Introduction

BERNARD COMRIE, MATTHEW S. DRYER,
DAVID GIL, AND MARTIN HASPELMATH

1 What and why?

The World Atlas of Language Structures (*WALS*) provides the reader with 142 maps showing the geographical distribution of structural linguistic features. It is a quite novel type of atlas. Linguists have long worked with maps showing the geographical distribution of languages (i.e. the areas where most of their speakers live), and a complete world atlas was published a decade ago (Moseley and Asher 1994; see also *Ethnologue*, Grimes 2000). For well over a century, linguists have also produced atlases that show the geographical distribution of linguistic features in the dialects of a language. *WALS* is the first feature atlas on a worldwide scale. It can be thought of as a kind of dialect atlas of the "dialects" of Human Language. But it differs from dialect atlases in an important way. While dialect atlases show the geography of substantive linguistic features (such as particular cognate sounds, or particular words), *WALS* shows only **structural features**, i.e. abstract features of the language system that can be compared across unrelated languages.

Linguists interested in **linguistic typology**—the systematic study of the ways in which the languages of the world vary structurally and of the limits to this variation—have recently begun to ask questions relating to the geographical distribution of different values for structural linguistic features. For instance, they may want to know whether languages with a particular word order in the clause, say subject–verb–object as in English (*the farmer killed the duckling*), are found only in one part of the world, whether they are distributed more or less evenly across different parts of the world, or whether some in-between scenario holds. Although previous work has been able to provide some answers to some questions of this kind, these answers have hitherto been by and large unsystematic, often reflecting more the intuitive feel that a particular linguist has for the geographical distribution of the feature in question rather than a consistent sampling of the world's languages in order to answer the question. *The World Atlas of Language Structures* aims to provide just this kind of systematic answer, since the authors of the individual chapters, each dealing with a particular linguistic feature, have set out to be as comprehensive (within their sampling limits) as possible in the mapping of variants for that feature across the languages of the world. The printed atlas provides a visual overview of this distribution, by using different coloured dots for the different feature values. In addition, the text that accompanies each printed map provides an explanation of the feature values and of their assignment, as well as discussion of patterns of geographical distribution and of the relevance of the chapter to theoretical issues. The interactive electronic version (on the accompanying CD-ROM) provides much more information, including the possibility of zooming in on particular geographical areas that seem of particular interest for a particular feature. It provides access to the original data and bibliographical or other sources that underlie the atlas's database. It enables the user to manipulate data, for instance to calculate what percentage of the world's languages have a particular feature value, or to see whether there is a correlation between particular values of different features.

Although the main users of *WALS* will be those interested in linguistic typology, the atlas is also relevant to the interests of other linguists. For instance, certain theoretical approaches have been criticized for being based too heavily on languages exhibiting particular geographically restricted feature values, and *WALS* will enable those interested in testing such criticisms to see whether they do indeed hold. Thus, much recent work on relative clauses has been based heavily on the kind of construction found in the major European literary languages, such as literary English *the man whom I saw*, and has been criticized for extending the analysis appropriate for such languages to other languages with radically different relative clause types. While the material on relative clauses in the atlas does not, of course, directly address the issue of whether a particular approach to the analysis of relative clauses is valid cross-linguistically, it does show that the distribution of the "European" type of relative clause (in the terminology used here: the relative pronoun strategy) is by and large restricted to Europe, thus at least calling into question theoretical approaches that rely on so geographically restricted a typological variant (see Chapters 122–123).

Both typologists and other linguists are interested in questions relating to **correlations among different features**, such as whether there is indeed a tendency for occurrence of prepositions (rather than postpositions) to correlate with verb–object order in the clause, and for postpositions to correlate with object–verb order; whether there is a tendency for noun–adjective order in the noun phrase to correlate with verb–object order in the clause, and adjective–noun order with object–verb order. The relevant *WALS* data, in particular the maps that can be generated in the electronic version by combining data from different individual maps, suggest that there is validity to the postulated correlation in the first case (adpositions and verb position), but not in the second (adjectives and verb position) (see Chapters 95–97).

Previous work in typology has provided extensive results regarding the ways in which languages vary structurally and regarding correlations among different features. As noted above, *WALS* provides significant further contributions to these areas, especially with the data provided by the electronic version. However, because of the maps, *WALS* provides an especially significant contribution to the field of **areal typology**, which seeks to establish whether particular geographical distributions are the result of language contact among neighbouring languages.

The maps vary with respect to the degree to which the features show areal patterns. On some maps, clear geographical patterns emerge. On Map 83, for example, one finds a clear distribution of the two orders of object and verb among the languages of Eurasia. In Europe, most of the languages are VO (placing the verb before the object), while to the east of this is a huge area covering much of Asia, where most of the languages are OV (placing the object before the verb). And in South-East Asia, and extending out into the Pacific, is a large area where VO again predominates. Other maps show much less areal patterning of this sort. Where one finds similarities of this sort within a particular geographical area, there are three sorts of explanation. One is that it is the result of contact between languages. There are a number of well-documented instances of relatively small linguistic areas, such as the Balkans and Mesoamerica, where many features are shared due to contact, irrespective of their genealogical classification. However, some of the *WALS* maps suggest the possibility of larger linguistic areas, such as one covering much of northern Eurasia.

