92=ADAMAWIC	3
93=UBANGIC	2
94=MELIC	2
95=KRUIC	1
96=AFRAMIC	13
97=DELTIC	2
98=BENUIC	11
99=BANTUIC	1

Zones

The second *layer* of classification, indicated by the first plus second digit of all codes, is composed of the one hundred zones listed on table above and on p. 300, representing the most useful referential division of each of the above geosectors and phylosectors into ten parts.

Within each phylosector, the component zones (or *phylozones*) are based on the known linguistic subdivisions of each of the *affinities* (or 'families') concerned, selected subdivisions being either combined or further divided to arrive at a total of ten referential parts. 5=Indo-European, for example, divides readily into ten phylozones, corresponding to so-called "branches" of the Indo-European wider affinity or "family", whereas in the case of 1=Afroasiatic a total of ten phylozones is arrived at by allocating more than one zone to the more complex Chadic "branch" of the Afro-Asiatic intercontinental affinity, representing three actual groupings within that branch.

Within the five geosectors, twenty-two of the fifty component zones are themselves *phylozones*, corresponding to wider or narrower affinities, as in the case of 00=Mandic in Africa, for example, or 41=Uralic in Eurasia. The remaining twenty-eight zones are *geozones*, corresponding to geographic groupings of languages which may (but do not necessarily) share a geo-typological relationship, as in the case of 43=Caucasus or 44=Siberia. These phylozones and geozones are not distributed evenly among the five geosectors, as indicated on the table of sectors and zones above (and in a more detailed format on p.300).

Each *phylozone* (among 50 in the *phylosectors* and 22 in the *geosectors*) covers a single *set* or a *grouping* of 2 or more related sets, or – in certain cases – a *reference area* of sets which share an external unity within the *affinity* of the relevant phylosector, but which do not form a discrete *grouping* within that phylosector (i.e. do not share any *internal unity*). Examples of such *reference areas* within a phylosector are found in the 9=Transafrican phylosector, where no less than five zones are best treated as reference areas: 90=Atlantic, 91=Voltaic, 92=Adamawic, 96=Aframic and 98=Benuic.

Each *geozone* (among 28 in the *geosectors*) covers a *reference area* of 2 or more sets which do not share any known or certain *internal (or external) unity*. It should be emphasised that this categorisation of geozones, in contrast to phylozones, is intended primarily as a warning that all component *sets* of languages within a single geozone should not be assumed to be linguistically related, although some of them may be.[40] The twenty-eight geozones together account for a total of 380 sets, in contrast to the 314 sets included within seventy-two phylozones. They thus account for a majority of the linguistic complexity of the modern linguasphere, although for only a small minority of the world's current population.

In the present English version of the Register, the names given to the one hundred zones are harmonised by the systematic use of the suffix -ic, and the zones of each sector are numbered in approximate geographical sequence, as far as possible from north to south and/or from west to east. 5=Indo-European, for example, is ordered in sequence from 50=Celtic (in northwestern Europe) to 59=Indic (in southern Asia).

The sectors and zones form a consistent system of reference covering the totality of modern languages in the world, to which any past or future system of historical classification may be specifically cross-referenced. A stable framework – or linguistic "workbench" - is thus provided by the Register, on which pieces of the historical jigsaw of distant linguistic relationships can be assembled and re-assembled as necessary. The underlying framework of reference will no longer need to be changed each time a new 'family-tree' of remoter affinities is proposed.

Within each of its one hundred zones of reference, the *Linguasphere Register* provides a scale of alphabetical classification, providing an adjustable assessment of relationships among the languages included in each zone: see Sections 2.3 and 2.4 below.

[40] As in the case of four sets within 46=Southasiatic geozone (46-D to 46-G, parts of the Mon-Khmer affinity)

2.3 Sets, Chains and Nets

The three upper-case (majuscule) letters of the alphabetical codes, e.g. 99-AAA-aaa, represent the (majuscule) letters of the alphabetical codes, e.g. 99-AAA-aaa, represent the upper-case (majuscule) letters of the alphabetical codes, e.g. 99-AAA-aaa, represent the inter-relationship of languages which are known – or assumed – to share *at least a substantial minority of their basic vocabulary*, i.e. in principal 25 to 30%+ of their vocabulary of common human experience (as measured by the use of phonologically related forms with the same meanings, using wherever possible the 200-item comparative *swadesh-list*).

More distant relationships, involving smaller proportions of vocabulary, lie outside the coverage of the alphabetical classification, but are referred to as *affinities* in the presentation of each zone in the Register (including the five *continental* or *intercontinental affinities* used as a basis for the referential phylosectors).

The majuscule code expresses the relative proximity of outer languages in terms of three successive layers of close relationship or *groups*. These successive layers (which may or may not contain more than one component) are designated by the terms *set*, *chain* and *net*, in the order of increasing proportions of basic vocabulary in common among any two or more *outer languages* classified together at each layer.

It should be stressed that the approximate percentages of basic vocabulary cited below - and on the table of layers on p. 297 - are indicative only. Items of vocabulary are the only element of language which can be roughly quantified for comparative purposes, but even in this case their use must be treated with caution. Apart from the problem of the availability of adequate and comparative wordlists, there are also inevitable problems of personal judgement, as involved in deciding when partially similar forms in two or more languages should be treated as items of vocabulary in common between them. The question of whether or not such items may be loanwords is a further issue, although this is of greater relevance in diachronic assessments than in the present synchronic classification of languages.

Sets The establishment of each *set* implies that all included languages are known - or assumed - to share at least a *substantial minority* of their basic vocabulary (in principal 25 to 30% or more), either collectively among every potential pair of languages or at least among a sequence of overlapping pairs of languages. The number of distinct sets within each zone is also a measure of that zone's relative complexity, varying from only one set (as in the case of all zones in sector 5=Indo-European), to a total of over twenty (as in the case of nine of the ten zones in sector 2=Australasia).

Chains The establishment of each *chain* implies that all included languages are known - or assumed - to share *around half or more* of their basic vocabulary (in principal 45 to 50% or more), either collectively among every potential pair of languages or at least among a sequence of overlapping pairs of languages.

Nets The establishment of each *net* implies likewise that all included languages are known - or assumed - to share *a substantial majority* of their basic vocabulary (in principal 65 to 70% or more), either collectively among every potential pair of languages or at least among a sequence of overlapping pairs of languages.

These three successive groups or layers of close relationship – *set, chain* and *net* – are coded respectively by the first, second and third majuscules (upper case letters) of the alphabetical classification, as set out on the next page and on p.297.

All three layers are applied systematically in the classification and coding of all the languages of the world, even where the component language or languages of a closely knit set may form also an identical chain and net.

The isolated outer language of [40] Euskara / Basque, for example, is recorded in the Register successively under set 40-A, chain 40-AA and net 40-AAA Euskera, reflecting the fact that this outer language has no linguistic relatives at the successive layers of net, chain and set within the zone 40=Euskaric, of which it is the only language.

The use of the Latin alphabet for coding inevitably limits each zone to a maximum of 26 sets and each set to a maximum of 26 chains (and so on), although this order of magnitude has in fact held for the classification of the world's languages. At a few points, especially in geosector 2=Australasian, the system has reached its limit, with two sequences of geozones (20= and 21=, and 25= to 27=), almost reaching their maximum of 26 sets per zone.

The *swadesh-list* of 200 items of basic vocabulary established by Maurice Swadesh in the 1950's (with a shorter form of 100 items) has already been used for comparative purposes in many parts of the world, and the relevant percentages have been used in the Linguasphere Register wherever possible for the establishment of sets, chains and nets. In other cases, the three layers have been applied in the manner which corresponds most conveniently or closely to the reported or presumed relationships among the relevant languages.

A major project for the future, and one to which potentially thousands of local observers will be able to contribute, is the collection of the same *swadesh-list* for as many languages and dialects as possible throughout the world, in order that they may be linked to the continually updated Linguasphere Register. It will then be possible to maintain the present alphabetical classification as a standard form of comparative assessment for close and immediate linguistic relationships in any part of the world. The acceptance of Maurice Swadesh's useful comparative list does not, however, indicate any confirmation of his much debated thesis that the basic vocabulary of all languages changes at some sort of fixed rate, like carbon-14, and may therefore be used as a method for prehistoric dating.

The use being made here of the *swadesh-list* is no more than as an approximate measure of the lexical distance between related languages spoken by present-day communities. Completion of the list can be undertaken by teachers and children in schools all over the world, and it is a happy coincidence that the title "Swadesh" should correspond also to the phrase meaning "our country" in one of the world's major languages, i.e. [59] Hindi+Urdu *swa desh*.

The sets, chains and nets constitute the third to fifth layers of classification and coding in the Linguasphere Register, as indicated by the majuscule (upper case) letters of all alphabetical codes recorded in column 1.

Alphabetical (upper-case) scale of linguistic proximity > SEE FULL TABLE on p.297
Each *set* comprises two successive layers of close relationship:
chain (within each *set*) and *net* (within each *chain*) = upper-case alphabetical code (-AAA-)

+ an **alphabetical code** comprising three **upper-case (majuscule) letters**	marking ➢ three **layers of close relationship**	***ideally*,** the following minimum of basic vocabulary may be shared by languages in the same *set*, *chain* or *net*	TOTALS
99-**A**	**SET**	**substantial minority** (*say* 25-30%+)	694
99-**AA**	CHAIN	**intermediate proportion**	1,410
99-**AAA**	**NET**	**substantial majority** (*say* 65-70%+)	2,694

2.4 Outer Languages, Inner languages and Dialects

Within each zone, and within each of its successive sets, chains and nets, every component language (in the broadest sense of an individual "language") is classified in terms of three successive layers of immediate relationship - outer language, inner language and dialect. Where the countable noun "language" is used in this discussion, it refers potentially to the successive layers of outer language and inner language.

No form of speech exists in isolation from all others, and individual languages exist not only for purposes of communication but also to mark social distances among speakers of different, including closely related, forms of speech.

Where those forms are particularly close, and more or less inter-intelligible, they are treated in the Linguasphere Register as component inner languages within the same outer language, each outer language being represented by the first lower case (minuscule) letter of the proximity scale and its component inner languages by the second. The optional layer of dialect is reserved for relatively minor variations within an inner language, usually dependent on geographical location, and is coded - wherever known to exist - by means of a third minuscule.

The three lower-case letters of the alphabetical code thus represent a scale of immediate relationship among spoken and written forms of a language which are *largely inter-intelligible*.

