Journal of Language Evolution, 2016, 19–29
doi: 10.1093/jole/lzw002
Introduction

OXFORD

Linguistic diversity and language evolution

Harald Hammarström*

Department of Language and Cognition, Max Planck Institute for Psycholinguistics, Wundtlann 1, 6500 AH Nijmegen, The Netherlands

*Corresponding author: harald.hammarstroem@mpi.nl

Abstract

What would your ideas about language evolution be if there was only one language left on earth? Fortunately, our investigation need not be that impoverished. In the present article, we survey the state of knowledge regarding the kinds of language found among humans, the language inventory, population sizes, time depth, grammatical variation, and other relevant issues that a theory of language evolution should minimally take into account.

Key words: word order; language inventory; language evolution; linguistic diversity; ergativity.

1. Introduction

Human language may be defined as a human-learnable communication system with conventionalized form-meaning pairs capable of expressing the entire communicative needs of a human society (cf. Hockett 1960 for a similar view and background).

With such a definition, two kinds of languages are attested as mother tongues, namely, spoken languages where form is acoustic and there is a vowel/consonant distinction, and, signed languages where form is given by constellations of the human body. Other kinds of language without native speakers are also attested, including whistled languages (form is acoustic but there is no vowel/consonant distinction and the signal is a free airstream formed by the lips), drummed languages (form is acoustic but there is no vowel/consonant distinction and the signal is produced by means of a drum), and written languages (form is symbolic). It turns out that all known whistled languages (cf. Thierry 2002; Gartner and Streiter 2006), all known drummed languages (cf. Stern 1957) as well as all known written languages actually (ever) used by a human society are renderings of a spoken language. That is, each one is representation of a spoken language at some level, such as phoneme, syllable, morpheme, word, or the like, possibly with imperfections but nevertheless systematic. A few sign languages in Aboriginal Australia are also of this kind, notably Warlpiri Sign language, as argued by Kendon (1988). They are unlike all other known (fully developed) sign languages in that they can be analyzed as mapping to the corresponding spoken language, and did not emerge by way of a sizeable deaf (sub-)community but via cultural practices of silence in a predominantly hearing community.

2. The Language Inventory

There are approximately 6,500 attested spoken languages that are (or were) *mutually unintelligible* with each other (see Fig. 1). But let us spell out more carefully what this number actually counts.

In order to be included in that count a language has to be mentioned in some publication in such a way that one can argue that it is different from, that is, mutually unintelligible, to all other languages. This can be argued with direct assessment of intelligibility or with actual linguistic data, that is, form-meaning pairs. For example, in very poorly known areas, one may have to resort to testimonies as to whether a now extinct ethnic group needed an interpreter or not when communicating with their neighbors. In most cases, however, we make use of direct data from the languages involved. For example, a standardized word list

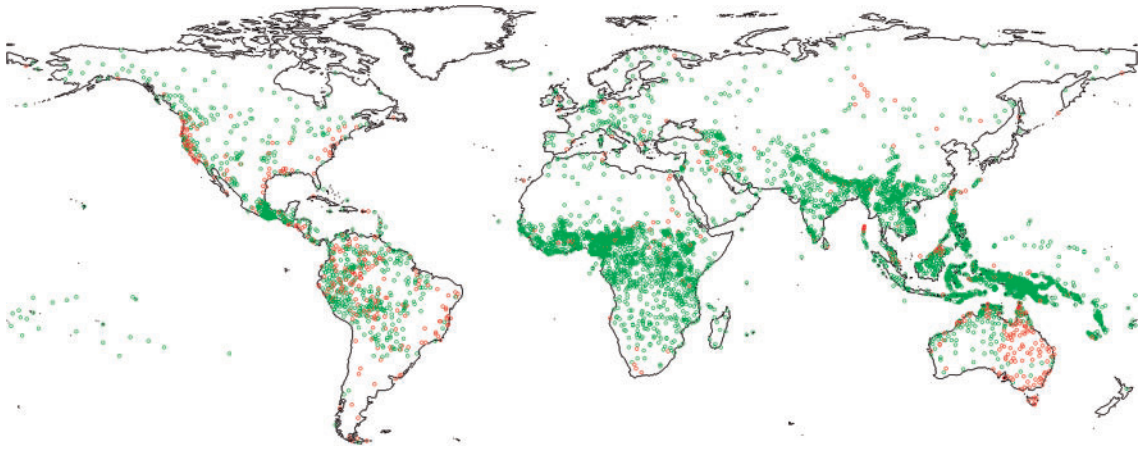

Figure 1. The language inventory rendered with one dot per language at the centre of its geographical location. A language is given a red dot if it is (known to be) extinct (Hammarström et al. 2015) and a green dot otherwise. (Colour online)

of 100 or 200-items of basic vocabulary (so-called Swadesh lists, cf. Tadmor et al. 2010) are commonly collected for the purposes of language comparison. It has been found empirically that if two languages share more than approximately 70 per cent of shared basic vocabulary (Wurm and Laycock 1962), they are likely intelligible. This number is not an exact physical constant, but simply one heuristic which is somewhat more objective and practical than various other alternatives (Casad 1974).

Second, the collecting of information regarding which languages exist(ed) is an extremely decentralized activity. The relevant information spans several centuries and involves missionaries, anthropologists, travelers, naturalists, amateurs, colonial officials, government censuses, and not least linguists. At least 40,000 bibliographical sources (thus, on average, six per language) went into the compilation of the language catalogs Glottolog and Ethnologue (described below). As can be expected, the evidence for the different languages varies tremendously, mostly according to region of the world.