The second possibility is that it reflects a genealogical relationship among at least some of the languages, involving a feature inherited from a common ancestor. For example, the fact that VO order predominates from Indonesia and the Philippines and extending eastward into the Pacific (but excluding much of New Guinea) reflects the fact that almost all of these languages belong to the Austronesian family and the VO order is apparently a feature that is inherited from Proto-Austronesian, the ancestor language from which all of the Austronesian languages have descended. Note that in many cases shared features within a geographical area may be partly genealogical and partly due to contact. When languages within the same family share certain features after a long period of time, contact among speakers of different languages in the family may have reinforced similarities that were features of the protolanguage so that after a long enough period of time, the fact that the features are shared may be as much due to contact as due to their common ancestry. It should also be emphasized that features shared among languages in the same geographical area which are classified genealogically as belonging to separate families may reflect deeper genealogical connections. At first sight, this sounds like a contradiction: if they are in separate families, then how can they have a deeper genealogical connection? The answer is that it must be understood that when linguists classify languages into language families, they generally intend groups for which there is considered to be strong evidence of common ancestry. But there is little doubt that many of these families are in fact distantly related to each other but that the available evidence means that, at least at this time (and perhaps at any future time), there is a lack of convincing evidence for just which families are in fact related to which other families. But the absence of convincing evidence of a genealogical relationship does not mean that such families may not share a few features that are retentions from a common ancestor.

The third possible explanation for shared features within a particular geographical area is that it is at least partly coincidence. Especially when a map shows primarily two values, as does Map 83, with VO and OV word order, there are bound to be geographical areas in which one of these orders predominates where the occurrence throughout this area is simply accidental, where there may be two or more regions within this area which are predominantly OV or VO due to contact or for genealogical reasons, but where the fact that these regions are the same as each other is purely coincidental. Along the north coast of Papua New Guinea are a number of languages in different families, including Austronesian, which are VO. While some of these families may have acquired VO order due to contact with Austronesian, this seems rather unlikely for one family, Torricelli, since, as noted by Foley (2000: 365), most of the languages of this family are inland and have apparently had relatively little contact with Austronesian languages. Thus the contiguity of the VO languages in the Torricelli family with a few VO languages in the Austronesian family seems most likely to be coincidental.

One way to determine whether a trait that is shared within a geographical area represents nothing more than coincidence is to see whether there are other unrelated traits which are also shared among the same (or a similar) set of languages. Evidence for multiple shared traits within a particular geographical area provides evidence for a **linguistic area**. While the electronic version provides the possibility of investigating correlations among linguistic features throughout the languages of the world, it also provides the possibility of investigating correlations within particular regions. If a number of feature values correlate in a particular part of the world, but not in the world as a whole, then this is particularly strong evidence that the current distribution in that part of the world is the result of contact among its languages or possibly deep genealogical relationship. Even brief perusal of the maps will reveal, for instance, that South-East Asia often stands out from neighbouring areas, and sometimes even from the world as a whole, with the result that the materials provide a more solid basis than heretofore for the recognition of South-East Asia as a linguistic area (see, for instance, Maps 13, 51, and 55). (In addition, they show how some feature values characteristic of South-East Asia have a certain distribution beyond this area, reflecting weaker but nonetheless palpable language contact between South-East Asia and neighbouring areas with percolation of some feature values.)

Even a single trait, if relatively common in one area and relatively uncommon outside that area, can provide evidence for a linguistic area. The example cited above, of relative pronouns being common in Europe—not only in Indo-European languages but also such Uralic languages as Finnish, Estonian, and Hungarian—but uncommon elsewhere in the world, provides evidence for Europe as a linguistic area.

We hope that *WALS* will provide the interested lay reader with an overview both of the typological diversity of the world's languages —including reference to phenomena that might be judged highly unusual by the standards of the world's best-known languages—and of the ways in which that diversity is often patterned geographically, with particular areas often being characterized by particular feature values or particular combinations of feature values (whence the term "areal typology").

Although *The World Atlas of Language Structures* is a pioneering effort, we do acknowledge the efforts of earlier linguists who, on the basis of much less extensive data and with far fewer resources, nonetheless achieved significant results in mapping the world's linguistic diversity. Among our intellectual forebears, particular reference should be made to Schmidt (1926).

2 Organization

2.1 The features. The World Atlas of Language Structures contains 142 chapters, each consisting of a world map (plus four blow-up maps) followed by a double page of text. Each of the 142 chapters shows the distribution of a particular linguistic **feature**, reflected in the chapter's title. In a few cases, a single text accompanies a bloc of several successive maps. These texts are numbered according to the map numbers, so that a few texts receive multiple numbers (e.g. Chapters 122–123, a single text accompanying two maps).

Each chapter was contributed by an author (or team of authors) who is an expert on the particular structural feature, and who collected the worldwide cross-linguistic data from published materials and other sources. Quite a few authors are responsible for more than one chapter. Altogether 55 authors made a contribution to this work.

The 142 features are grouped thematically into the following eleven sections: phonology, morphology, nominal categories, nominal syntax, verbal categories, word order, simple clauses, complex sentences, lexicon, sign languages, and other. A complete listing of the features is provided in the Contents. As suggested by the section titles, the features span all of the major areas of language structure.

The first eight sections, phonology, morphology, nominal categories, nominal syntax, verbal categories, word order, simple clauses, and complex sentences, encompass the major structural domains of grammar. Within each of these sections, the features included provide a broad coverage of the most important subdomains within each of these structural domains. These eight sections constitute the core of the atlas, containing a total of 128 chapters.

The remaining three sections, containing a total of fourteen chapters, are of a more variegated nature. In the lexicon section, broad coverage would have been impractical; instead, a somewhat arbitrary choice of features provides a sample of the kinds of patterns that can be observed in this domain. In the sign-language section, too, the two features that are included represent arbitrary choices within the domain of sign-language morphosyntax. Finally, the last section touches on two diverse domains that are generally considered marginal to linguistic structure: paralinguistic sounds and writing systems.

The choice of features included in the atlas was governed by several competing considerations. As suggested above, the features were chosen to represent as many as possible of the major domains and subdomains of language structure. However, the desire to be as inclusive as possible came up against some practical considerations.

First, the requirement that each map represent a geographically and genealogically adequate sample of the world's languages entailed that the data for each map be obtained primarily from previously published descriptions of each individual language, typically in the form of a reference grammar. However, this limited the choice of features to those for which information is available in a typical reference grammar, and in a theory-neutral form facilitating cross-linguistic comparisons.