Unlike the layers of close relationship, no attempt is made to assign a scale of shared vocabulary to each layer of immediate relationship or *idiom*, since approximate percentages of common vocabulary – useful as a yard-stick between say 20% and 80% - are of little interest at either extreme of the statistical scale. (Just as percentages of common vocabulary approaching 0% leave too much scope for chance resemblances, so figures approaching 100% are too much affected by the ready flow of vocabulary among immediately related forms of speech.)

The only scale of terms available in English to cover a range of immediate linguistic relationships is in the dichotomy of "language" versus "dialect", unlike the trichotomy available, for example, in [51=] Français (French) or [52=] Deutsch (German): *langue* or *Sprache*; *dialecte* or *Dialekt*; *parler* or *Mundart*. The Register therefore creates a distinction between "outer language" and "inner language" in English, although it cannot be stressed enough that these are relative terms and cannot be treated as absolutes. [41] This distinction also corresponds approximately to the official dichotomy observed in India between 18 major "scheduled languages" and their 75 component "mother-tongues" (of more than 10,000 speakers each).[42] For the equation of the Register's outer languages with India's scheduled languages, see zones 49, 59 and 73.

The English term "idiom", on the other hand, is used where necessary in the Register to describe any form of speech assigned to one or more of these three layers of immediate relationship, outer language, inner language or dialect. It thus corresponds to the general use of the term "group", on a wider scale, to describe any of the three layers of close relationship.

[41] In the first (1997) preview edition of the Register, the layers of *outer language* and *inner language* were distinguished by a re-defined use of the terms "tongue" and "language", but discussion with colleagues made it clear that this terminological innovation would create more difficulties than it solved (especially since it is the reverse of Indian usage).

[42] "scheduled languages" as included in the 8th schedule to the Constitution of India, amended in 1992, with "mother-tongues" of more than 10,000 speakers as listed in Vijayanunni 1997.

Alphabetical (upper-case) scale of linguistic proximity > SEE FULL TABLE on p.297			
Each *net* comprises two or three successive layers of immediate relationship: *outer language*, *inner language* and (optionally) *dialect* = lower-case alphabetical code (-aaa)			
+ two or three **lower-case (miniscule) letters**	marking ➢two or three **layers of immediate relationship**	up to *three* layers of relative proximity composed of largely inter-intelligible spoken (and/or written) *idioms*	TOTALS
99-AAA-**a**	**Outer language**	= **basic demographic unit**	4,994
99-AAA-**aa**	**inner language**	= **basic unit of classification**	13,840
99-AAA-**aaa**	dialect (*as required*)	= **local, social or written variety**	(< 8,881)
(uncoded)	(voice)	= **the total linguistic repertoire and competance of each person in any language or languages**	6,000,000,000

The Register remains as flexible as possible in its definition of the concept of individual languages, and in its use of this new distinction between *inner* and *outer languages*. It is not a question of forcing all idioms into a rigid mould, but of providing greater flexibility in the analysis of the complex and varied relationships existing between the different forms of so many individual languages, both spoken and written. As applied in the Register, the term "dialect" (see below) refers especially to the distinctive pronunciation of a particular language as spoken (or written) in a particular locality or region, or within a particular social group, with normally some characteristic items of vocabulary or morphology. More radical differences of language, on the other hand, which have sometimes fallen under the extended use of the term "dialect", such as divergences among many traditional, localised varieties of [52=] Deutsch / German and Nederlands / Dutch, or of the [51=] Romance languages, can now be treated more appropriately as "inner languages".

In the same way, it is now possible to draw clearer attention to the different categories of "inner" language which frequently exist side by side within "outer" languages characterised by a strong written as well as spoken tradition, as is the case with many official national languages in the world. The new trichotomy of layers of communication also makes it easier to deal consistently with cases in which the recognition of one or more languages in a continuum depends on political circumstances, such as the abandoned "unity" of [59=] Hindustani after 1947 in partitioned India and of [53=] Srpsko-hrvatski (Serbo-Croat) after 1992 in ex-Yugoslavia. In both these cases, as far as the Register is concerned, one is dealing with an outer language composed of inner languages (each with its dialects), both before and after the events of political separation.

The Register records instances of a regular relationship in the 20th century between the following categories of written and spoken language, as apparent in such languages as [45=] Nihon-go (Japanese), [51=] Français (French), [52=] Deutsch (German) and [52=] English:

- the *formal*, official variety, in which speech is modelled on (and often read from) the written word;
- the *traditional* spoken varieties, associated with specific localities or areas;
- the *generalised* or mainstream colloquial variety of the formal language, encouraged by modern education and television, which has often largely replaced (but been locally influenced by) the traditional varieties of the same language;
- and, very often also, the vigorous development or survival of one or more *counterstream* varieties, associated with less privileged urban populations including alienated youth, especially in or near capital cities.

The parallel use of two (or more) of these varieties by the same speaker - i.e. co-existing within the same voice - has been described as “diglossia”, although there has been a variety of redefinitions of this term since it was first proposed by Ferguson[43], not to speak of the proliferation of other terminology in this area, including “register”, “code”, “speech style”, “biglossia”, “triglossia” and “multiglossia”. The general use of the term *diglossia* to describe “any functional compartmentalisation by a society of its linguistic resources” seems now well established and is probably the most useful, if one stresses its importance as a synchronic phenomenon whatever its historical or cultural background.[44]

It is worth noting at this point that the universal and inherent characteristics of speech itself provide the best parameters for the observation and analysis of individual languages, rather than categories based on specific social, cultural and political events or priorities. Looking back from the end of the 20th century at the sociolinguistic discussions of the post-colonial period, only twenty to forty years ago, [45] one is struck at how rapidly the whole geopolitical and telecommunicational context of the linguasphere has since changed (or rather “telecommunicational and geopolitical”, the former having been the driving force behind the second).

The practical application in the Register of the alphabetical scale and of the triad of terms *outer language*, *inner language* and *dialect* may be illustrated by reference to the well-known set of [52=] Norsk+Frysk (Germanic) languages within the 52=Germanic phylozone, and to the complex case of [52=] Deutsch+Nederlands in particular. The following lines may be read in conjunction with the relevant section of the Register (Volume Two, pp. 411-39).

52=Germanic zone

- The languages of the Indo-European *intercontinental affinity*, covered by the initial digit of *phylosector* 5=Indo-European, are further subdivided and coded in the Register so that all Germanic languages are covered by the initial digits of *phylozone* 52=Germanic, comprising a single *set*.
- For historical or diachronic purposes, the Germanic languages have been traditionally classified as North Germanic (Scandinavian, including Icelandic), West Germanic (German, Dutch, Frisian and English) and East Germanic (the extinct language of Gothic).
- For the purposes of contemporary or synchronic purposes, however, the Germanic set of languages (52-A Frysk+Norsk or Frisian+Norwegian) is subdivided in the Register into three current *chains*: Nordic (52-AA Norsk+Svenska), Anglic (52-AB English+Anglo-Creole) and Continental West Germanic (52-AC Frysk+Deutsch or Frisian+German); plus one further *chain* to cover East Germanic (52-AD 'Gothic'), extinct since 16th century.
- The Continental West Germanic *chain* (52-AC Frysk+Deutsch) is subdivided into two current *nets*: Frisian (52-ACA Frysk+Frasch) and German+Dutch (52-ACB Deutsch+Nederlands).
- The complex continuum of German+Dutch (52-ACB Deutsch+Nederlands) *idioms* is classified as a sequence of nine closely related *outer languages* (52-ACB-a to 52-ACB-i)
- Within the Middle German *outer language* (52-ACB-d Deutsch-C.), for example, Luxemburgish (52-ACB-db Letzebürgesch) is treated in the Register as an *inner language*, together with other traditional Franconian *inner languages* in Germany and eastern France. Formal German (52-ACB-dl Hochdeutsch-F.) is likewise treated as an *inner language* within Middle German.

[43] See Ferguson 1959
[44] See Khubchandani 1997 pp.132-148.
[45] See for example Fishman, Ferguson & Das Gupta 1968 and Whiteley 1971.

- Within the Upper German *outer language* (52-ACB-e Deutsch-S.), Alsatian is treated as an *inner language* (52-ACB-el Elsässerditsch), together with other traditional Upper German *inner languages* like Swabian (52-ACB-ej-Schwäbisch) or Carinthian (52-ACB-ee Kärntnerisch). As in the case of Luxemburgish, Swabian and Carinthian, Alsatian is itself composed of several localised *dialects* (e.g. 52-ACB-elb Strossburjerditsch or Strasbourgeois, i.e. Strasburg urban dialect).
- Swiss German (52-ACB-f Schwytzertütsch), on the other hand, is treated as an *outer language* in its own right, with no less than eleven constituent *inner languages* and over forty *dialects* (reflecting the linguistic fragmentation typical of mountain valleys).
- A wide variety of expatriate German idioms, spoken in every continent, are classified together as a *notional outer language* (52-ACB-h Auswanderungsdeutsch), comprising no less than twenty-two *inner languages.*

Questions of 'dialect'

The traditional dichotomy of “language(s)” and “dialect(s)”, in the consideration of immediately related *idioms*, has frequently implied that a dialect is not only subordinate to a language but also in some way inferior in quality or correctness. The term “dialect” has also been applied sometimes in a deprecatory sense to specific languages, implying that they are unwritten or in some way “undeveloped”. As a result, in a discussion of African languages over thirty years ago, it was proposed that the use of the term “dialect” be abandoned, [46] because it had been so ambiguously employed on its long journey from Ancient Greece (in its ultimate derivation from Greek *dialektos*). But the place of the modern word in English and French and other languages is too well established for this proposal to have been taken seriously.

Instead, the Register now retains the use of the term “dialect” to describe the last of its three layers of immediate relationship, but only in situations where three rather than two layers are preferable in the description of a particular linguistic situation, or where sufficient data are available to permit a presentation to this degree of detail. In very general terms, an attempt is made to limit the use of the term “dialect” to the narrower applications of that term, especially differences relating largely to phonology and phonetics, or “accent” , with small variations of vocabulary and/or syntax. One may define *accent* in this context as the pronunciation of an idiom, characteristic of the geographical, linguistic and/or social background of the voice or voices concerned, including those distinctive features of pronunciation which have been carried over from one idiom to another, within the personal history of an individual voice or the inter-generational history of a speech community. Such history has often involved the replacement of one idiom by another or the absorption of a traditional idiom into a more generalised idiom of the same outer language (with an “accent” being the last vestige of the idiom replaced).