As of 2015, the entire landmass has been surveyed for spoken languages at one time or another, with very few exceptions. The least well-surveyed areas include the northern and southern foothills of Indonesian Papua, the Nigeria-Cameroon borderland, the Javari river area (Brazil-Peru border area), pockets of the Democratic Republic of Congo and its border to Angola, the border area of Arunachal Pradesh (India) and China and the area around where Chad–Sudan–Central African Republic meet. Many regions of the world are or were politically difficult to survey for western scholars and are thus known only mainly from older surveys, for example, Myanmar and Libya. The world is so well-surveyed for languages that there are numbers, distribution and sometimes further information on peoples living without permanent contact with the outside world—one ethnic group in North Sentinel Island (Sarkar and Pandit 1994), an unknown number in Indonesian Papua and some seventy ethnic groups in South America, compare the overview in Brackelaire and Azanha (2006). Languages completely new to the scientific community continue to be discovered every year, but these are typically languages spoken by a (usually aging) fraction of an ethnic group who otherwise speak a known language, and that is how earlier surveys were never alerted to it. Apart from completely new languages, hundreds of revisions to the language inventory are made every year following newly collected information or more careful scrutiny of older data.

The survey situation with respect to nonspoken languages is very different. Sign languages taught in schools tied to nation states where deaf children from the same country are brought together are easy to track. But so-called village sign languages, that is, sign languages developed in a rural setting where a large proportion of the inhabitants is (often hereditarily) deaf, have not been systematically surveyed. Those reported in the literature are those that happened to catch the attention of sign language researchers. There is reason to believe there are proportionately more village sign languages in sub-Saharan Africa than in, for example, Western countries due to the prevalence of Bacterial Meningitis (Molesworth et al. 2002). Similarly, lists of initiation languages, whistled languages, ritual languages, and secret languages can be expected to be incomplete, as they are only documented on the random occasion of an interested researcher.

The figure of 6,500 languages defines languages in what may be termed the structural sense, that is, based on

whether sets of form-meaning pairs amount to intelligibility with other sets of form-meaning pairs. Laypersons, as well as some (socio-)linguists, typically operate with a logically different conception of language which is based on speakers' self-identification. We may call this the *political* language. If one asks 'what language do you speak and what other people speak your language?' the answer may or may not correspond to the structural sense of language. Sociopolitically, a speaker may identify with a larger set of varieties than those intelligible to his/her idiolect, or with a more restricted set of varieties. If the political and structural language do not coincide, it is usually the political one that takes precedence in the speakers' conception. For example, as Dixon (2002: 5) explains concerning the Western Desert area of Australia:

For the people themselves it is the tribal dialect (= political language) that has a name (in all but a very few instances)—for example, Pitjantjatjarra, Yankuntjatjarra and Pintupi in the western deserts area. Speakers of Pitjantjatjarra, Yankuntjatjarra and Pintupi recognise that these are mutually intelligible and—once the linguistic sense of the term 'language' is explained to them—acknowledge that they are dialects of one language. But this language had no name, in traditional times. There is now an accepted label. 'The Western Desert language' currently in use, by Aborigines and non-Aborigines, to describe a chain of dialects, each mutually intelligible with its neighbours, which extends over one and a quarter million square kilometres (one-sixth of the area of Australia).

If they do not coincide, the political language is much more often a smaller set of varieties than the structural one, which means that the number of languages in the sense of political language, is higher than 6,500—perhaps up to thrice as high.

There are currently two global-scale continually maintained inventories of the (signed and spoken) languages of the world:

***Ethnologue** (18ed): http://www.ethnologue.com contains speaker numbers, detailed locations, and other metadata*
***Glottolog** (2.6): http://www.glottolog.org contains sources of data on languages and a more principled (Hammarström 2015: 733–4) classification*

Both databases recognize something close to the structural definition of language but with deviations toward the political definition of language, and thus have inventories of around 7–8,000 languages (Hammarström 2015). The major difference is that Ethnologue carries metadata but lacks systematic sources for the information given, while Glottolog points to the primary sources for metadata. The Ethnologue languages inventory is indexed by three-letter codes which form the iso-639-3 standard. Glottolog is indexed by four letter + four digit glottocodes which identify any variety above or below the language level, that is (sub-)families and dialects. Both Ethnologue and Glottolog allow full download of their underlying databases and there is a straightforward mapping between glottocodes and iso-639-3 codes.

As should be clear from the preceding discussion, the language inventory as represented in the above numbers and databases is entirely dependent on there being a written record, if even by a traveler. For example, many languages in the Amazon went extinct in the past few centuries, and for many of them we have scraps of data from travelers ascertaining their previous existence. For eastern Brazil, where the obliteration took place earlier, such information is much more scarce leaving the list of languages we can assert much shorter. But by analogy with the neighboring regions where we have more data, quite possibly in eastern Brazil there were many more that never made it into the written record. Similar calculations could be brought over into prehistory. If a language is left alone it would take approximately 1,000 years for it to reach unintelligibility with its former self.[1] One could then take archaeological information and infer how many languages would have been spoken in any place in the past, and obtain a far higher number than 6,500 (Pagel 2000).

Some geographical conditions on language density have begun to be investigated. A general trend, in harmony with biological diversity, is that the nearer to the equator the more languages per (land) square kilometre (Nettle 1999; Gavin and Stepp 2014). There is also the correlation that impassable terrain such as mountains, forests, and swamps would harbor more language density, but this relation is complicated by the lower population density often found in such areas (Axelsen and Manrubia 2014).