Secondly, the magnitude of this work, coupled with the desire to achieve a complete product within a limited time frame, entailed that the bulk of the data in the atlas derive from work already conducted over the course of a lengthy period, extending back for years and in some cases even decades. Thus, the choice of features was largely determined by what was already on offer from the contributing authors.

Due to these practical considerations, many features of current or potential future interest to linguistics had to be excluded. If the reader is disappointed that his or her favourite feature is not to be found in the atlas, chances are that this feature simply has not been described for a sufficiently large number of diverse languages to have warranted its inclusion. Hopefully, the absence of such features from the atlas will motivate future linguists to go out and collect the necessary data.

2.2 The maps. The great majority of maps show two hundred languages or more. Map 83 ("Order of object and verb") shows the greatest number of languages (1370 languages), while the two maps on sign languages (Maps 139 and 140) show a much smaller number, for the simple reason that linguists have only recently begun to study the grammatical structure of sign languages in a comparative perspective. On average, the maps show 409 languages. This is less than 10 per cent of the world's languages,[1] so the picture that we see in this atlas is far from complete. However, not more than 10–15 per cent of languages have been described comprehensively, and many hundreds of languages are still completely or almost completely unknown. But both descriptive and comparative linguistics have made enormous progress in recent decades, and these efforts are reflected in the current work. Altogether 2559 languages, somewhat less than one half of the world's languages, occur somewhere in the atlas—we call these the ***WALS* languages**. More than 6700 books and articles have been consulted by the authors and the relevant bibliographical references can be accessed in the electronic version of the atlas. In addition to the maps and accompanying texts, the atlas contains a genealogically organized list of the 2559 languages (the Genealogical Language List, by Matthew S. Dryer, on p. 584), with additional information on geographical location and alternative names, to facilitate identification of each language. At the end of the atlas is a language (and language family) name index. Issues having to do with the identification and designation of languages and language families are discussed in detail in §3 below.

The two map pages for each feature consist of a single world map (omitting unpopulated Antarctica) and below it four blow-up maps of areas with a particularly high density of languages. The world map is Pacific-centred, a representation that will look unfamiliar to many readers from Europe, Africa, and the Americas. A Pacific-centred map was chosen for two reasons. First, unlike the Atlantic Ocean, the Pacific Ocean has many small islands on which different languages are spoken. A Europe–Africa-centred map would have meant showing some Polynesian languages on the right and others on the left. Second, human populations spread out of Africa into Eurasia and across the Bering Strait into the Americas. Some linguists have suggested that similarities between Eurasian and American languages to some extent still reflect ancient population movements, and these similarities can only be captured in a Pacific-centred map. Contacts across the Atlantic have been much too recent to have an effect on the linguistic picture as the atlas shows it.

The world map is in Robinson projection, a projection which distorts shape, area, scale, and distance to create attractive average projection properties. The scale is 696 km per cm at the equator, 606 km per cm at 30 degrees south or north of the equator, and 352 km per cm at 60 degrees south or north of the equator. The shapes of regions north and south of 45 degrees latitude (including much of Europe and North America) are significantly distorted in this projection. However, for the purposes of displaying the world's languages, traditional rectangular Mercator maps would be unsuitable because they compress the areas around the equator, which are precisely the areas with the highest density of languages (see Nettle 1999 for documentation and discussion of this fact). Moreover, the ellipse in the Robinson projection better reflects the essentially round character of the earth than rectangular maps.

The four blow-up maps below the world map show exactly the same languages and symbols as the main map, but because of the greater resolution of the maps, many language dots that are covered by other dots on the main map are now distinguishable. Many chapters provide information on so many languages that even on the blow-up maps, many language dots are still underneath other language dots and hence invisible on the printed maps. Atlas users can turn to the interactive electronic version and zoom in closely enough that all languages can be identified. Which language dots are on top of others and thus visible in the printed version is largely random, though we have tried to make sure in some cases that languages which are of particular interest for that map (e.g. those that are mentioned in the accompanying text) will not be hidden underneath other languages.

Three chapters are somewhat special with respect to the maps. The two chapters on sign languages (Chapters 139–140) show pictures of signs instead of blow-up maps, and the chapter on writing systems (Chapter 141) shows differently coloured areas rather than differently coloured dots, as well as specimens of written texts in different systems instead of blow-up maps.

2.3 The feature values. Each feature is associated with a set of **feature values**, forming the basis for distinguishing between languages of different types. The simplest maps show just two different values. For example, Map 107 on passive constructions shows two language types, those possessing a passive construction and those lacking a passive construction. Most maps distinguish between three and five values, but some distinguish up to nine values (Chapters 33, 49, and 51). Although language types can usually be subdivided into many more subtypes, the values have been limited to at most nine in *WALS* because more types are difficult to represent by different colours, and users would find them hard to distinguish anyway. Each value is represented by a unique symbol, most often characterized by a particular colour, but sometimes also by a particular shape. Each symbol contains a three-letter code (the ***WALS* code**) that uniquely identifies the language (see §3.5 below, and the *WALS* Code Index at the end of the atlas).

Within each chapter, the feature values are indicated in abbreviated form in a small box on the map itself (in the southern Indian Ocean), and in expanded form in a larger box (called the **feature-value box**) within the accompanying text. For each feature value, these two boxes also show the colour and shape of the associated dot, and the number of languages characterized by the feature value in question. In addition, the text of each chapter provides a detailed description of the set of feature values, including the criteria that were used to assign feature values to individual languages.