Needless to say, the use of the term “dialect” in the *Register* should not be confused with its usage in the dual language/dialect terminology in many of the *Register*’s sources. In the past, even among linguists, there has been great variation in applying the traditional dichotomy of language versus dialect, and yet the naïve question “is it a language or a dialect?” is posed or implied sometimes even in professional sources, as though there were an absolute distinction between the two.

In Europe, the tendency has been for “dialect” to be applied more narrowly to varieties of well documented languages, in situations which have been treated in the Register under the layers of “dialect” or “inner language”, but the word “dialect” has been employed on the whole more loosely in primary and secondary sources on languages in other parts of the world.

[46] Dalby 1966, p.171-173

For this reason, it has often appeared justified to treat so-called "dialects" among, say, Asian and Oceanian languages, as *inner languages* or even *outer languages* for the purposes of the Register. It is also clear that the enumeration of localised dialects in the narrower "European" sense has never been undertaken in some parts of the world, and that the *Register* will become considerably more detailed once this has been done. Under sector 5=Indo-European, for example, the listing of inner languages and dialects among European languages (see zones 50=Celtic to 53=Slavic) is more detailed than among those of South Asia (59=Indic, in particular) and it is evident that more detailed surveying of localised "dialects" in that region – from rural area to rural area, and often from caste to caste - will lead to a considerable increase in the complexity of that sector. Even so, over 50% of the inner languages in the 5=Indo-European sector are already subdivided into two or more dialects, in contrast to less than 2% of the languages in the 2=Australasian geosector, where individual languages are spoken by an average number of voices lower than in any other sector, and where documentation on minor variations is in any case sparse.

It must also be stated that the listing of dialects within an inner language does not necessarily imply that all voices speaking that inner language can readily be classified under any one of those listed dialects, which may not yet account for the totality of the language in question.

On the subject of labelling "dialects", and of distinctions between immediately related idioms in general, it should be recognised that one is dealing with the fluidity of language at its most extreme. Traditional dialectology runs into the ground at this very point, as it becomes evident that any feature of any language is liable to have its own independent boundary or "isogloss". This is where the continuum of the linguasphere becomes most obvious, and where even individual voices are liable to be mobile between adjacent idioms.

Specific phonological features or "diagnostic" forms have often been employed in dialectological studies to establish more precise boundaries among neighbouring idioms, but their relevance often varies along the line of the chosen isogloss, according to the number of other features which happen to follow a similar course for parts of that line. [47]

Boundaries of close relationship

In fact, the most frequently useful feature in determining boundaries among related idioms is the natural lie of the land. Valleys and plains allow the free movement of languages and linguistic features, but it is in mountain valleys that the distinctions among neighbouring languages become most obvious. Watersheds and mountain ridges are among the greatest dividers of language.

Serious attempts have also been made to establish scales for the measurement of relative intelligibility among immediately related idioms,[48] but here also it becomes difficult to establish consistently firm ground. It is generally accepted that degrees of inter-intelligibility are influenced by a wide variety of extraneous factors, including the acuity of hearing, relative intelligence and previous linguistic experience of the person whose comprehension is being judged, as well as reciprocal feelings among speakers of immediately related languages, and the subject-matter actually under discussion.

It must be once again stressed that there can be no absolute definition of any of the three terms *outer language*, *language* or *dialect*, since their application depends not only on degrees of relative inter-intelligibility but also the relative complexity of any given linguistic environment. Thus although both [52=] Deutsch (German) and [79=] Han-yu (Chinese), in all their respective varieties, are presented in terms of these three layers of immediate relationship, it is clear that Chinese embraces an even wider range of internal variety than does German.

[47] The classic example is that of the so-called "Rhenish fan" (*Rheinischer Fächer*) in Germany, where there is a "fanning-out" of the distinctive isoglosses marking the northern limits of the High German sound-shift. See Barbour & Stevenson 1990, pp.87-88; Russ 1990, pp.136ff.

[48] See, for example, Casad 1974.

Another significant feature of the linguasphere, clearly apparent from the Register, is the way in which small isolated languages more often than not contain deep divisions within themselves, as though the voices of each internal community need to be able to define themselves in terms of their closeness to voices of at least one other related, but distinct, form of speech.

In Europe, for example, such internal division is found within such isolated and relatively "endangered" ethno-linguistic entities as:

> [40=] Euskara or Basque (with nine inner languages and over thirty dialects, extending across the Pyrenees from Vizcaya to the Pyrénées-Atlantiques);
>
> [41=] Saame or 'Lappish' (with three outer languages, nine inner languages and fourteen further dialects spoken across northern Scandinavia and into Russia);
>
> [50=] Gaeilge+Gàidhlig or Gaelic (with a sequence of five inner languages and up to thirty dialects spoken from the south coast of Ireland to the northern Hebridean islands of Scotland);
>
> [51=] Rumantsch+Grischun with Nones+Cadorino or Ladin (with two outer languages and eight inner languages spoken in the Alpine valleys of Switzerland and Italy);
>
> [52=] Frysk+Frasch or Frisian (with four outer languages, ten inner languages and double that number of dialects - spoken on or near the North Sea coast and islands, from the Netherlands to Schleswig-Holstein); and
>
> [53=] Serbska+Serbšćina or Sorbian (with two inner languages and a dozen dialects, spoken in a small area south of Berlin).

For an exemplification of the way in which the Linguasphere classification has been applied to the detailed configuration of languages in other areas of the world, see any of the individual zones of the Register, and particularly the discussion in many of the headings to individual zones. A synopsis of the zones within each sector is presented in Volume Two (pp. 16-35).

Finally, it should be mentioned that a potential by-product of the Register's listing of inner languages within outer languages relates to the planning and application of machine-translation. It is obvious that the problems confronting the design of a bilingual translation program are considerably less when the program is *translingual* (i.e. covering languages classified within the same net), and especially when the two languages are immediate enough in relationship to be treated within the same *outer language*. In other words, help in the development of published literature in a previously unwritten or little written language can be expected to benefit from the machine-translated version of literature already existing elsewhere in the same outer language, depending of course on the availability of a programmer with the necessary linguistic knowledge. Research in Mexico has already led to the creation of computer programs which can move a text largely by substitution between two very closely or immediately related languages, in contrast to the general reformulation required by translations between unrelated or distantly related languages.[49]

Endangered languages

An *endangered* speech community is one whose common language is in demographic decline, the voices of that language being no longer replaced in proportion to those dying. The present size of an endangered speech community is not a factor in its definition as "endangered".

> *An endangered language may be defined as an outer language,*
> *the voices of which are no longer being replaced in a balanced proportion.*

Publicity has been generated in recent years for the defence of "endangered" languages and their communities of voices, and this is in every way to be welcomed. The most reliable indicator that a

[49] For early development in this field, see Weber, McConnel et al. 1990, based on work by Bill Mann and David Weber in the 1970's and subsequent use in the field of [84] Quechua languages in the 1980's.

speech community is in danger of extinction is when the annual number of infant voices acquiring the language of that community is substantially less than the number of voices dying in the same year. This criterion applies to all languages, whatever their present total of voices, but a *seriously* endangered language can be regarded as one with present participation of below 1,000 voices, which will be extinct or moribund (less than 100 voices) within a generation (30 years) if present trends continue.

On the other hand, it can be frequently observed that there is a point in the numerical decline of a speech community which has a strong effect on certain individual voices, who take up an active role in the subsequent defence of their community's language, which may not actually be their mothertongue. These voices often succeed in leading a movement of linguistic "revival" within their community as a whole (for the majority of whom the language may already have become a symbolic marker of their community, rather than a living means of communication). For such leading voices, the language concerned is not always a primary language, as in the not infrequent case of a "missed generation", where children have acquired from their grandparents a language which had not been learned or retained by their intervening parentsSuch movements for the strengthening of an endangered language involve not only political and educational action, but – perhaps most importantly - the recording and creation of literature in that language.

Notable modern examples of movements to arrest the decline of languages in western Europe[50] include all those listed above as examples of internally divided ethno-linguistic entities (under Boundaries of close relationship): [40=] Euskara (Basque), [41=] Saame (Lappish), [50=] Gaelic, [51=] Romantsch with Ladin, [52=] Frysk+Frasch (Frisian) and [53=] Sorbian

Although some past assessments of the proportion of "endangered" languages in the world have been based on absolute numbers of voices, the overview of the linguasphere provided by the Register suggests that the viability or "survival" of any form of language is less a question of its numbers of voices as of its overall environment - geographic and social, linguistic and educational, cultural and political, economic and technological, religious and military.

A recent Unesco publication from Australia[51] has rightly drawn attention to the fact that "the question of large or small numbers of voices [of a language] is quite relative", an Australian language with more than 1,000 voices being regarded as "large", but a language in India with 10,000 or more voices as "small". The same publication, however, is one of a number of recent sources which appear to seriously overestimate the number of languages currently in danger.

At the opening of his presentation, the editor (op.cit. p.1-2) writes:

> *"According to our estimates there are between 5,000 and 6,000 languages spoken in the world today – most of them in several dialects…… the last three hundred or more years have seen a dramatic increase in the death and disappearance of languages, at a steadily increasing rate in many parts of the world leading to a situation in which today 3,000 or more languages or so still spoken, are now endangered, seriously endangered or dying, with many other still viable languages already showing signs of being endangered… Basically, any language of a community which is not learned any more by children, or at least by a large part of the children of that community (say at least 30 per cent) should be regarded as "endangered" or at least "potentially endangered"…*

This statement wrongly implies that at least half, and perhaps more, of the modern languages of the world are no longer being learned by a third or more of the children of each speech community concerned. Fortunately, this estimate is a considerable exaggeration, as can be seen by comparison with the proportion of outer languages – less than 9% - recorded in the *Register* as having become extinct during the 20th century (i.e. with a demoscale of 0 in column 5).

[50] See Stephens 1978.
[51] Wurm & Baumann 1996, p.2.

It can be argued that some languages may have become extinct in less accessible areas of the world, without the fact having yet been recorded, but this is unlikely to raise the figure beyond 10%.

It must also be pointed out that a startling number of the known cases of language-extinction during the 20th century have occurred in one country, Australia, which – according to preliminary calculations based on the Register - accounts for *just over one third of all cases of extinction of outer languages recorded for the world as a whole*.