3. Diversity of Language Populations

Judging from the present inventory of languages, language populations (Lewis et al. 2015) can consist of up to a billion speakers (English, Chinese, etc.) and go down to one speaker, if not already extinct. On all

1 Needless to say, this is not a physical constant but some kind of average obtained from known cases. Conditions under which a faster or slower speciation is to be expected are known (see, e.g. Bakker 2000).

Table 1. Median number of speakers across conventional macro-areas

Macro-area	Median no. of speakers
Eurasia	735
Australia	87
Africa	807
North America	299
South America	225
Greater New Guinea	643

continents, the median number of speakers is below 1,000 (see Table 1). Most languages with a very low speaker number (less than 100) represent languages in a declining stage, that is, the low speaker number does not reflect a stable state, but an ongoing shift to another more widely spoken language often seen as more prestigious and/or economically advantageous. Such a shift starts with bilingualism in one generation, broken transmission to some later generation and finally no transmission at all to the latest generation, leaving the language alive only as long as the oldest members of the early generation. A large number of languages are somewhere in this process and thus labeled *endangered languages*. The languages are witnesses to the world's linguistic diversity and the gradual disappearance is alarming (Evans 2009) especially if they disappear without sufficient documentation (grammar, dictionary, and texts). The Ethnologue language inventory contains endangerment information for all languages (though this is not always up to date). The website www.endangeredlanguages.com aims to collect resources and track the status of endangered languages.

The smallest linguistic populations attested in a *stable* state, that is, with full intergenerational transmission at least until the modern era are Masep (∼40, Indonesian Papua, Clouse et al. 2002), Marori (∼50, Indonesian Papua, Arka 2012), Mor (∼60, Indonesian Papua, own fieldwork), and Gurr-Goni (∼70, Australia, Green 2003). Thus, a number of three dozen presents itself as an empirical lower limit as to how small a language can be and still survive. Naturally, the speakers of these small languages are all (at least) bilingual. The smallest predominantly monolingual communities can be found in the Amazon forest (e.g. Zuruwaha ∼140 speakers, Suzuki 1997: 13). However, such communities are often runaways from political turmoil in the wake of the rubber boom era (∼1900) and their monolingualism probably reflects their recent post-rubber boom situation, not a longer tradition of monolingualism.

There are no worldwide figures for bilingualism and multilingualism but impressionistically, bilingualism from childhood would appear to have been more the norm than the exception. Cases where entire communities speak five or more languages are known from sub-Saharan Africa (Lionnet 2010: 2; Lüpke 2013) where knowing that many languages is necessary for everyday social and economic activities. It remains to be seen if there are any cognitive constraints on human multilingualism. At present, nowhere is the sociopolitical situation such that interaction in even more languages, for example, a dozen or so, is necessary for daily life.

It will come as no surprise that large speaker populations can be linked to nation states and empires, and that essentially only small speaker populations are found in the dwindling fraction of societies less affected by the modern frontier. Before the advent of agriculture beginning some 12,000 years ago (Diamond 1997), we must expect the speaker populations to have been similar to those of hunter-gatherer societies today. Given the marginalization of present-day hunter-gatherers, a faithful and/or generalizable speaker number distribution is difficult to produce, but would tend toward an average of a 1,000 or less and a ceiling within sight (a hunter-gatherer language of over, e.g. 100,000 speakers would be absurd to imagine in prehistory).

4. Genealogical Diversity

For centuries, linguists interested in the history of languages and their speakers have been primarily occupied with finding *language families*, that is, sets of languages that resemble each other so much—mainly in basic vocabulary—that one must assume they derive from a common ancestor (Campbell and Poser 2008). The most well-known of all language families is Indo-European (∼580 languages) native to a wide area stretching from the British isles to Bangladesh, including Dutch, English, Kurdish, Greek, French, Armenian, Hindi, etc. The largest language family in terms of number of member languages is Atlantic-Congo (∼1430 lgs) covering most of sub-Saharan Africa. The largest language family in terms of geospatial distribution is Austronesian (∼1274 lgs) stretching from Hawaii to Madagascar. The smallest language families have only one member language (also called *language isolates*), such as Basque (France–Spain), Etruscan (extinct, Italy), or Hadza (Tanzania). The current understanding of demonstrated families admits no less than 424 families in the count of Hammarström et al. (2015). The genealogical diversity is unevenly distributed across continents, see Table 2.

The demonstration of language families largely (but not exclusively) rests on the comparison of basic vocabulary. Because of vocabulary replacement the signal

Table 2. Numbers of languages and families (including isolates) across macro-areas

	No. of languages		No. of families	
Greater New Guinea Area	1797	28.0%	127	29.9%
South America	490	7.6%	109	25.7%
North America	558	8.7%	71	16.7%
Africa	1845	28.8%	50	11.7%
Eurasia	1423	22.2%	35	8.2%
Australia	292	4.5%	32	7.5%
	6409	100%	424	100%

decays when tracing proto-stages into the past. Even under the most optimistic estimates of stability (Pagel et al. 2013), at some point, too little is left to find even deeper genealogical relations. Therefore, in spite of the tremendous interest by amateur and professional linguists for finding 'new' deep language families, actual successes are few. The rate of vocabulary replacement is not regular like the half-life of a radioactive isotope but neither is it completely random (Holman et al. 2011). An estimate of time-depth is possible thanks to a number of calibration points, either historical, inscriptional or archaeological (in the rare event that one can convincingly argue a link between an archaeological entity and a (proto-)language). The deepest families recognized as 'demonstrated' in, for example, Hammarström et al. (2015) are expected not to exceed 10,000 years. Unless there is a breakthrough in the way language families are demonstrated—that goes beyond basic vocabulary—this time limit cannot be improved upon: if languages share too little vocabulary we would not accept they are related while if languages share a lot of vocabulary, we do not think the relation is old.[2]

Time depths such as 10,000 years have no chance of shedding light on what forms language might have had at the time of its emergence. Also the distribution of genealogical diversity is not directly indicative of time of original settlement by *Homo sapiens* of different continents. As per Table 2, the most diverse areas are South America and Greater New Guinea which differ dramatically in the age of settlement—11,000 for South America (Waters and Stafford 2007) vs 49,000 years ago for New Guinea (Summerhayes et al. 2010) and, Africa, presumably the continent of origin, exhibits much less diversity. In essence, genealogical diversity can at any point in time be obliterated by the expansion of a random language, or one fueled by a technological advantage (such as agriculture) of its speakers.