One general requirement on the feature values of each chapter is that they be exhaustive. What this means is that for each feature, each and every language under consideration must be assigned one of the feature values; there are no "blank" cases of languages that do not have a feature value. In order to meet this requirement, many chapters include one or more feature values that account for various cases involving nonapplicability, indeterminacy, or mixing. For example, many maps characterize the different kinds of a certain construction, such as the associative plural (Chapter 36), but also include an additional feature value corresponding to the absence of the construction in question. Similarly, many of the word order maps (Chapters 81–97), such as noun–adjective order, also include an additional feature value corresponding to the absence of a dominant word order.[2]

Each feature value is associated with a dot of distinctive colour and, in some maps, also shape. The choice of symbols is intended to reflect, as accurately as possible, the logic underlying each set of feature values. For example, if feature values form an ordered set, then their symbols should ideally also form an ordered set. Similarly, feature values that are conceptually closer to each other should ideally be associated with symbols that are closer to each other. In many cases, however, it proved impossible to come up with a set of symbols that would be completely isomorphic with the logic of the feature values. Nevertheless, the reader will notice a number of colour and shape schemes which recur throughout the atlas,

[1] According to *Ethnologue* (Grimes 2000, vol. 1, p. 846), there are currently 6809 living languages.

[2] However, the exhaustiveness requirement is violated in Maps 14–17 on stress, which are silent on languages lacking stress. There was no way around this, because these chapters are based on a database collected earlier (StressTyp) that did not include stressless languages.

underscoring the presence of particular logical patterns each of which is shared by a variety of different maps.

No attempt has been made to make chapters by different authors that overlap in their features consistent, and it is not hard to find inconsistencies between certain chapters. For example, there are chapters by three different authors (or sets of authors) which deal with nominal case. There are some languages which are shown as lacking case by one author and as having case by a second author. These inconsistencies can arise for a number of reasons. Some occur because different authors use different criteria for identifying case, these differences hopefully being clear from the text accompanying the maps. Some occur because different authors classify particular borderline cases differently. Some occur because different authors use different sources and the sources describe the phenomenon differently. And some occur simply because of errors, either in an author's interpretation of their source or simply due to a coding error. In fact, quite independently of inconsistencies between chapters, there are bound to be errors of coding that will only be discovered after this atlas has been published. The editors will keep a **catalogue of reported errors** on the internet at <http://www.wals.info/>

3 The languages

3.1 The WALS samples. There is a total of 2560 languages which appear on at least one map in the atlas. Some of these languages (265 in number) appear on only one map, while some, such as English, appear on most of the maps. There are 175 languages which appear on at least eighty maps, and 424 languages which appear on at least forty maps. The choice of which languages to include on particular maps was the choice of individual authors. However, there is a set of 100 languages (hereafter the 100-language sample) which authors were asked to include on their maps if at all possible, and a further 100 languages which authors were encouraged to include on their maps (hereafter these two sets of 100 languages together are referred to as the 200-language sample).

A general desideratum for a good language sample is that it maximize both genealogical and areal diversity. Samples which include too many languages from one area of the world or too many languages from one family can provide a misleading picture of the relative frequency of different types of languages. Typological studies in the past have often included a disproportionate number of Indo-European languages or of languages of Europe or Eurasia. While Eurasia has a larger landmass than any other continental region in the world, fewer than 20 per cent of the languages of the world are spoken on mainland Eurasia (i.e. excluding the languages of Indonesia and the Philippines and other islands). In fact, there are more languages spoken on the island of New Guinea than in mainland Eurasia. Furthermore, as a number of the maps in this atlas show, there are patterns of similarity among languages of Eurasia that one does not find elsewhere in the world. For example, Map 97 shows that the vast majority of the OV languages of Eurasia (i.e. ones that place the object before the verb) place the modifying adjective before the noun. From this, linguists in the past erroneously concluded that this was a normal feature of OV languages. But as Map 97 shows, this is not true outside of Eurasia, where OV languages more often place adjectives after the noun. Only by using samples of languages which include many languages from outside Eurasia can we avoid making erroneous inferences of this sort.

Maximizing genealogical and areal diversity were major considerations in constructing the 100- and 200-language samples. However, there were a number of other considerations that played a role in constructing these samples that would not generally play the same role in constructing samples of languages. First, most of the languages of the islands of the Pacific fall within the Oceanic branch of the Austronesian family and thus are closely related to each other. For instance, one would normally not include more than one of these languages in a sample of 100 or even 200 languages. However, because the sample used here is for an atlas, we decided that we ought to include more of these languages, since otherwise there would be few dots on the maps in the Pacific. For this reason, there are two Oceanic languages in the 100-language sample and seven in the 200-language sample. Similar considerations led to the inclusion of three Bantu languages in the 100-language sample and five in the 200-language sample. Without these, many of the maps would have shown few languages in sub-Saharan Africa, and the majority of those shown would have been non-Bantu languages that are in some ways atypical of this region. A second consideration that would not normally play a role in constructing a language sample is that we felt that we ought to include a number of the major languages of Eurasia, even when this meant including pairs of languages which are too close genealogically to be otherwise included in a sample of 100 or 200 languages, including English and German, French and Spanish, and Modern Hebrew and Egyptian Arabic.

A further consideration in choosing languages for the 100- and 200-language samples was the ready availability of detailed grammatical descriptions. In most cases, the choice of a language over genealogically related languages was based on the availability of detailed descriptions. Some of the languages that were included in the samples are ones for which there is no detailed description but for which an expert on the language was willing to answer questions from authors (see §4 below). Some of the languages in the 200-language sample were chosen primarily for the purposes of maximizing genealogical or areal diversity, despite the fact that the available descriptions of these languages are somewhat meagre, thus making it impossible for many authors to include them on their maps. One language in the 200-language sample, Minica Huitoto, appears on only 32 maps; however, this was because we eventually realized the need to distinguish this language from other Huitoto languages and some authors in attempting to include Huitoto used sources for one of these other languages.