The Register also highlights the fact that that the majority of languages which have become extinct during recent centuries, or which are in danger of becoming extinct in the 21st, are languages spoken or formerly spoken by peoples whose original way of life has been destroyed, primarily semi-nomadic and/or hunter, gatherer and fishing communities. Australia is the greatest modern hecatomb of such languages, followed by Amazonia and North America, the Arctic and parts of Southwest and Northeast Africa.

During humankind's long paleolithic era of slow expansion and development (over 90% of human life on earth), all languages were necessarily dependent on the survival of small and fragile communities of hunters, gatherers and fishers. The progressive elimination of such communities during recent millennia, however, has been a result of humankind's pursuit of development - first the agricultural and pastoral conquest of the physical and biological environment, and more recently its industrial and physical conquest and cumulative pollution.

It is salutary to reflect that the ecological balance between humankind and the natural environment; as maintained by traditional 'Aborigine' communities in Australia, may be taken to represent a higher level of 'civilisation' than that of their conquerors from the British Isles… if success in civilisation is measured by its capacity to prolong; rather than shorten, human occupation of this planet.

The adoption by hunting and gathering communities, or by their scattered descendants, of a form of the language of the "conquerors" of their traditional environment is illustrated by the Pygmy and "Twa" populations of central Africa and by the "Agta" groups in the Philippines (submerged in the pre-colonial era by largely Bantu-speaking and by Austronesian-speaking agriculturalists), and in the era of colonial expansion by the linguistic and cultural absorption of many Australasian, Arctic and Amerindian communities, who now speak varieties of European languages, notably English.[52]

It is not surprising, therefore, that some of the strongest warnings about languages in peril should have come from scholars working in areas like Alaska and Australia.[53] That such warnings may have exaggerated the numerical extent of the tragedy must not be allowed to draw attention away from their reality, and from the real needs of the often isolated communities concerned.

Languages still spoken by hunting and gathering or proto-agricultural communities today are typically spoken by a few hundred voices or less, and presumably always have been. The most important area of survival of such communities is the island of New Guinea, the unique linguistic complexity of which may be all that remains of a lost world, a *paleolinguasphere* which was far more linguistically diverse than to-day's linguasphere. The linguistic survival of any such tiny community depends quite simply on its social and economic survival. It may be broken up or destroyed, either by the traditional dangers of war or famine or disease[54] or environmental disaster [55], or by the more recent dangers of encroaching "development".

---

[52] See entries in the *Index* below, for the linguistically absorbed "Pygmy / Pygmoid / Twa" and "Agta" communities (now speaking languages in the 03=, 07= and 99= phylozones, and in the 31= phylozone); and see geosectors 2=, 4= (zone 43=), 6= and 8= for relevant Australasian, Arctic and Amerindian communities.

[53] Notably by Professor Michael Krauss and Professor Stephen A.Wurm.

[54] Such as the speech-communities of northeastern Africa, some of which may have been totally destroyed in the man-made and natural disasters of the late 20th century.

[55] Such as the 34-Sissano speech community (34-BAA-b) overwhelmed in 1998 by a tidal wave in northern New Guinea.

Short of confining the surviving voices of such communities to an anthropological “reserve”, however, and denying their children access to the benefits other children enjoy, there seems little one can do to arrest the demise of tiny speech-communities under pressure, apart from opposing the more rapine aspects of development and from recording as much as of their present speech and heritage as one can. Apart from the social protection and support of the individuals and communities concerned, a major transnational cultural priority should be to organise the extensive video-filming of speakers of all languages in immediate peril, and to sub-title such video-recordings in an appropriate international language.

A worldwide archive of this endangered heritage would not only be a valuable addition to the permanent human record but would also provide a base for reviving a subsequently extinct language, if the descendants of its last speakers one day wished to do so. As children from such communities are scholarised, it would seem appropriate to assign the task of video-recording their own languages and cultures to individual teachers and their classes, together with responsibility for preparing sub-titles in a transnational language. The multiple educational benefits of such a programme, if it could be financed by public or private agencies, are obvious.

Apart from the languages of semi-nomadic and/or hunting-gathering communities, there does not seem to be evidence to suggest that an important proportion of all the world’s languages are in danger of being extinguished by English or by any other widespread language in the immediate future. What does seem to be certain, however, is that some *inner languages* and *dialects* are declining in use as a result of pressure from a socially 'preferred' standard within the same *outer language* or *net* of closely related outer languages. The decline in the use of localised 51=Romance languages in Europe under pressure from standardised national varieties is a visible example: Walloon or Norman under pressure from French, Aragonese under pressure from Spanish, Lombard or Ligurian under pressure from Italian, etc.

This is not only true of major national and international languages but also of so-called ‘‘minority’’ languages. A striking image is that of an elderly lady in North Wales, speaking a rural dialect of [50=] Cymraeg (Welsh), who regrets that she does not “speak proper Welsh, like they do on the telly”, or of a farmer in South Wales who speaks with tears in his eyes of the only time he heard his own local dialect on TV - “our Welsh” - when the voices “seemed to come out of the box for the first time” and sit with him in his own parlour.

Similar cases can be quoted for other minority languages in Europe, such as [40=] Basque or [50=] Gaelic or [51=] Catalan, and it is most important that those who fight for the rights of speakers of their own languages against the power and status of state-languages should not forget the equal rights of speakers of localised varieties of their own languages, often threatened by the standardised variety. A most positive development in that respect is the way in which the francophone community in Belgium encourages the development of every variety of minority Romance language, not only in Belgium itself but anywhere in the world - notably by the publication of a literary journal, which publishes poems and texts in any Romance idiom «of less expansion».[56] Similarly, in educational texts published on and in the Provençal language, examples are provided of dialects of that language in different parts of Provence and surrounding areas, school-children being encouraged to write down their own local versions of the sample texts wherever these are at variance with those provided.[57]

A further aspect of the question “what is an endangered language ?” relates to the fact that very few languages exist in total isolation from any other surviving linguistic relative. In the case of a large proportion of the languages of the world, including the examples just discussed, we are not dealing with absolute entities but with points on a sliding scale of communication, be it 51=Romance, 53=Slavic, 59= Indic or [44=] Turkic or 79=Sinitic or 99=Bantuic, or any one of the many smaller chains and nets of closely related languages.

---

[56] *micRomania: Littératures en langues romanes de moindre expansion* (ed.Jean-Luc Fauconnier; rue de Namur 600, Châtelet, Belgium).

[57] See Vouland 1988.

Only a small minority of the world's speech-communities are in complete linguistic isolation, and although the loss of any community's form of speech is regrettable, it is clear that the so-called ''death'' or ''survival'' of an individual language is not something cut and dried like the extinction or preservation of a zoological species, as has sometimes been asserted. It is more realistic, and certainly less emotive, to consider that the natural extinction of a speech community – as opposed to its physical destruction by human or natural means - is an aspect of the general development of the linguasphere, with its continuing currents and counter-currents of convergence, divergence and submersion.

As already argued, every speech community deserves to have the opportunity of preserving its own form of speech, whatever its relationship to the language of any other community. From the point of view of preserving humankind's heritage of linguistic diversity, however, priority should clearly be given to the careful recording of endangered languages which are not closely related to any other surviving language. It is for this reason that the admittedly emotive symbol of a skull and cross-bones ι is employed in the Register only for complete *nets* of languages which have become extinct during the 20th century.

The employment of the same symbol ι for complete *chains* and *sets* which have become extinct, means that the relative degree of loss in the overall diversity of the linguasphere is measured by the number of these symbols which occur together (one for a net, two for a chain, and three for a set). Through the use of this symbol of extinction, even a superficial glance at the pages of the Register demonstrates the degree to which the solitary geozone 29=Transaustralia (covering most of Australia) accounts for a major proportion of the total reduction of linguistic diversity in the world during the century now closing.

By definition, however, the numbers of speakers of *seriously* endangered languages – many still in Australia - are very small, and represent only a fraction of 1% of the world's population. The total of *all* endangered languages remains to be assessed, following the above definition, but – as already pointed out - is substantially less than has sometimes been claimed.

It would be unfortunate therefore, in human as opposed to linguistic terms, if a legitimate concern for the defence of seriously endangered languages were to be used as an excuse for not devoting major investment to the communicational and educational development of the intermediate languages spoken by a much more important percentage of humankind.

In this respect, it is worth reminding ourselves that some of the impetus behind the movement to save "endangered" languages – and to unintentionally exaggerate their number – stems from the pastoral romanticism of the 19th century, as Susan Gal has observed:[58]

"Announcing the extinction of cultures, languages, and dialects at the moment they are first described by outsiders has been a rhetorical construct central to Western ethnography… It is also a central thread in European dialectology and national folklore, the source of much early evidence about language death. Practitioners in these disciplines looked to the countryside for archaic, unchanging and therefore authentic cultural elements to use in defining national cultures and buttressing nationalism… All of these scholarly enterprises are within the Western literary tradition of the "pastoral", a rhetorical convention which continually looks back, often nostalgically and for moral guidance, to a lost but supposedly more pristine, rural, homogeneous, and authentic past.[59]"

[58] Gal 1989, p.315-316.
[59] citing R,Williams, "Base and superstructure in Marxist cultural theory", *New Left Review*, 87, 1973, pp.3-16

2.5 Linguistic Nomenclature

The symbol # introduces any note on nomenclature in column 3.

Among the first priorities of the Linguasphere Register has been the need to establish a firm basis of unambiguous reference names in the Latin (Roman) script, rather than to seek to provide a precise phonemic or phonetic transcription of each name, or even a precise transcription of the relevant script. It is proposed that the Linguasphere Register be subsequently extended to include also citations of reference names in other scripts, for the languages concerned, and where necessary in the International Phonetic Alphabet or equivalent.[60]

- Column 2 of the Register contains the chosen *reference name* for each layer of classification, from sector to dialect; wherever possible, at the layers of languages and dialects, this is the *autonym* or 'own name' used within the language concerned.

- Column 3 contains aliases or 'other names', including alternative reference names (together with other information, as discussed in the following section Categories of Data).

For the conventions covering the indexing of reference names and aliases, please see the introductory note to the Index in this volume (p. 121).

Typography of Reference Names in column 2 of the Register

When listed as entries in column 2, all reference names for layers wider than an outer language (from sector to net) are printed in capitals, whereas all names for layers of immediate relationship (from outer language to dialect) are printed in lower-case letters. The reference names of outer languages are distinguished by an upper-case initial, as opposed to a lower-case initial for inner languages and dialects. The reference names of dialects (the only optional layer in the system) are in smaller face than all other reference names in column 2, and are also the only reference names not printed in bold in that column.