2 It is not known whether the corresponding vocabulary decay generalizations hold (or hold less) also for sign languages, since there is insufficient documentation of any sign language family or sign language passed down from generation to generation (Fischer 2015).

Table 3. Present status of grammatical description of the world's languages. Figures computed from the bibliography of Hammarström et al. (2015)

Most extensive description		No. of lgs	
Long grammar	~300 pages and beyond	1,134	17.7%
Grammar	~150 pages	891	13.9%
Grammar sketch	~50 pages	1,602	25.0%
Phonology	A phonological description or similar	711	11.1%
Wordlist or less	A short wordlist of less	2,071	32.3%
		6,409	100%

5. Structural Diversity

By definition, all human languages can express the same set of meanings, but they differ endlessly in their ways to do so. Looking at only a few languages one might easily get the impression that there are only a few options a grammar might have, but this view breaks down quickly when considering an increasing number of languages (Evans and Levinson 2009). Linguists and nonlinguists alike have typically theorized about human language as well as its emergence without due appreciation of the actual diversity attested, let alone the potential diversity of languages gone extinct. Partly, this is due to the asymmetry or even lack of information available on minority languages. However, thanks to increased documentation and organization in the past decade, we are now in a much better position to map and ultimately understand the diversity. A conventional way of documenting a language is the so-called Boasian trilogy of a grammar, text collection, and dictionary.[3] Focusing on grammatical description, Table 3 gives number for how many of the languages of the world we currently have grammatical descriptions of various lengths. For almost half of the languages of the world, we lack a grammatical description of any kind, let alone more extensive descriptions.

On the level of sounds used in spoken languages, some languages make extensive use of pitch for lexical contrasts, while other languages use none. Some languages break down their words into only eleven

3 This convention is now a century old. A more modern version would also include at least conversational data and multi-modal data as well (Woodbury 2011).

segmental sounds while others bin the combinatorics so that they obtain over 100 segmental sounds. For a long time it was thought that every language makes use of distinctive sounds passing air through the nose, either nasal stops or nasal vowels, but in the 1990s languages in the Lakes Plain area of Indonesian Papua were found that have no nasals either phonemically or phonetically (Clouse 1997). A small minority of languages, mainly in Southern Africa, make use of distinctive sounds where air is drawn *into* the mouth, so called click consonants, but there is little beyond speculation to suggest that these sounds carry primordial clues (Güldemann and Stoneking 2008).

All humans have the same body parts, yet even then, language divide up the lexical space differently. Some languages have a simplex lexeme than covers the hand and the arm (needing further specification if a distinction needs to be expressed) while others have one for each (Brown 2005). Similarly, while we have the same machinery for sensory perception, some languages have elaborate sets of simplex smell terms, while other languages conventionalize very few (Majid and Burenhult 2014).

Perhaps the most interesting kinds of variation is to be observed in the domain of grammar. Present-day humans face a world, communicative needs, perceptional capabilities, and organization of thought that is presumably congruent to that of the hominids who first evolved language. Therefore, we will review some conspicuous examples below.

In the temporal dimension, some languages obligatorily make the speaker distinguish between five past tenses (Payne 1985: 240), yet some other languages allow the speaker to leave this unspecified, understood from the context or optionally specified by a word like 'yesterday' (Dol 2007).

Some languages divide all referents into classes/genders which show different agreement patterns. In some languages, the speaker is forced to mark the gender overtly all over the sentence, while other languages have minimal marking redundancy. Many class/gender systems systematically distinguish natural gender (masculine/feminine) but vary in their ways to do so. Sometimes referents that do not have natural gender (e.g. a stone) have a class of their own, sometimes all are feminine or all masculine, and sometimes each item is more or less arbitrarily assigned. In some languages, mixed groups default to masculine agreement, in other to feminine agreement, and so on.

Perhaps the most conspicuous distinction concerns that of speech-act participants, that is, you versus me, and so on. Typically, grammars have simplex and obligatory distinctions between 1/2/3 person and also some obligatory distinction in the number of participants, such as the pronoun system of Dutch (Table 4). One might think that this is the only kind of 'sensible' system of oppositions, but many languages would disagree. For example, Indonesian (Table 5), has an inclusive/exclusive distinction in that a different word for 'we' is used depending on whether it includes the person being addressed. Gula Sara, a Central Sudanic language of Chad, differs further in that it has a special simplex form for 1 + 2 person singular (you and me) as well as an inclusive/exclusive distinction (Table 6). It also seems that, historically, a 'plural' ending containing a *g* might have been added to simplex forms of the categories person 1, 1 + 2, 2 and 3 ('minimal') to form the 'plurals' 1 exclusive, 1 inclusive, 2 and 3 ('augmented').