The choice of which languages to include in the 100-language sample and which to include in the 200-language sample was based on the following considerations. Languages which are more well known were normally placed in the 100-language sample. Languages with more readily available or more detailed descriptions were also placed in the 100-language sample. However, we also attempted to maximize genealogical and areal diversity in both samples so that in some instances, a language was placed in the 200-language sample but not in the 100-language sample if there was already a language that was close genealogically or geographically in the 100-language sample.

The following is a list of the languages in the 200-language sample, with those in the 100-language sample marked with an asterisk. The languages are organized in the same fashion as the entire set of *WALS* languages in the Genealogical Language List, by family, subfamily and genus (see the introduction to the Genealogical Language List on p. 584 for an explanation of the notion of genus). Names of genera are placed in italics.

Khoisan:
Central Khoisan: *__Khoekhoe__
Northern Khoisan: **Juǁ’hoan**

Niger-Congo
Adamawa-Ubangian: *__Sango__
Atlantic
Northern Atlantic: **Diola-Fogny**
Benue-Congo
Bantoid: **Kongo, *Luvale, Nkore-Kiga, *Swahili, *Zulu**
Defoid: *__Yoruba__
Igboid: **Igbo**
Gur: ***Supyire, Koromfe**
Kru: *__Grebo__
Kwa: **Ewe**
Mande
North-West Mande: **Bambara**
Kadugli: *__Krongo__

Nilo-Saharan
Central Sudanic
Bongo-Bagirmi: *__Bagirmi__
Lendu: **Ngiti**

Eastern Sudanic
Nilotic: ***Lango**
Nubian: **Dongolese Nubian**
Surmic: **Murle**
Fur: **Fur**
Kunama: **Kunama**
Maban: **Maba**
Saharan: **Kanuri**
Songhay: ***Koyraboro Senni**

Afro-Asiatic
Berber: ***Middle Atlas Berber**
Chadic
East Chadic: **Kera**
West Chadic: ***Hausa**
Cushitic
Beja: **Beja**
Eastern Cushitic: ***Harar Oromo**
Southern Cushitic: **Iraqw**
Semitic: ***Egyptian Arabic**, ***Modern Hebrew**

Basque: ***Basque**

Indo-European
Armenian: **Eastern Armenian**
Baltic: **Latvian**
Celtic: **Irish**
Germanic: ***English**, ***German**
Greek: ***Modern Greek**
Indic: ***Hindi**
Iranian: ***Persian**
Italic: ***French**, ***Spanish**
Slavic: ***Russian**

Uralic
Finno-Ugric
Finnic: ***Finnish**
Ugric: **Hungarian**
Samoyedic: **Nenets**

Altaic
Mongolic: ***Khalkha**
Tungusic: **Evenki**
Turkic: ***Turkish**

Yukaghir: **Kolyma Yukaghir**

Yeniseian: **Ket**

Chukotko-Kamchatkan: ***Chukchi**

Nivkh: **Nivkh**

Ainu: **Ainu**

Japanese: ***Japanese**

Korean: ***Korean**

North-West Caucasian: ***Abkhaz**

Nakh-Daghestanian
Daghestanian
Avar-Andic-Tsezic: **Hunzib**
Lak-Dargwa: **Lak**
Lezgic: ***Lezgian**
Nakh: **Ingush**

Kartvelian: ***Georgian**

Burushaski: ***Burushaski**

Dravidian
Dravidian Proper: ***Kannada**
North-West Dravidian: **Brahui**

Sino-Tibetan
Chinese: ***Mandarin**
Tibeto-Burman
Baric: **Garo**
Bodic: **Ladakhi**
Burmese-Lolo: ***Burmese**
Karen: **Kayah Li**
Kuki-Chin-Naga: **Bawm**, ***Meithei**
Lepcha: **Lepcha**

Hmong-Mien: ***Hmong Njua**

Tai-Kadai
Kam-Tai: ***Thai**

Austro-Asiatic
Munda: **Mundari**
Mon-Khmer
Aslian: **Semelai**
Khasi: **Khasi**
Khmer: **Khmer**
Palaung-Khmuic: **Khmu'**
Viet-Muong: ***Vietnamese**

Austronesian
Paiwanic: ***Paiwan**
Eastern Malayo-Polynesian
Oceanic: **Drehu**, ***Fijian**, **Kilivila, Kiribatese, Maori, Paamese**, ***Rapanui**
South Halmahera-North-West New Guinea: **Taba**
Western Malayo-Polynesian
Borneo: ***Malagasy**
Chamorro: ***Chamorro**
Meso-Philippine: ***Tagalog**
Sulawesi: ***Tukang Besi**
Sundic: **Karo Batak**, ***Indonesian**

West Papuan:
North-Central Bird's Head: ***Maybrat**

Sentani: **Sentani**

Border: ***Imonda**

Torricelli
Kombio-Arapesh: ***Arapesh**

Sepik
Sepik Hill: ***Alamblak**

Lower Sepik-Ramu
Lower Sepik: **Yimas**

Trans-New Guinea
Angan: **Hamtai**
Asmat-Kamoro: ***Asmat**
Binanderean: **Suena**
Dani: ***Lower Grand Valley Dani**
Engan: ***Kewa**
Madang: ***Amele**, **Kobon, Usan**
Mek: **Una**
Wissel Lakes-Kemandoga: **Ekari**

Marind
Marind Proper: **Marind**

Dagan: ***Daga**

Solomons East Papuan: ***Lavukaleve**

Australian
Bunuban: ***Gooniyandi**

Daly
Western Daly: **Maranungku**

Gunwinyguan
Nunggubuyu: **Nunggubuyu**
Yangmanic: **Wardaman**
Iwaidjan: ***Maung**
Mangarrayi: ***Mangarrayi**
Pama-Nyungan: ***Martuthunira**, ***Ngiyambaa**, **Pitjantjatjara, Yidiny**

Tangkic: ***Kayardild**
Tiwian: ***Tiwi**
West Barkly: **Wambaya**
Wororan: **Ungarinjin**