These typographical conventions are unnecessary outside column 2 and the Index, and do not apply, for example, to citations of language-names in column 3 (where all aliases begin with a lower-case initial, and where citations of language-names in the notes require no specific typographical convention).

Linguistic prefixes and/or suffixes included as part of a reference name (e.g. language-prefixes in the [99=] Bantu set) are printed in normal type, even where the rest of the relevant name is in bold . They may be omitted in citations of a reference name, provided their omission does not lead to ambiguity (as verifiable from the Index). The *linguasphere key* of the Swahili language, for example, may thus be cited either using the full indigenous form of its reference name as [99=] Kiswahili, or using its abbreviated ('English') form [99=] Swahili.

Hyphens have been used extensively in column 2, to indicate the known or apparent lexical and morphological boundaries within a reference name, and also to ensure that each reference name is held together as a typographic unit.

---

[60] In Mann & Dalby 1987 (see pp.206-218 “Writing African languages”), reference names for African languages were transcribed in the expanded character-set of the African Reference Alphabet. Such a transcription has been considered too complex to serve as part of a global system of referential names, but thought is currently being given to the adaptation of this phonemic alphabet (and of the related International Niamey Keyboard, *op.cit.* p. 217-8) to global transcriptions. It has the advantage of being directly linked to the standard Latin or Roman alphabet (on a 2 letters to 1 letter basis) and also of being closely related to the International Phonetic Alphabet . The International Niamey Keyboard (Clavier international de Niamey) is being promoted by the Observatoire Linguistique as the transnational *Linguasphere Alphabet*, not as a replacement for other scripts but as a metascript enabling them to be transliterated into a common transnational form whenever required.

One of three symbols may be prefixed or suffixed to a reference name in column 2 (replaced in two cases, in the online edition, by a raised Greek capital letter), as follows:

- a book symbol (🕮 > represented online by the raised Greek capital ᴮ for βιβλίο “book”) prefixed to a reference name indicates that that idiom has been “read only” (silently or aloud) during the 20th century, i.e. is a written historical source preserved from earlier centuries;

- a writing hand (✍ > represented by the raised Greek capital ᴦ for γραφή “writing”) prefixed to a reference name indicates that that the form and/or content of that idiom, when spoken, are modelled on the written (normally standardised) word;

- a raised star (*) suffixed to a reference name indicates that entry in the Register is provisional and remains to be confirmed.

Within the alphabetical scale of linguistic proximity, reference names (from set to dialect) have been selected and standardised on the following principles:

- reference names are standardised in the basic Latin script of 26 letters, wherever possible in a form already established in that script;

- reference names may include accents and diacritics, but no supplementary (non-Latin or 'phonetic') letters

- a reference name with one or more accents or diacritics remains unique and unambiguous within its zone, even if the accents or diacritics are omitted;

- wherever possible, reference names are autonyms ('own names') of languages and/or peoples (see above)

- for reference names covering more than one idiom, combinations of two language-names are preferred to artificial cover names, especially wherever such cover names are based on only one of the component language-names;

- every reference name is unique and unambiguous within its own zone (but may be repeated for two or more successive layers, if their content is identical);

- where the same linguistic name covers two or more non-identical cells within the same zone, unique reference names may be created by the use of standardised suffixes (see below);

- the parts of a compound reference name are either hyphenated or written without a break;

- a normal hyphen is replaced with an additive hyphen (+) whenever the link is between two separate parts.

A series of capitalised suffixes is employed, for layers of immediate relationship, to distinguish reference names which would otherwise be identical in the same zone. This includes the distinction of languages and dialects by points of the compass or by an aspect of their usage. The suffixes are as follows:

- (directional suffixes, in columns 2 & 3) -N north(ern); -E east(ern); -S south(ern); -W west(ern); -C central; plus combinations, eg -NW northwest; -CW west central, etc

- (other suffixes, in column 2) -A "proper" (name); -F formal or standard; -G generalised; -L liturgical / pre-modern literary; -M middle; -U urban; -V vehicular

In the interests of standardisation, it is proposed to retain these capitalised suffixes in reference names, whatever alternative languages of presentation may be used for future editions of the Register. Although the single-letter suffixes are based on English usage, they can also be regarded as international or at least inter-European. All are valid abbreviations for the equivalent French terms also, except -W, which is valid also in German.

Occasionally, two or more otherwise identical reference names in the same zone have been distinguished by the suffixing of a numeral 2 to the second such name.

In some cases, especially where alternative spellings are current in the Latin script, different forms of the same linguistic name may be used as contrastive reference names in successive rows, even where the content of those rows is identical. For certain South American languages, for example, international or anglophone spellings in k- or ch- often correspond to Spanish or Portuguese spellings in c- or x-. In such cases, the former spelling has often been retained in the reference names of wider layers and the latter spelling in the reference names of the inner languages themselves.

It is recommended that the use of a Latin transcription of autonyms, even for well-known major languages, become acceptable as alternatives to established names in English, French, Spanish, etc., even in the context of a presentation in one of those languages. When the speakers of what was to become [52=] English first invaded the land which was to become England, they called the original inhabitants and their language "Welsh" (meaning literally «foreigner»!). It is scarcely reasonable that this event should justify the permanent precedence of the name Welsh over the modern autonym of the language, i.e. [50=] Cymraeg. Or likewise for other language names used by former invaders and colonisers, where these are at variance with the autonyms used by the speakers of those languages in the 21st century.

At the same time, since freedom of expression is the watchword, it is here suggested that individuals should feel free to use whichever form of a language name (or both) which they consider appropriate to the context in which they are writing or speaking. Even in the present text, a flexible approach has been adopted to this question, with either [50=] Cymraeg or [50=] Cymraeg (Welsh) or [50=] Welsh being acceptable and unambiguous alternatives.

The establishment of inner languages within outer languages, and the use of suffixes (see above) enable the Register to make the necessary distinction between (1) idioms dominated by the written word (and distinguished in the printed *Register* by the symbols 🕮 or ✍, ot by ᴮ or ᴳ online), and (2) all closely related vehicular and/or regional idioms, represented essentially by the spoken word.

The few reference names designating outer languages and inner languages which are in English (in column 2, other than for English itself, e.g. [12=] 'Syro-Mesopotamian') should be regarded as interim proposals, and the Observatoire Linguistique will be grateful to receive suggestions for more appropriate, indigenous forms and, where required, for more consistent transliterated forms.

Typography of Reference names outside column 2 of the Register

When cited outside column 2, the names of layers of worldwide reference (sectors and zones) are prefixed with the relevant numerical code plus a double hyphen (e.g. 9=Transafrican or 50=Celtic).

In the citation of reference names outside column 2, the typographical conventions prescribed for use in column 2 do not apply. It is important to emphasise that the Linguasphere Register does not seek to recommend or impose any convention or rule on how a language name should be chosen or written in any context outside the Register itself, other than suggesting that it should not be deprecatory and that it should be unique in its Latin transcription among the language names used within the relevant Linguasphere zone.

Aliases ('other names') in column 3 of the Register

Alternative reference names for all layers except dialect are printed first, in bold type, with the same rules of typography as for reference names in column 2 (except for quotation marks). Since most English and other 'foreign' cover names are treated as alternative reference names, they are normally cited without quotation marks in column 2, apart from lexical items such as "wider".

All other aliases or 'other-names' (alternative linguistic or ethnolinguistic names in column 3) are distinguished by the use of a lower-case initial and always precede any basic information included in the same row.

Commas are used to separate names which are presumed to be approximate equivalents or synonyms, whereas semi-colons (;) or a series of points are used, respectively, to separate independent names in a list or in a sequence. A final series of points (…) represents an open-ended or incomplete list. A raised star * is suffixed to an item of information which is provisional and/or remains to be confirmed.

Quotation-marks:

> Translations into English in column 3 of meaningful aliases or reference names, are enclosed within «angled or so-called 'French' quotes».
>
> Aliases which are disapprobatory or considered otherwise inappropriate are enclosed within 'single quotes'.
>
> Aliases, including scientific group-names in the language of presentation (i.e. English) but excluding alternative reference names, are enclosed within "double quotes".

Notes in the language of presentation (i.e. English in this edition) are normally in italics.

Foreign-names (exonyms), quoted from 'third' languages other than the language in question or the language of presentation (i.e. English), are preceded by the linguasphere key of that language. For example, under [52=] Schwytzertütsch (Swiss German), the Français (French) exonym is recorded as:

in [51=] *Français:* suisse-alémanique, suisse-allemand;

or under [12=] Shamerit (Litugical Samaritan), the Ivrit (Hebrew) exonym is recorded as:

in [12=] *Ivrit:* Shomronit.

CONVENTIONS FOR THE TYPOGRAPHY OF NAMES IN THE INDEX ARE DISCUSSED ON P.121.

2.6 Categories of Data

(as recorded in columns 3-4 of the Register, excluding nomenclature and statistics)

It should be recalled (see section 2.1 above):

> - that the central objective of the framework edition of the Linguasphere Register has been to record all known variants of living and recorded languages in the world and to classify them in terms of their closest relationships;
>
> - and that it now awaits the input of correspondents around the world in order to ensure that every language is accurately registered, including adequate information on its nomenclature, location, written forms, current use and relative number of speakers (see section 2.8 below).

Notwithstanding this limited objective, the foundation edition already presents several categories of basic information in columns 3 and 4 of the Register, with the intention that these should provide an initial framework for its subsequent expansion. A number of icons, as described below, have been introduced in this edition and it is proposed that these should be refined and extended, not only to classify a full range of basic data, but also to provide links to parallel sources of textual information and audi-visual materials on individual languages.

The brief selective data already provided in the present edition of the Register include the following categories, for which the coverage is not yet to be regarded as complete:

> - (in column 3) on **scripts**, on the **location(s)** or epientre(s) of areas where a language is spoken, on **interlinguistic** relationships, and on **speech communities** (including bilingualism; mobility, etc.);
>
> - (in column 4) on the **nation-states** & **provinces** where a language is spoken and on the official national status of a language.

Location

(as recorded in columns 3 and 4 of the Register)

A cross within a circle ⊕ introduces a note in column 3 on the area where a language is or has been spoken, normally during the 20th century, as represented by the names of one or more localities which constitute the "epicentre(s)" or spot locations of that language. Epicentres and other place-names (toponyms), together with occasional personal names (anthroponyms), are always listed with an initial capital, as opposed to linguistic names, and place-names used as linguistic names, which are listed with an initial lower-case letter in column 3.