Every language must have conventions for expressing actions. The simplest division is between actions which require one participant, for example, 'walk' as in 'she walks', and those which require two participants, for example, 'chase' as in 'she chases him'. The participant in the one-participant clause is conventionally labeled S and the agent and patient in the two-participant clause are labeled A and P, respectively. Most languages have

Table 4. The pronoun distinctions in Dutch (Donaldson 1997)

	Singular	Plural
1	ik	wij
2	jij	jullie
3	hij/zij	zij

Table 5. The pronoun distinctions in Indonesian (Sneddon 1996)

	Singular	Plural
1	saya	kami (*EXCL*) kita (*INCL*)
2	kau	kalian
3	dia	mereka

Table 6. The pronoun distinctions in Gula Sara (Nougayrol 1999: 106)

	Minimal	Augmented
1	má	zígī (*EXCL*)
1+2	zé	zégēgē (*INCL*)
2	í	ség
3	nén	dég

conventional ways of marking the participants (so that the hearer can recover who did what to whom), either by ordering or by markers on the participants themselves. The question is now how languages chose to share the marking across the two kinds of clauses. One (uncommon) possibility is that there is no sharing between the two types, that is, that S, A, and P are marked in three different ways. But the most common possibility is that the S and A are marked the same, as opposed to P. This type is called nominoaccusative and encodes the worldview that there is always one actor, but for some kinds of clauses there are also participants acted upon. A third (not uncommon) possibility is that the S and P receive the same marking. This type is called ergative and encodes the mirror-image worldview. Curiously, languages which have ergativity, typically do not have it throughout the language, but only in certain tenses, or only for nonpronominal participants (Dixon 1994).

Probably because it is relatively easy to ascertain, the most energetically investigated element of grammatical variation concerns the *order* of element in a clause. In the transitive clause, we have an agent (A), patient (P), and verb (V). Most languages have a specific 'neutral' order, such as English AVP 'the dog kills the cat', while other languages allow alternative orders depending on, for example, the tense, while yet others allow any order[4] (see Dryer 2005 for a more detailed definition). Conventionally, orders are often presented using the letters S instead of A and O instead of P, even though the terminology with A and P is the more appropriate one given that the roles are semantically defined. Table 7 shows the frequency of the various orders in the languages of the world in terms of raw numbers of languages[5] and by the majority value per language family. Figure 2 maps the values geographically.

The data on constituent order illustrate the basic challenge for all researchers of grammatical diversity on a global scale. While the logical possibilities are inhabited, the distribution is not random at all—if it were, we would expect essentially uniform frequencies in Table 7. Languages from the same family are not independent, and language family sizes vary considerably, so the raw number of languages having a property is not a good indicator of any intrinsic preference for a certain property. However, when we stratify by language family (Table 7), there is still a very skewed, if not more skewed, distribution. One can similarly eliminate other potential sources of nonindependence, such a horizontal transfer (borrowing), by areal stratification. Once such confounds are arguably eliminated, one is left with a nonrandom fact, or, tendency, requiring explanation. Such patterns that reoccur across families and areas are termed *statistical universals*. Traditionally, explanations of statistical universals are sought in terms of processing machinery preferences (Hawkins 2014) or functional pressures related to communication (Haspelmath 2008).

Table 7. Raw counts of basic constituent orders in languages and families (majority value per family) across the world, adapted from Hammarström (2013)

	No. of languages		No. of families	
SOV	2,275	43.3%	239	56.6%
SVO	2,117	40.3%	55	13.0%
VSO	503	9.5%	27	6.3%
VOS	174	3.3%	15	3.5%
NODOM	124	2.3%	26	6.1%
OVS	40	0.7%	3	0.7%
OSV	19	0.3%	1	0.2%
Total datapoints	5,252		366	
No data	2,284		58	
Total	7,536		424	

Turning our interest to language evolution, the principle of uniformitarianism suggests that, in want of indications to the contrary, the earliest language was subject to the same functional pressures. This idea would entail that the maximum likelihood hypothesis is that proto-world was an SOV language[6] contra, for example, Jackendoff (1999), and by extension to other grammatical characteristics, probably quite unlike English.

Logically, there are other possible interpretations of linguistic features reoccurring across families and areas. One with much stronger implications for the evolution of language is the idea that such patterns reflect the proto-world stage rather than a cognitive/functional universal (Newmeyer 2000; Gell-Mann and Ruhlen 2011; Maurits and Griffiths 2014). On this hypothesis, many families have SOV order because they inherited it from

4 Such languages, whenever needed, mark who did what to whom in other ways than through a conventional order.

5 For ease of comparability with other databases and figures, the counts are presented in terms of the slightly inflated (in comparison to strict mutual intelligibility) language inventory of Hammarström et al. (2015).

6 At least when assuming that proto-world was a spoken language. In raw numbers, SOV is almost as common as SVO in (Kimmelman 2011: 8)'s survey of twenty-four signed languages. However, a more complete survey and an attempt to count independent cases are needed before we can pronounce a preferred word order for sign languages.