Eskimo-Aleut: ***West Greenlandic, Central Yup'ik**

Na-Dene
Athapaskan: **Navajo, *Slave**
Tlingit: **Tlingit**

Haida: **Haida**

Algic
Algonquian: ***Plains Cree, Passamaquoddy-Maliseet**
Yurok: **Yurok**

Iroquoian: ***Oneida**

Yuchi: **Yuchi**

Muskogean: ***Koasati**

Tunica: **Tunica**

Caddoan: ***Wichita**

Siouan: ***Lakhota**

Kiowa-Tanoan: ***Kiowa**

Keresan: ***Acoma**

Uto-Aztecan
Aztecan: **Tetelcingo Nahuatl**
Cahita: ***Yaqui**
Numic: **Comanche**
Takic: **Cahuilla**

Wakashan: ***Makah**

Salishan: **Squamish**

Kutenai: ***Kutenai**

Sahaptian: **Nez Perce**

Penutian
Miwok: **Southern Sierra Miwok**
Tsimshianic: **Coast Tsimshian**

Oregon Coast
Coosan: **Hanis Coos**

Hokan
Pomoan: **Southeastern Pomo**
Yuman: ***Maricopa**

Karok: ***Karok**

Oto-Manguean
Chinantecan: **Lealao Chinantec**
Mixtecan: ***Chalcatongo Mixtec**
Otomian: ***Mezquital Otomí**

Mixe-Zoque: ***Copainalá Zoque**

Mayan: ***Jakaltek**

Chibchan
Aruak: **Ika**
Rama: ***Rama**
Talamanca: **Bribri**

Choco: **Epena Pedee**

Barbacoan: **Awa Pit**

Tucanoan: ***Barasano**

Huitotoan: **Minica Huitoto**

Warao: ***Warao**

Yanomam: ***Sanuma**

Peba-Yaguan: ***Yagua**

Panoan: **Shipibo-Konibo**

Quechuan: ***Imbabura Quechua**

Aymaran: **Aymara**

Arawakan: ***Apurinã**

Cariban: **Carib, *Hixkaryana**

Tupian
Tupi-Guaraní: ***Guaraní, Urubu-Kaapor**

Macro-Ge
Ge-Kaingang: ***Canela-Krahô**

Trumai: **Trumai**

Chapacura-Wanhan: ***Wari'**

Mura: ***Pirahã**

Arauan: **Paumarí**

Tacanan: **Araona**

Cayuvava: **Cayuvava**

Matacoan: ***Wichí**

Guaicuruan: **Abipón**

Araucanian: ***Mapudungun**

Alacalufan: **Qawasqar**

Chon
Chon Proper: **Selknam**

Creoles and Pidgins: **Ndyuka**

While the 100- and 200-language samples could be used as samples for other typological studies, a genealogically more balanced sample, with only one language per genus, would remove Kongo, Luvale, Nkore-Kiga, Zulu, Koromfe, Modern Hebrew or Egyptian Arabic, German or English, French or Spanish, Bawm, Drehu, Kilivila, Kiribatese, Maori, Paamese, Rapanui, Karo Batak, Kobon, Usan, Ngiyambaa, Pitjantjatjara, Yidiny, Central Yup'ik, Navajo, Passamaquoddy-Maliseet, Carib, and Urubu-Kaapor. Three languages (Maybrat, Makah, Kutenai) were included because of the availability of language experts to answer questions from chapter authors or because copies of unpublished descriptions were made available to authors, and therefore might not be included in a sample used for future typological studies.

3.2 Language versus dialect. There are a number of instances in which distinct dialects are distinguished in the atlas. For example, *WALS* distinguishes four dialects of Inuktitut, the Eskimo-Aleut language of northern Canada. In a more extreme case, Map 54 shows a number of different dialects of German (because dialects vary with respect to the phenomenon investigated). In other cases, what is shown as a language is probably in fact a set of closely related but mutually unintelligible languages. An example of this is Bikol, spoken in the Philippines. No systematic attempt has been made here to distinguish dialects of the same language from different languages. Where authors submitted separate data for more than one dialect of a language, we generally maintained the distinction, and where authors submitted data using a name that covers a number of mutually unintelligible languages, we endeavoured to ascertain which language their data was based on, but in some cases we were unable to do so, and in some cases, their data was based on more than one source, where the sources describe different varieties.

We originally attempted to have a set of *WALS* languages in which no language was a variety of another *WALS* language. While we were able to minimize this, it proved impossible to apply the principle consistently. For example, as just mentioned, one map shows a number of varieties of German, while all other maps that include German simply show *German*, without specification of a particular variety. A more typical example is provided by Irish. Here, most authors provide data for Irish, without specifying a variety. But one author submitted data for Donegal Irish and another author data for Munster Irish (since the dialects differ from each other in some ways). The *WALS* set of languages also occasionally includes pairs in which one is a variety of the other, where the relationship is one of language to set of closely related languages rather than one of dialect to language. For example, while we generally distinguish different Huitoto languages, a few authors submitted data where it was not clear which Huitoto language their data was based on and for their maps, the *WALS* language is simply called *Huitoto.*

3.3 Language names. Many languages are known in the literature under different names. We have attempted to choose names for *WALS* that are the names by which the languages are currently known. This means that our name is occasionally different from that used in some sources on the language, where we have reason to believe that the name used in the source does not conform to more recent usage. For example, older sources on O'odham refer to the language as *Papago*, but we use the more recent name *O'odham*. Older names are often considered offensive by communities in which the language is spoken. In the Genealogical Language List on p. 584 we also give the name of the language as it is listed in *Ethnologue*, as best we can determine. The Language Name Index also includes the *Ethnologue* names as well as other names used in sources for the language (such as *Papago*).