Column 4 includes a list of the nation-states and administrative areas (provinces, states, regions, counties, etc.) where a language is spoken.

Names of nation-states are given in the language of presentation (in this case English), whereas names of administrative areas are cited wherever possible in the appropriate official language(s). The names of provinces have not yet been standardised, since they are subject to change by current national governments and may therefore appear with different names in sources from different periods.

For a relatively long period within the 20th century (between 1945 and 1989) political frontiers appeared to provide a stable framework of reference for the study of human society. More recent events, however,

have shown that such frontiers can in fact change overnight in a way that linguistic realities cannot (even though they may in some cases follow suit, as in parts of eastern and central Europe after 1945). The years since 1989 have provided many examples of the instability of nation states, which the on-going compilation of the *Register* endeavoured to take accoount of as they occurred (the break-up of the USSR and Czechoslovakia, the reduction of Yugoslavia and Ethiopia, with the creation of new nation-states, and the re-uniting of Germany, etc.). The internal administrative boundaries of individual states are even less stable, and there may be some inconsistencies in the nomenclature of specific provinces (or counties, etc.) in column 4, as a result of the inclusion of data referenced to different periods tn the administrative history of a particular country.

A cautious approach has also been adopted in repect of certain national names which have been recently changed. It does not appear reasonable to amend all reference sources each time a government decides to change the name of its own country, especially if such a government has not been democratically elected and/or there has been no change in the geographical area covered by such a name. In the present foundation edition, the former Zaire (now the Democratic Republic of the Congo) is referred to as Congo/Zaïre, and the former Burma (now Myanmar) is still referred to as Burma.

A flag υ before the name of a nation-state (or province) indicates that the relevant language currently has official status there. A separately bracketed flag (υ) indicates that the idiom no longer has the official status which it once had in the relevant nation-state. A technical problem led to omission of the flag symbol υ at some places in column 4, but an alternative source of information on the official languages of all nation-states is available from the National Index on pages 287 ff. below.

Scripts

(as recorded in column 3 of the Register)

An italicised open book 🕮 introduces a note on the script or scripts used for the language or languages in question, with or without a note on the approximate date from when that language has been written.

Although the compiler has been professionally concerned with the study of scripts[61], it has seemed preferable to exclude the reproduction of scripts (other than Latin) from this first edition. As the *Register* becomes a collective work after 2000, however, it is anticipated that future editions will reproduce the reference names of languages in the appropriate script in column 3, and in certain cases also their transcription in the International Phonetic Alphabet or equivalent.[62]

Steps are also being taken to conclude the developmental work on an extended Latin script, begun in Africa in the 1980's under the auspices of UNESCO and the Agence de la Coopération Culturelle et Technique. [63] This International Niamey Keyboard (INK) or *Linguasphere Alphabet* provides a pair of alternative characters for each of the 26 letters of the standard alphabet, allowing their precise phonological value to be established in terms of each language or non-Latin script being transcribed.

[61] See Dalby 1967, 1968, 1981, 1984, 1986

[62] In Mann & Dalby 1987 (see pp.206-218 "Writing African languages"), reference names for African languages were transcribed in the expanded character-set of the African Reference Alphabet. Such a transcription has been considered too complex to serve as part of a global system of referential names, but thought is currently being given to the adaptation of this phonemic alphabet (and of the related International Niamey Keyboard, *op.cit.* p. 217-8) to global transcriptions. It has the advantage of being directly linked to the standard Latin or Roman alphabet (on a 2 letters to 1 letter basis) and also of being closely related to International Phonetic Alphabet . The International Niamey Keyboard is being promoted by the Observatoire Linguistique as the transnational *Linguasphere Alphabet*, not as a replacement for other scripts but as a metascript enabling them to be transliterated into a common transnational form whenever required.

[63] Dalby 1984

Note on Languages and Speech Communities

(as recorded in column 3 of the Register)

The following categories of information are still of a framework nature, and will deserve considerable expansion in future editions.

Occasional linguistic notes are introduced by the symbol ➢, including notes on areas where further data are required (as for certain idioms recorded under [84=] Quechua-C., for example).

Revolving arrows X introduce comparative notes on linguistic relationships and/or classification, or on the translinguistic influence of one language on another. Of particular importance under this category are notes on transitional idioms or on the blurring of boundaries between languages of adjacent groups (labelled *transition to* in both cases), since these draw attention to cases where the necessarily rigid classification of the *Register* conceals cases where languages are in a transitional or ambiguous position. An example of a transitional idiom is provided in Europe by the Romance inner language [51=] Benasqués and an example of transition among adjacent nets of languages is provided in Africa by [99=] Kikoongo+Kiluba (Bantu-Inner-West) and Kiswahili+Ikiruundi (Bantu-Inner-East), within a series of such transitional cases in the 99=Bantuic phylozone.

A paragraph mark ¶ introduces an occasional note on the relevant community of voices, including such questions as their mobility (e.g. pastoral nomads, or hunter-gatherers), their migration, or their extinction or near-extinction.

The symbol E introduces a note on the *bilingualism* or *translingualism* of specific speech communities, with the implication that the majority of voices in that community are bilingual or translingual in the alternate language cited in the note.

2.6 Statistics

(including scale of voices in column 5 of the Register)

The scale of voices records a single-digit representation (or *democode*) of the numbers of speakers of a particular language, estimated at the end of the 20th century and expressed as orders of magnitude:

0 = "extinct during the 20th century" or (0) = "extinct before the 20th century"
1 = "less than 100 voices at the end of the 20th century"
2 = over 100 voices ("numbered in hundreds")
3 = over 1,000 voices ("numbered in thousands")
4 = over 10,000 voices ("numbered in tens of thousands")
5 = over 100,000 voices ("numbered in hundreds of thousands")
6 = over 1,000,000 voices ("numbered in millions")
7 = over 10,000,000 voices ("numbered in tens of millions")
8 = over 100,000,000 voices ("numbered in hundreds of millions")
9 = over 1,000,000,000 voices ("over one billion")

The Linguasphere scale of *voices* or speakers, although covering broad categories of population, are more useful for comparative purposes than precise census figures (in the minority of cases where these are available for specific languages), since points on this scale – known as *democodes* - can be established or estimated with reasonable accuracy for all the outer languages of the world. The *Register* presents a democode or estimated democode for every outer language in the world, and from these democodes it has been possible to establish a unique demographic assessment of the relative importance of all component languages in the linguasphere, at the end of the 20th century. A first global review has also been made of languages extinct since around the year 1900, and of those which are in danger of early extinction after the year 2000.

For comparative purposes, the *scale of voices* has been calculated also for the total number of speakers of all languages within each phylozone and geozone (as recorded not only in column 5 of the Register but also in the Synopsis of Sectors and Zones presented on pages 16-35 of Volume Two).

The scale 1-9, as applied to living languages, can thus be read in both directions: rising to 9 as a measure of their relative demographic importance, and rising to 1 as a measure of their relative danger of extinction.

Outer languages are more appropriate as demographic units than inner languages, since the boundaries among inner languages are frequently fluid, with many individual voices in permanent transition between two or more neighbouring or overlapping inner languages. Just as it is easier to set approximate physical boundaries to outer languages than among a cluster of adjacent inner languages, so it is easier to establish approximate totals of voices for the former rather than the latter. In appropriate cases, however, democodes have been included also for certain inner languages and dialects, but are printed in smaller, lighter type to avoid any confusion with the codes of outer languages.

A raised star * is suffixed to an (estimated) figure or *democode* which is provisional and remains to be confirmed. Where the population of a language is near the dividing line between two democodes, that language is assigned a higher democode with a star.

A skull-and-crossbones ☠ indicates the extinction during the 20th century (i.e. since 1900) of the only or last known language within a particular net, chain or set. The degree of loss of a particular group of languages is readily apparent from the co-occurrence of one, two or three of these symbols in column 4.

A bullet ● indicates the recorded extinction of the only or last known languages within a particular net, chain or set during the period 1500-1899.

Preliminary Estimates on the Major Languages of the World

(as recorded on pages 291-294 and 300)

The provisional table of the world's major languages in 1999-2000, presented on pages 291-294, is a preliminary assessment based not only on available estimates already published, but also on the latest population figures available for each country. The table should be regarded as a framework which may be progressively refined on the basis of informed documentation and estimates received for each country and major language of the world.

Megalanguages and Macrolanguages

For practical purposes, a statistical distinction is made between the category of 12 *megalanguages*, each comprising more than 100 million voices (or speakers), and the category of of 76 *macrolanguages* (including the 12 *megalanguages*), each comprising more than 10 million voices. As world populations rise, more languages reach and pass the total of 10 million voices, although the difficulties involved in estimating totals for each language should not be underestimated.

The table of major languages relates to spoken languages only, since the end of the 20th century has seen the spoken word achieve powers of global transmission not previously dreamed of. This new situation requires new strategies of communication, and new methods of analysis and presentation. It is particularly important that the spoken languages dealt with here should not be confused with corresponding written languages – not only because of varying literacy rates, but also because of the differential distribution of certain writing systems. [53=] Croatian and Serbian (formerly "united" as Serbo-Croat) may still be treated together as a pair of inter-intelligible spoken languages but continue to be written in different alphabets (Roman and Cyrillic). In Asia also, [59=] Hindi and Urdu overlap as spoken languages but are written in different scripts (Devanagari and Perso-Arabic). In contrast, the major [79=] Chinese languages (not readily interintelligible when spoken) are the "same" language in writing, through the use of identical meaning-based characters. For this reason, the estimates below take no account of those able to read a particular language. They also exclude the large number of speakers of Japanese and older speakers of [45=] Korean, for example, who are able to comprehend much written Chinese without needing to be acquainted with any spoken form of that language (because of the application of Chinese characters to the writing of their own languages).

The estimates in the Table of major languages relate not only to the *primary* voices of each language but also to its *alternate* voices. The first of these categories includes all who acquire "native" or "native-like" competence in a language, most frequently their *primary language* ("mother-tongue"), whereas the category of alternate voices includes all those whose speaking of the language in question is influenced by their knowledge of one or more other languages, especially their own primary language. These two categories replace the earlier tripartite distinction sometimes made between "mother-tongue speakers", "second-language speakers" and "foreign-language speakers" (where the distinction between "second" and "foreign" was based largely on the political status of a language in a particular country). It should be emphasised that alternate voices are a frequent medium of enrichment for a language, and should not therefore be regarded as "secondary" in status to primary voices.