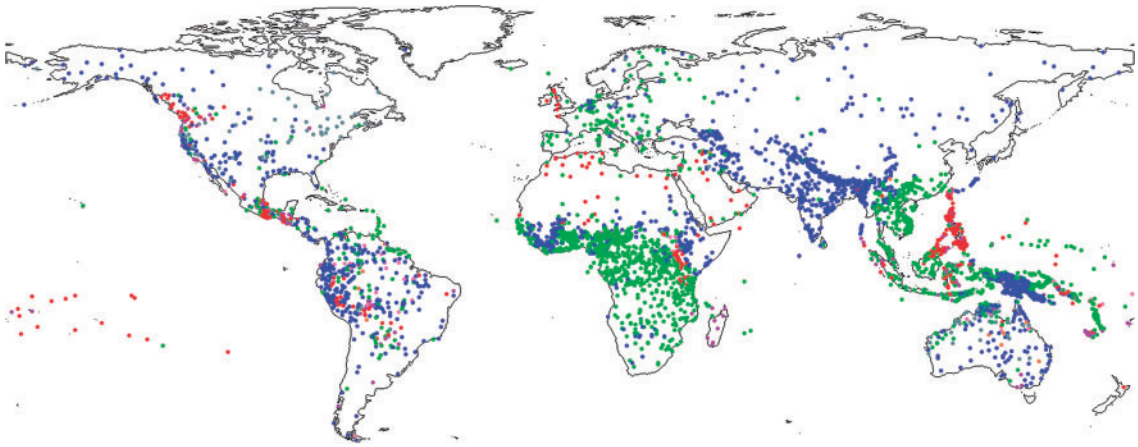

Figure 2. Geographical distribution of basic constituent orders in languages across the world, adapted from Hammarström (2013). Legend: SOV blue, SVO green, VSO red, VOS purple, NODOM gray, OVS yellow, OSV orange. (Colour online)

proto-world (which was SOV by coincidence), and those who do not have simply drifted away from SOV during the time after proto-world. It follows from this view that given more time, all languages will eventually have lost the trace of the original SOV state. Defenders of the classical view, that universals reflect cognitive-functional pressures rather than remnants of proto-world, emphasize that the age of known language families is very shallow compared to that assumed for proto-world and that a lot of change has happened within the time frame of those families. The argument for the proto-world view is that the changes that have happened recently seem to be in a consistent direction away from SOV (Newmeyer 2000; Gell-Mann and Ruhlen 2011; Maurits and Griffiths 2014).

6. Investigating Diversity for Language Evolution

On the classical view, language evolution happened so long ago and because of the volatility of language change and population dynamics, even if we knew everything about the present-day languages, little could be said about language evolution. But this view does not have to be definitive. In particular, in recent times, large databases have been amassed that enables us to study the implications of linguistic diversity in quantitative terms. In the past, linguistic data of various kinds, if available at all, has been rather fragmented and it is only in the last decade that the use of worldwide scale databases has become practical. It also remains almost completely unexplored what can be learned, with respect to language evolution, from combining linguistic data on a global scale with data from other disciplines such as archaeology, ethnography, and genetics (cf. Holman et al. 2015).

The following is a selection of databases of worldwide scope which are publically available and accessible and cover the domains of phonology, lexicon, and grammar respectively:

PHOIBLE http://phoible.org:

PHOIBLE Online is a repository of cross-linguistic phonological inventory data, which have been extracted from descriptive sources. The 2014 edition includes 2,155 inventories covering 1,672 distinct languages (Moran et al. 2015).

ASJP http://asjp.clld.org/:

The Automated Similarity Judgment Program (ASJP) has a database of 40-item word lists of basic vocabulary for 6895 varieties covering 4,401 distinct languages. The word lists are transcribed in a simplified but uniform transcription system.

WALS http://wals.info:

The World Atlas of Language Structures (WALS) has 192 multistate grammatical features sparsely filled in for 2,679 languages (Dryer and Haspelmath 2013). A subset of 200 languages are densely filled in. The features were individually designed by experts on the respective domain of grammar (and binned into maximally six feature values owing to the original publication as an atlas with no more than six colors on a map).

Some novel investigations along these possibilities are Roberts et al. (2014) who seek possible (but not necessary) reflections of Neanderthal admixture in language, Nichols (2008) who calculates spread rates of human migrations reflected in language families, and Wichmann and Holman (2009) who investigate the effect of population size on the rate of language change.

However, a few caveats are in order. A few direct global-scale studies did not lead to a breakthrough in our understanding of language evolution.

So far, no robust correlates of societal type has been found in grammar. In particular, the grammars of hunter-gatherer languages are not different from the more recent agriculture-based societal types (Bickel and Nichols 2016). The famous quotes by (Sapir 1921: 22, 234) 'The lowliest South African Bushman speaks in the forms of a rich symbolic system that is in essence perfectly comparable to the speech of the cultivated Frenchman' and 'When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam' hold empirically as true now as a century ago. Another study (Atkinson 2011), tracing the origin of language through a series of founder effects, falls short on the lack of a genuine link between phoneme inventory size and population size (Moran et al. 2012), and also fails to correct for multiple testing of 2,500 different potential points of origin in the significance values attributed to the correlation between phonemic diversity and distance from the putative origin.

It should also be kept in mind that large databases of linguistic diversity are approximations. For example, traditional grammatical feature divisions are akin to what a simple zoological questionnaire would be to the biological reality. That is, it is like asking, for each animal, does it have wings? can it fly? and so on, while in reality, there are many different kinds of wings (e.g. large, small, membranous, tegmina, and some arbitariness on what defines a wing in the first place) and degrees of flying (high altitude, duration, and so on), that are not captured by the questionnaire.

7. Conclusion

Language shows variation along a large number of dimensions that are relevant for any hypothesis on language evolution. In the present article, we surveyed the language inventory, population sizes, time depth, grammatical variation, and other relevant issues that a theory of language evolution should minimally take into account. Traditionally, language evolution is thought to have happened so far in the past that little could be inferred about language evolution from present-day information on language. But this view does not have to be definitive. In particular, in recent times, large databases have been amassed that enables us to study the implications of linguistic diversity in quantitative terms.