When names of languages involve two or more words with a modifier followed by a head in the usual English name for the language, where the head denotes a language or language group and the modifier identifies a particular variety of that language or language group, we have two ways of referring to the language. In the chapter texts we use the usual name, with the modifier preceding the head, but in the Genealogical Language List and in the electronic version, we place the head first with the modifier following in parentheses. Thus, what is called *Chalcatongo Mixtec* in chapter texts is called *Mixtec (Chalcatongo)* in the list and in the electronic version. Note that we follow this convention both when the head denotes a language and the modifier identifies a dialect of that language (e.g. *Irish (Donegal)*) and when the head denotes a language group and the modifier identifies a particular language in that group (e.g. *Mixtec (Chalcatongo)*).

There are a number of cases that may look like instances of this, but where the head does not denote a language or language group of which the modifier identifies a variety. For example, Upper Kuskokwim remains in that form because it is not the upper variety of a Kuskokwim language (there is no Kuskokwim language or language group); rather, it is a language spoken in the vicinity of the Upper Kuskokwim river. Similarly, Tümpisa Shoshone remains in that form since it is not a variety of Shoshone, but just a closely related language. Some language names in English already occur in the form Head Modifier because their name reflects the syntax of some other language in which the modifier follows the head. These remain in the form Head Modifier, without parentheses. An example of this is Hmong Daw. Also we retain the order Modifier Head if the Head denotes a type of language, such as creole, pidgin, or sign language, rather than a genealogical category. An example of this is Berbice Dutch Creole.

There are many instances of homophonous language names, where two languages in different parts of the world happen to have the same name. Where two such languages occur in the set of *WALS* languages, we generally disambiguate them by adding a modifier of the form *in* plus country name in parentheses, as in *Baka (in Cameroon)* and *Baka (in Sudan)*, where the former is a Niger-Congo language and the latter is a Nilo-Saharan language. We sometimes employ this usage for only one language, where we use a different name for the second language, simply because the ambiguous name is sometimes used by others as a name for the second language. An example of this is *Mono (in United States)*, where *Mono* is also the name used by *Ethnologue* for the Austronesian language we call *Mono-Alu*. Occasionally this approach does not suffice to disambiguate a language name when both languages are spoken in the same country. In this case, we place the name of the language family in parentheses. An example of this is *Motilón (Chibchan)*, where the family is needed to distinguish it from the Carib language Yukpa, which is sometimes known as *Motilón* and which is also spoken in Colombia and Venezuela. Note that there are many instances of homophonous language names where we do not add a modifier because the other languages with this name are not *WALS* languages. For example, *Ethnologue* lists two languages called *Bulu*, one spoken in Cameroon, the other in Papua New Guinea. Since only the former is a *WALS* language, we simply call this language Bulu.

While our names include standard diacritic symbols (as in *á*, *ä*, or *ã*), we avoid using superscripts or symbols that are not standard symbols. Thus we use *Yidiny*, rather than *Yidin*^{y} or *Yidiɲ*.

3.4 Locations of languages. The languages are represented on the maps as dots, rather than as regions, but it should be borne in mind that many languages are spoken over areas larger than the dots. We attempted to locate the dots somewhere near the centre of the region where the languages are spoken, although in some cases this was difficult because the region in which the language is spoken is discontinuous. In these cases, we generally located the dot within the larger region in which the language is spoken. In some cases, the location of the dot is based on the location of a major city, town, or village in which the language is spoken. For example, Egyptian Arabic is located in the vicinity of Cairo, rather than in the middle of Egypt. For most languages, the location of the dots is based on their location on maps in Moseley and Asher (1994) or *Ethnologue* (Grimes 2000). For languages spoken in Canada and the United States, the location is based on Goddard (1996). For languages spoken in Nepal, the location is based on maps in Bradley (1997). For many of the languages spoken in Australia, the location is based on Tindale (1974). In some instances, the location is simply based on an explanation in the specific sources for the language. Future work will probably make clear that the location of some of the dots is inaccurate. Where we find inaccuracies after *WALS* is published, we will post information at http://www.wals.info/.

Note that in identifying the location of languages, we use locations prior to European colonial expansion. This means that the dot for English is located in England, and not in some other country where English is spoken. Similarly, Spanish is located in Spain, despite the fact that the majority of speakers are in the Americas. Thus, the languages shown in the Americas are for indigenous languages and for creoles and sign languages (since instances of the latter two types in the Americas only have locations in the Americas). Analogously, indigenous languages that are now spoken in locations different from where they were spoken at the time of European contact are located in their location at the time of European contact rather than their present location (in contrast to the practice in *Ethnologue*). For example, a number of indigenous languages of the United States are now only or primarily spoken on reservations in Oklahoma, often far from where their speakers originally lived. An example of this is Yuchi, originally spoken much further east, in what is now Tennessee.

3.5 Three-letter codes. The languages are identified on the maps by means of a three-letter code (***WALS* code**). A list of the *WALS* codes and the languages they denote is given at the end of this atlas. Readers who find these three-letter codes difficult to read will find them easier to identify on the zoomable maps of the elctronic version. We considered using the three-letter codes employed by *Ethnologue* but decided not to, both because there are many instances in which the languages in the atlas either represent varieties of languages in *Ethnologue* or correspond to an entire set of languages in *Ethnologue*, and because we wanted to use three-letter codes that are more mnemonic. The Genealogical Language List on p. 584 does give the *Ethnologue* three-letter code for each *WALS* language, as best we can determine. In assigning three-letter codes to languages, we first attempted to use the first three letters of the language name, unless the language name includes two or more words, in which case we attempted to use three letters based on initial letters in the different words. In other cases (where there is more than one language with the same three initial letters), we used the first three consonants in the name of the language (ignoring vowels). In some cases we had to use some other sequence of letters appearing somewhere in the name of the language and in a few instances we had to resort to adding as the third letter some letter that does not appear in the name of the language at all. But we always use the first letter in the *WALS* language name as the first letter in the three-letter code. With language names which are represented as *Modifier Head* in the chapter texts but as *Head (Modifier)* in the Genealogical Language List and in the electronic version, we use the latter for determining the three-letter code so that the first letter in the three-letter code matches the first letter in the head. For example, the three-letter code for Chalcatongo Mixtec, called *Mixtec (Chalcatongo)* in the Genealogical Language List and in the electronic version, is *mxc*. For languages whose names consist of two letters, such as Ik, we use a two-letter code rather than a three-letter code.