In many cases, totals of primary voices alone would have presented a false picture of linguistic reality – only 6 million primary voices for [99=] Swahili, for example (as opposed to 55 million including alternate voices) or only 400 million primary voices for [52=] English (as opposed to around 100 million primary plus alternate).

Estimates of primary speakers may be based to a large extent on the latest census figures for relevant nation states, updated to allow for population growth, although specific linguistic totals are not available for the majority of countries. The assessment of alternate speakers, on the other hand, can only depend on the personal judgement of informed observers, and also on where one draws the line in terms of adequate competence in a language. Pupils who study a foreign language for several years in school, for example, cannot always be counted as "alternate" speakers of that language in any real sense. If it were possible to measure and count every "voice" concerned, it would be reasonable to include only those with the ability to "get by in a language" – to follow the main points of a televised drama or news bulletin, for example, or to ask and reply to straightforward questions about everyday subjects. There is great scope for the careful sampling of primary and alternate speakers of individual languages in specific countries and communities.

Of considerable but hitherto neglected importance is the fact that most of the world's principal languages belong to one of a small number of *nets* of closely related languages. In most cases, such relationships enable speakers of one language to acquire, with relative ease, an at least partial understanding of another language in the same network. Outer languages classified in the same net are often partially inter-intelligible, and a characteristic of many *nets* is that they include a proportion of "translingual " voices, able to negotiate with relative ease between two or more constituent outer languages.

For certain languages in the table below, especially macrolanguages, some allowance is made for voices alternating between languages in the same net, as in the ability of many [53=] Ukrainian and Belarussian voices to adapt themselves to Russian, of [53=] Czech and Slovak voices to Polish, or of [51=] Portuguese voices to Spanish (although less in the reverse direction, in each of these cases). A major example is provided by Hindi, reinforced by the popularity of [59=] Hindi-speaking films in South Asia. Many millions of voices alternate regularly between Hindi and their own languages in the same Indic network, often within the same conversation or same sentence.

For the purpose of assessing realistic totals, some pairs of sequences of very closely related languages have been treated together as a single unit. Examples from the following tables include [59=] Hindi+ Urdu, [31=] Malay+ Indonesian, [53=] Serbian+ Croatian, [53=] Czech+ Slovak, [52=] Swedish+ Danish+ Norwegian, and [52=] Dutch+ Flemish+ Afrikaans. This is not to imply that such paired languages are the 'same' language but that they are sufficiently close to form an extended transnational speech community.

Megalanguage describes a language with an estimated total in excess of 100 million *primary* and *alternate voices*. Of the twelve megalanguages spoken in the world at the end of the 20th century, no less than eight are classified within the Indo-European language-family. All twelve are national state-languages, and all include a more or less important intercontinental diaspora of speakers.

Although the identity and relative order of these twelve languages is reasonably certain, the figures cited for their estimated number of voices are by nature only very approximate. The following independent estimates have taken totals from other sources into account, and the demography of the countries involved although the most difficult estimate is that relating to [52=] English. By their nature, in an era of increasing telecommunication and travel, the number of voices for most megalanguages is still expanding, and this is particularly true for English, throughout the world. Some estimates in non-English language sources are as low as 500 million, and in some English-language sources as high as 1500 million or more.

The median figure presented here for English, of approx. one billion (1000 million) primary and alternate voices in the year 2000, has been scaled down to exclude more limited speakers of the language, although it is clear that the demand to learn English and the consequent supply of teaching are increasing rapidly in many parts of the world. It is likely that the total numbers of alternate (especially young) speakers of English are now expanding worldwide at a rate of well above a million a month, having already exceeded the estimated 400 million primary speakers of the language.

That humankind should be seeking and developing a worldwide vehicular language is in keeping with the evolution of planetary society towards a globalised speech community. On the other hand, the natural diversity of that community will depend on the development of English as a "service-language" of a multilingual society, and not as the vehicle of a dominant monolingual culture. This will be measured by the relative position of English in the category of primary speakers, any significant rise in which would indicate a major danger to humankind's linguistic diversity.

An interesting feature of the Table of major languages is that [79=] Chinese (Putonghua or "Mandarin") and [52=] English appear more or less "neck and neck" as the planet's two most spoken languages, breaking the barrier of around one billion speakers each at the turn of the millennium, but closely followed by [59=] Hindi+Urdu (including many alternate voices native to closely related languages like Panjabi).

In terms of geographical spread, English already occupies an undisputed position in the world, and a steady expansion in its learning and use around the globe will see its taking an increasing lead over all other languages during the early part of the 21st century. Moreover, within a decade or two, it is reasonable to assume that there will be more speakers of English in Asia than in any other continent: India is an increasingly important source of literary creativity in English, and China will not be far behind. Asia will play an important role in helping to ensure that English serves as a transnational auxiliary-language, against a multilingual background.

Modern English (an Indo-European language) does not belong to one of the 12 nets of closely related languages referred to above, but it stands in an important relationship with two of them, [52=] Deutsch+Nederlands (part of the Continental West Germanic chain) and [51=] Italiano+Português (or West Romance, part of the Romance chain and set). Its origins as a Continental Germanic language during the first millennium are reflected in the fact that Old English is today more easily accessible to a speaker of Modern German or Dutch than to a speaker of Modern English. The grammatical system of English has since been simplified and greatly changed, while a major proportion of its present-day vocabulary has been received from French or from other languages of the [51=] Romance chain, including Latin itself. English is therefore now best described as a "Romanised Germanic" language, within the Indo-European family.

In terms of its place in the world, [52=] French also is in a category by itself. Measured in terms of primary speakers only, it falls below the level of one hundred million speakers and lies in 12th position among the twelve megalanguages. And yet it is the only language which currently provides an alternative to English as a transnational vehicular around the world, being used as an official or semi-official language in 44 countries in five continents.

Among other languages estimated to fall within the category of *macrolanguages* (totalling more than 10 million *primary* and *alternate* voices each), the large majority serve as an educational and often also administrative medium for one or more nation-states or component federal states. See the Index of Countries on pp.287-90 below.

It should be noted that no less than eight macrolanguages in the [79=] Chinese net are estimated separately in the Table of major languages. A large proportion of their component voices now alternate between their own Chinese language and the official national or Pu-tong-hua (= "commonly understood language", known also in English as "Mandarin").

The macrolanguages which belong to the Indic network, and which are spoken in India and/or Pakistan, stand in a similar relationship to the Hindi+Urdu megalanguage, although this is not imposed educationally with the same vigour as the national language in China.

English has not retreated from the dominant position it occupied in the colonial period in South Asia, and is currently widely studied in English-medium schools frequented by children of the more favoured social classes and castes. Although the political and social situation differs from that in Chins, it is clear that the return of Hong Kong to China (as a linguistic "Trojan horse") has accelerated the demand for English-medium teaching in that country also.

Arterial Languages

It was during the final preparation of this framework edition of the Register that it was decided to adopt a relative as well as an absolute measurement of major languages in the modern world. Whereas macrolanguages and megalanguages are defined in terms of absolute numbers, the concept of *arterial languages* is defined in terms of the current population of the world, whatever that may be at any time.

> ● *An arterial language is defined as a language understood by one percent or more of humankind, in other words by at least sixty million speakers and hearers in 1999/2000, when the world population is estimated to have topped six billion for the first time.*

The future balance of linguistic diversity in the world will be measured, not only by the degree to which the national and local use of macrolanguages remains unimpaired by the wider use of English and other megalanguages, but also by the degree to which the *arterial languages* do not themselves impair the use of less widely spoken languages within their own areas of influence.

A fact of some considerable interest is that no less than 12 of the world's 28 *arterial* languages belong to only 3 nets (and 3 sets) of languages, these being [51=] West Romance, [59=] Indic and [79=] Sinitic, and have thus been significantly influenced by only three classical written languages: [51=] Classical Latin, [59=] Sanskrit and [79=] Classical Chinese.

2.8 Expanding the Register

At first sight, the present volume may appear to have been compiled for the use of specialists, but who are these specialists?

In practice, millions of intelligent observers throughout the world are specialists on the areas and languages surrounding them, and the help of such observers, including students and teachers in schools and colleges and universities of every continent, will be vital in improving and updating the Linguasphere Register. They can also play a key role in developing and applying a global philosophy of speech and education. It is hoped that the stimulation of transnational dialogue will also contribute to the future development of the philosophy of speech presented in these pages.

It can be argued that knowledge of two or more languages, and an access to differing views on the world, form part of the educational rights of every child, and that the eradication of monolingualism and illiteracy should be a dual objective for the decades ahead. For this goal to be realised, full and accurate information will be needed on every language and speech community in the world, a task which future revisions and expansions of the present Register will endeavour to provide with increasing levels of accuracy and detail. There is particular need for additional data on the thousands of minority speech-communities within the world's major cities, and on the electronic networks which now support and unite linguistic minorities scattered over wide areas of the globe.

From this point onwards, therefore, the Linguasphere Register becomes a collective transnational project, to ensure its ever greater precision and comprehensiveness. Speakers and observers of all languages, both individually and as members of institutes and organisations, private companies and schools, are earnestly invited to share the task of ensuring that subsequent editions of the Register present as full and as accurate a survey as possible of the linguasphere in the new millennium.

A member of any speech community should be able to find in the Linguasphere Register an accurate reference to her or his own language or own variety of language. Wherever this is not currently the case, however, it is requested that supplementary information should be forwarded to the Observatoire Linguistique, if possible by e-mail register@linguasphere.info or by post to Observatoire Linguistique, 18bis rue de Fontaine-Guérard, 27360 Pont St.Pierre, France (or) to Observatoire linguistique, Hendypost, Hebron, SA34 0XT Wales (GB). Responsibility and credit for the accuracy and completeness of future editions, in respect of each language and community, will lie with those who have access to the necessary information and local perspectives. Among new information now being collected are data on recorded or estimated levels of literacy and plurilingualism among voices in each language, together with the extended use of the demoscale (scale of voices) which is already provided for the numbers of voices of each outer language in the world. With its associated Mapbase, the Register will also seek to record salient facts on the current situation of individual speech communities, especially in cases where their survival or welfare gives cause for concern.

In this way, the data presented in the Register will have direct application to the work of humanitarian agencies and progressive governments, in the same way that those agencies and governments will be themselves able to provide new or revised data for future updated editions. It is important to recall that the ultimate speech community is the living human population of this planet, more united by the importance of speech in the personal and social life of each person than divided by anything else inherited from the past.