Acknowledgements

This article has benefited from comments by Dan Dediu, Simon Greenhill, and Bart de Boer. The usual disclaimers apply.

Funding

This research was made possible thanks to the financial support of the Language and Cognition Department at the Max Planck Institute for Psycholinguistics, Max-Planck Gesellschaft, and a European Research Council's Advanced Grant (269484 "INTERACT") awarded to Stephen C. Levinson.

References

Arka, I. W. (2012) 'Projecting Morphology and Agreement in Marori, an Isolate of Southern New Guinea'. In: Evans Nicholas and Klamer Marian (eds.) *Melanesian Languages on the Edge of Asia: Challenges for the 21st Century*, pp. 150–73 (Language Documentation & Conservation Special Publication 5). Honolulu: University of Hawaii Press.

Atkinson, Q. D. (2011) 'Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa', *Science*, 332/6027: 346–9.

Axelsen, J. B. and Manrubia, S. (2014) 'River Density and Landscape Roughness are Universal Determinants of Linguistic Diversity', *Proceedings of the Royal Society of London B: Biological Sciences*, 281/1784: 1–9.

Bakker, P. (2000) 'Rapid Language Change: Creolization, Intertwining, Convergence'. In: April McMahon, Colin Renfrew and Robert Lawrence Trask (eds.) *Time Depth in Historical Linguistics* (Papers in the prehistory of languages 2), pp. 585–620. Cambridge: McDonald Institute for Archaeological Research.

Bickel, B. and Nichols J. (2016) 'There is no Significant Typological Difference between Hunter-Gatherer and Other Languages'. In: Patrick McConvell, Tom Güldemann and Richard Rhodes (eds.) *The Language of Hunter-Gatherers: Global and Historical Perspectives*. Cambridge: Cambridge University Press.

Brackelaire, V. and Azanha G. (2006) 'Últimos pueblos indígenas aislados en América Latina: Reto a la supervivencia'. In: *Lenguas y tradiciones orales de la Amazonía. ¿diversidad en peligro?*, pp. 313–67. La Habana: Casa de las Américas.

Brown, C. H. (2005) 'Hand and Arm'. In: Comrie Bernard, Matthew S. Dryer, David Gil and Martin Haspelmath (eds.) *World Atlas of Language Structures*, pp. 522–5. Oxford: Oxford University Press.

Campbell, L. and Poser, W. J. (2008) *Language Classification: History and Method*. Cambridge: Cambridge University Press.

Casad, E. (1974) *Dialect Intelligibility Testing*. Norman, OK: SIL.

Clouse, D. et al. (2002) Survey report of the north coast of Irian Jaya. SIL International, Dallas. SIL Electronic Survey Reports 2002-078 <http://www.sil.org/silesr/abstract.asp?ref=2002-078>.

—— (1997) 'Toward a Reconstruction and Reclassification of the Lakes Plain Languages of Irian Jaya'. In: Karl J. Franklin (ed.) *Papers in Papuan linguistics No. 2* (Pacific Linguistics: Series A 85), pp. 133–236. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Diamond, J. (1997) *Guns, Germs and Steel: the Fates of Human Societies*. London: Cape.

Dixon, R. M. W. (1994) *Ergativity* (Cambridge Studies in Linguistics 69). Cambridge University Press.

—— (2002) *Australian Languages: Their Nature and Development* (Cambridge Language Surveys). Cambridge University Press.

Dol, P. (2007) *A Grammar of Maybrat: a Language of the Bird's Head Pennisula, Papua Province, Indonesia* (Pacific Linguistics 586). Canberra: Pacific Linguistics.

Donaldson, B. (1997) *Dutch: a Comprehensive Grammar*. London: Routledge.

Dryer, M. S. (2005) 'Order of Subject, Object, and Verb'. In: Comrie Bernard, S. Dryer Matthew, Gil David and Haspelmath Martin (eds.) *World Atlas of Language Structures*, pp. 330–3. Oxford: Oxford University Press.

—— and Haspelmath, M. (2013) *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology <http://wals.info> accessed 1 Oct 2015.

Evans, N. (2009) *Dying Words: Endangered Languages and What They Have to Tell Us*. John Wiley & Sons.

—— and Levinson, S. (2009) 'The Myth of Language Universals: Language diversity and its importance for cognitive science', *Behavioral and Brain Sciences*, 32/5: 429–92.

Fischer, S. D. (2015) 'Sign Languages in their Historical Context'. In: Claire Bowern and Bethwyn Evans (eds.) *The Routledge Handbook of Historical Linguistics*, pp. 443–65. New York: Routledge.

Gartner, B. and Streiter O. (2006) 'Smart Messages: The Whistled Languages of La Gomera (Spain), Antia (Greece) and Kuşköy (Turkey) – state of research and open questions'. In: Mathias Stuflesser, Andrea Abel and Magdalena Putz (eds.) *Mehrsprachigkeit in Europa: Erfahrungen, Bedürfnisse, Gute Praxis. Tagungsband*. Bozen: Eurac.

Gavin, M. C. and Stepp, J. R. (2014) 'Rapoport's Rule Revisited: Geographical Distributions of Human Languages'. *PLoS One*, 9/9(e107623): 1–8.

Gell-Mann, M. and Ruhlen M. (2011) 'The Origin and Evolution of Word Order'. *PNAS: Proceedings of the National Academy of Sciences of the United States of America*, 10: 1–16.

Green, R. (2003) 'Gurr-goni, a Minority Language in a Multilingual Community: Surviving into the 21st Century'. In Joe Blythe and R. McKenna Brown (eds.) *Maintaining the Links: Language, Identity and the Land: Foundation for Endangered Languages, Conference, 7th, Broome, 22–24 September, 2003*, pp. 127–34. Bath, UK: Foundation for Endangered Languages.