4 The data sources

The maps of the *World Atlas of Language Structures* are largely based on published primary sources that provide information about the languages in question. These include full grammars and dictionaries, but also more specialized articles that are confined to particular aspects of the language structure (e.g. only the phonological structure, or only certain syntactic constructions). Unpublished dissertations have also been used as sources, because these are often accessible to typologists. In a few cases (for languages like English, Spanish, or Russian), the authors have relied on their own knowledge of the language. Secondary sources, i.e. published typological surveys based on primary sources, have also occasionally been used where it seemed hard to avoid, although it is now generally recognized that comparative linguists should ideally work with primary sources.

Many dialect atlases have worked with an entirely different method of data collection, based on questionnaires. These are drawn up by the atlas editors and filled in by different fieldworkers on location for each data point. This is practical for dialect atlases because the fieldwork is restricted to a relatively small area, but for an atlas of global scope like *WALS*, this method would have required a budget a thousand times larger. Moreover, the questionnaire method is problematic for more sophisticated cross-linguistic work, because identifying certain phenomena (e.g. iambic rhythm or applicative constructions) in different languages requires detailed knowledge of the phenomenon in question.

The editors briefly considered the possibility of basing the atlas on questionnaires sent to a set of experts who know their respective languages so well that they would be able to answer structural questions without additional fieldwork. However, it quickly became clear that while this method would have the advantage of showing a uniform sample of languages on all maps (as in dialect atlases), it would be difficult to find enough experts willing to collaborate on such an enterprise, and it would not be wise to leave the rich data sources of published descriptions untapped.

As was mentioned in §3.1, the editors encouraged the authors to try and provide data on a fixed sample of 100 core languages and a further sample of 100 additional languages. For quite a few of these languages, the editors contacted experts, asking them to serve as consultants for the *WALS* authors; a list of their names is given in the Acknowledgements. Some authors also contacted other experts and received relevant data by personal communication from them.

More than 58,000 data points are shown on the *WALS* maps. Of these, about 2000 (or 3.4 per cent) are based on personal communications from experts, about 700 (or 1.2 per cent) are based on the authors' own knowledge or own data, and about 400 (or 0.7 per cent) are based on secondary sources.

For each data point (i.e. language-feature pair), the source used by the author is given in the interactive electronic version. For many of the data points, the electronic version also provides an example of the phenomenon in question.

5 The Interactive Reference Tool

The interactive electronic version (on CD-ROM) of the *World Atlas of Language Structures* contains the entire database on which this atlas is based and allows the user to display the data in a variety of ways, to conduct automatic searches, to export data and maps, and to create compound features based on the standard 141 features of the printed version.

The data in the electronic version of the database should be thought of as an appendix to *WALS* and its individual chapters. Thus, scholars who make use of these data in their research must refer to all the relevant chapters and give credit to their authors. It is not sufficient just to refer to "the electronic version of *WALS*".

Users of the electronic version can customize the map in various ways: show major cities and country names, remove country boundaries and rivers, and replace the light green/light blue base map by a topographic map showing altitude levels. The language dots can be shown in five different sizes, and the language name can be shown either as the three-letter *WALS* code inside the symbol, or in full to the right of the symbol. The colours and shapes of the symbols can be changed. When the mouse pointer moves over the dot, the full name is shown, and when clicking on a dot, a window with further information on the language opens (including the data source). Users can also zoom in on areas with high dot density, closely enough to see all dots separately, and drag on a map to see adjacent areas. Maps can be exported and printed, and various user-defined selections can be saved for future use. In some chapters, an example is provided for each data point.

The electronic version allows users to manipulate the standard features in two ways: values can be removed (if they are not of interest in a certain context), and several values can be merged into a single value. For instance, the five values of Chapter 1 (small, moderately small, average, moderately large, large) can be reduced to three (below average, average, above average) with just two mouse clicks.

Users can search for language names, genus names, family names, country names, and even for text in bibliographic entries. It is possible, for instance, to find and display all languages beginning with *X*, all languages belonging to the Austronesian family, all languages spoken in Colombia, or all languages described by Jeffrey Heath. On the electronic maps that only show languages (without giving information about the features), different dot colours may stand for different families or different genera.

Most importantly for comparative linguists, users can create their own compound features. For example, a linguist may want to know whether the existence of tone in a language is correlated with the type of syllable structure. Both features have three values (tone: none, simple, complex; syllable structure: simple, moderately complex, complex), so by combining them, one gets nine possible values:

No tones AND Simple syllable structure
No tones AND Moderately complex syllable structure
No tones AND Complex syllable structure
Simple tone system AND Simple syllable structure
Simple tone system AND Moderately complex syllable structure
Simple tone system AND Complex syllable structure
Complex tone system AND Simple syllable structure
Complex tone system AND Moderately complex syllable structure
Complex tone system AND Complex syllable structure

The program automatically creates a compound feature with these nine values, shows the number of languages for each value, suggests a symbol for each value, and displays a map of the compound feature. More complex ways of creating compound features are also possible and are described in detail in the electronic version.

6 Disclaimer

In identifying the status of speech varieties as languages or dialects, in assigning names to languages and dialects, in identifying countries, and in locating languages in countries, we have been guided solely by practical considerations and by current scholarly practice. In no instance should our usage be taken as implying a particular political stance or as insulting the speakers of a particular speech variety.