2.9 Acknowledgements

The Observatoire Linguistique and the compiler of the Linguasphere Register express their deep gratitude to the very many friends and colleagues and correspondents who have contributed to the preparation and compilation of this Register over so many years. The compiler accepts sole responsibility for all errors and omissions which remain in this edition, and offers his apologies to any person whose name may have been inadvertently omitted from this important section of the Introduction.

Two friends and colleagues have supported the compiler of the Linguasphere Register throughout this endeavour, and their names are included on the title-page as contributing editors: David Barrett of Regent University, Virginia, and Michael Mann of the School of Oriental and African Studies in London. Philip Baker has been a source of excellent council throughout, while Alison and Christopher Dalby, Jonathan Downs, Nima Gadhia, Barbara Kahana and Antony Sanderson have contributed valuable editorial support and advice, especially in the later stages of the work.

The continuous encouragement, interest and criticism provided by David Barrett since the original design of the Register have ensured that the compiler never wearied of his task. They have enabled the development of the Register and of its system of classification to be kept continually under review, and have encouraged the compiler to search always for improved solutions and styles of presentation. Without David's sustained friendship and advice, the Register would never have been completed.

Michael Mann's support dates back to the preparation of the *Language Map of Africa* in the 1970's [64], and his continued interest in the evolving project of the Register has been of major importance in terms both of scientific advice and of friendship. He has helped guide the compiler through the essential applications of computing to the project, and has been the only person to provide detailed comments and criticisms on the entire contents of the Register. Most importantly, in the final stages of compilation, he has designed and applied a program for the production of the Index to the Linguasphere Register, as included in this volume.

As an editorial consultant, Philip Baker of the University of Westminster has been closely involved with the planning of the Register and the Linguasphere Mapbase since their beginnings, and has been a source of reliable criticism and advice. He has successfully developed a parallel project initiated by the

---

[64] Culminating in the publication of Mann and Dalby 1987

[64] Gal 1989, p.315-316.

[64] citing R,Williams, "Base and superstructure in Marxist cultural theory", *New Left Review*, 87, 1973, pp.3-16

[64] In Mann & Dalby 1987 (see pp.206-218 “Writing African languages”), reference names for African languages were transcribed in the expanded character-set of the African Reference Alphabet. Such a transcription has been considered too complex to serve as part of a global system of referential names, but thought is currently being given to the adaptation of this phonemic alphabet (and of the related International Niamey Keyboard, *op.cit.* p. 217-8) to global transcriptions. It has the advantage of being directly linked to the standard Latin or Roman alphabet (on a 2 letters to 1 letter basis) and also of being closely related to the International Phonetic Alphabet . The International Niamey Keyboard is being promoted by the Observatoire Linguistique as the transnational *Linguasphere Alphabet*, not as a replacement for other scripts but as a metascript enabling them to be transliterated into a common transnational form whenever required.

[64] See Dalby 1967, 1968, 1981, 1984, 1986

[64] In Mann & Dalby 1987 (see pp.206-218 "Writing African languages"), reference names for African languages were transcribed in the expanded character-set of the African Reference Alphabet. Such a transcription has been considered too complex to serve as part of a global system of referential names, but thought is currently being given to the adaptation of this phonemic alphabet (and of the related International Niamey Keyboard, *op.cit.* p. 217-8) to global transcriptions. It has the advantage of being directly linked to the standard Latin or Roman alphabet (on a 2 letters to 1 letter basis) and also of being closely related to International Phonetic Alphabet . The International Niamey Keyboard is being promoted by the Observatoire Linguistique as the transnational *Linguasphere Alphabet*, not as a replacement for other scripts but as a metascript enabling them to be transliterated into a common transnational form whenever required.

[64] Dalby 1984

[64] Culminating in the publication of Mann and Dalby 1987

Observatoire Linguistique in 1994, in the form of a cartographic survey of nearly 300 languages spoken by school-children in Greater London, principally languages from South Asia and Africa (Baker 2000).

The support of geographers in the development of the Linguasphere programme has been of particular importance, especially that provided over several years by two of Europe's leading geolinguists, Roland Breton of France and Colin Williams of Wales, both of whom have contributed prefaces to the present volumes.

The Observatoire Linguistique is indebted to the School of Oriental and African Studies (SOAS, University of London), and to Tony Allan of the Department of Geography in particular, for the academic support which the School has provided throughout the planning and preparation of the Register, including the in-house printing of the preview editions in 1997 and 1998.

Yasir Mohieldeen, cartographer at SOAS of the first sheet of the Linguasphere Mapbase (covering the languages of Africa), Barbara Kahana, web-master of the Linguasphere Website (www.linguasphere.info), Yuvi Kahana, designer of the website, Christopher Dalby, responsible for statistics, and Graham Morse, who undertook the cover design of the present edition, have made their own specialised contributions to the Linguasphere programme. Warm thanks are due also to Binita Desai Lakhia in India and to Peter Wildbur in England for their contributions to the evolving design and layout of the Register. Bob O'Shea and his team at Lindsay Ross International are warmly thanked for their creative and insightful contribution to the final tasks of printing and distributing the Linguasphere Register.

The Observatoire Linguistique expresses its gratitude to Herrmann Jungraithmayr and to John Bendor-Samuel for having invited the first presentations and discussions of the principles of the Linguasphere Register at the Universität Johann Wolgang von Goethe (Frankfurt-am-Main, June 1990) and at the Summer Institute of Linguistics (Dallas, May 1991).

To these and to the many other friends and colleagues, and institutions, who have contributed ideas, information and direct support, the compiler expresses his deepest gratitude. His gratitude extends especially to the personal friends in several countries who have played an important part in establishing the Observatory as a transnational research network, or who have encouraged the ideas and research which have led to its creation and development. Each of the following continents has played an important role.

Africa was where the research leading to the Register had its beginnings, and the personal support of the poet Léopold Sédar Senghor since the early 1970's, when he was President of Senegal, has provided a major source of encouragement and good counsel. He became the first honorary president of the Observatoire Linguistique (Linguasphere Observatory) when it was created in the 1980's. Among other friends and colleagues from or in Africa, acknowledgements are due especially to Abdoulaye Barry, Thomas Decker, Djibril Diallo, Augustin Gatera, Paul Hair, Mubanga Kashoki, Béthoule Lambiotte, Nhlanhla Maake, Kahombo Mateene, Davidson Nichol, Augusta Omamor, Washington Omondi, Karim Turay, Ernst Westphal, Kay Williamson and Zachary Zachariev.

North America was where the concept of the Observatoire Linguistique was first formulated in 1983, in discussions with Grant McConnell and his colleagues, including Jean-Denis Gendron and Bill Mackay, at the then International Centre for Research on Bilingualism in Quebec. Support and advice over many years was received from such friends as Carleton Hodge, Gus Liebenow and Karl Voegelin at the University of Indiana, from the 1960's, through to younger colleagues in the 1990's, such as Konrad Tuchscherer at the University of Boston. We particularly appreciated advice and support received from Barbara Grimes, John Bendor-Samuel and other colleagues at the Summer Institute of Linguistics at Dallas, on the occasion of a consultative visit to them with David Barrett in 1992, as well as support received from the Foreign Mission Board in Richmond, Virginia, and from Emile Martel at the Canadian Embassy In Paris

Europe has been where the Observatoire Linguistique was subsequently established from 1983, and where its first supporting association was registered as a non-profit organisation in France (Association normande de l'Observatoire linguistique). Among other friends in France, Thierry Arnold, Philippe Blanchet, Françoise Carriol, Catherine Dalby, Gérard Galtier, Denise and Harry Hoffman, Geraldine

Levacher, Jacques Nemo, Pierre Raffin, Jean-Claude Rémy, Christiane Seydou, Nicole de Verneuil contributed to the development of the Observatoire and its Register, while Stélio Farandjis, André Martinet and Henriette Walter gave their distinguished support to its meetings. In Wales, where the research-base of the Observatoire was transferred in 1995, we are indebted especially to Colin Williams at Cardiff University of Wales and to Medwin Hughes and Dai Rogers of Trinity College, Carmarthen, for their academic and personal welcome, and to friends such as Humphrey Humphries, Martin Lloyd, Wyn Owens and Nerys Wyn Parry for their support in the creation of the Wylfa Ieithoedd, the Welsh association of the Observatory. In England, support and advice have been received over four decades from friends and colleagues at the University of London, especially at the School of Oriental and African Studies, including "Goosh" Andrzejewski, David Appleyard, Guy Atkins, Hugh Baker, Terry Bishop, Margaret Bryan, Julia Davidson, Philip Denwood, Bruce Ingham, Mike Farmer, Malcolm Guthrie, Ian Hancock, Arthur Hatto, Dick Hayward, Hiroto Hoshi, Arthur Irvine, Ulrich Kratz, Harry Norris, John Okell, Freddie Parsons, Nigel Phillips, Cyril Phillips, Patrick Quow, Renate Sohnen-Thieme, Archie Tucker, Denis Winston and Jae Hoon Yeon. Elsewhere in Britain, continental Europe and the Middle East, sincere thanks are due to Andrew Dalby, Sigmund Brauner, Oesten Dahl, Jean-Luc Fauconnier, Robin Gaff, François Grin, Rainer Hauer, Laurent Hendschel, Paul Lefin, Emma Minton, Baruch Podolski, Raymond Renard, John Ryle, Leslie St.James, Robert Wlattnig and Petr Zima.

South Asia is where the development of a comprehensive philosophy of the linguasphere was completed during the 1990's, our warm acknowledgements being due especially to Dilip Chitre, Ajay Dandekar, Ganesh Devy, Lachman Khubchandani, Dinesh Mahulkar, Ramaji Naik Nimbalkar, Pramod Pandey, Shashank Pujari, Nagamma & Ramakrishna Reddy, and Suhnu Ram Sharma. Ganesh Devy and members of the Bhasha Research Centre in Baroda were responsible for establishing and supporting the Observatoire's links with India and Dilip Chitre for providing the inspiration of an Indian poet.

Grateful acknowledgements are due also to the growing number of correspondents around the world who are now submitting comments on and additions to the Register, either by e-mail or by post, some of which were in time to be incorporated in the present framework edition, and others of which will contribute to future versions. The names of all individuals and institutions who make a direct scientific contribution in this way are being carefully recorded and will be published in future editions of the Register and on the Observatoire Linguistique website.