Güldemann, T. and Stoneking M. (2008). 'A Historical Appraisal of Clicks: a Linguistic and Genetic Population Perspective'. *Annual Review of Anthropology*, 37: 93–109.

Hammarström, H. (2013) *The Basic Word Order Typology: Universality, Genealogy, Areality*. Paper Presented at the Association for Linguistic Typology.

—— (2015) 'Ethnologue 16/17/18th Editions: A Comprehensive Review'. *Language*, 91/3: 723–37.

—— et al. (2015) *Glottolog 2.6*. Jena: Max Planck Institute for the Science of Human History. <http://glottolog.org> accessed 14 Oct 2015.

Haspelmath, M. (2008) 'Parametric Versus Functional Explanations of Syntactic Universals'. In: Theresa Biberauer (ed.) *The Limits of Syntactic Variation*, pp. 75–107. Amsterdam: John Benjamins.

Hawkins, J. A. (2014) *Cross-linguistic Variation and Efficiency*. Oxford: Oxford University Press.

Hockett, C. F. (1960) 'The Origin of Speech'. *Scientific American*, 203: 88–111.

Holman, E. W., et al. (2011) 'Automated Dating of the World's Language Families'. *Current Anthropology*, 52/6: 841–75.

—— et al. (2015) 'Inheritance and Diffusion of Language and Culture: A Comparative Perspective'. *Social Evolution & History*, 14/1: 49–64.

Jackendoff, R. (1999) 'Possible Stages in the Evolution of the Language Capacity'. *Trends in Cognitive Sciences* 3/7: 272–9.

Kendon, A. (1988) *Sign Languages of Aboriginal Australia: Cultural, Semiotic and Communicative Perspectives*. Cambridge: Cambridge University Press.

Kimmelman, V. (2011) 'Word Order in Russian Sign Language. An Extended Report'. *Linguistics in Amsterdam*, 4: 1–55.

Lewis, P. M., et al. (2015) *Ethnologue: Languages of the World*, 18th edn. Dallas: SIL International.

Lionnet, F. (2010) *Laal: an Unclassified Language of Southern Chad*. Paper presented at the Workshop on Language Isolates in Africa, Lyon, December 3–4, 2010.

Lüpke, F. (2013) 'Multilingualism on the Ground'. In: Friederike Lüpke and Anne Storch (eds.) *Repertoires and Choices in African Languages*, pp. 13–76. Berlin: DeGruyter Mouton.

Majid, A. and Burenhult N. (2014) 'Odors are Expressible in Language, as Long as You Speak the Right Language'. *Cognition*, 130/2: 266–70.

Maurits, L. and Griffiths T. L. (2014) 'Tracing the Roots of Syntax with Bayesian phylogenetics'. *PNAS*, 111/37: 13576–81.

Molesworth, A. et al. (2002) Where Is the Meningitis Belt? Defining an Area at Risk of Epidemic Meningitis in Africa'. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 96/3: 242–9.

Moran, S. et al. (2012). 'Revisiting Population Size vs. Phoneme Inventory Size'. *Language*, 88/4: 877–93.

—— et al. (2015) *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology <http://phoible.org> accessed 1 Oct 2015.

Nettle, D. (1999) *Linguistic Diversity*. Oxford University Press.

Newmeyer, F. J. (2000) 'On the reconstruction of "Proto-World" word order'. In: Michael Studdert-Kennedy, Chris Knight and James R. Hurford (eds.) *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pp. 372–90. Cambridge: Cambridge University Press.

Nichols, J. (2008) 'Language Spread Rates and Prehistoric American Migration Rates'. *Current Anthropology*, 49/5: 1109–17.

Nougayrol, P. (1999) *Les Parlers Gula: Centrafrique, Soudan, Tchad* (Collection Sciences du Langage). Paris: Centre National de la Récherche Sciéntifique.

Pagel, M. (2000) 'The History, Rate, and Pattern of World Linguistic Evolution'. In: Michael Studdert-Kennedy, Chris Knight and James Hurford (eds.) *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pp. 391–416. Cambridge: Cambridge University Press.

—— et al. (2013) 'Ultraconserved Words Point to Deep Language Ancestry across Eurasia'. *Proceedings of the National Academy of Sciences of the United States of America*, 110/21: 8471–6.

Payne, D. L. (1985) *Aspects of the Grammar of Yagua: A Typological Perspective (Peru)*. Los Angeles: University of California doctoral dissertation.

Roberts, S. G. et al. (2014) 'Detecting Differences between the Languages of Neanderthals and Modern Humans'. In: Erica A. Cartmill, Seán Roberts, Heidi Lyn and Hannah Cornish (eds.) *The Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10), Vienna, Austria, 14-17 April 2014*, 501-502. New Jersey, NY: World Scientific.

Sapir, E. (1921) *Language*. London: Harcourt, Brace & World.

Sarkar, J. K. and Pandit T. N. (1994) 'Sentinelese'. In: T. N. Pandit and B. N. Sarkar (eds.) *Andaman and Nicobar Islands* (People of India XII), pp. 184–7. Madras: Anthropological Survey of India.

Sneddon, J. N. (1996) *Indonesian: A Comprehensive Grammar* (Routledge Grammars Series). London & New York: Routledge.

Stern, T. (1957) 'Drum and Whistle "Languages": An Analysis of Speech Surrogates'. *American Anthropologist*, 59/3: 487–506.

Summerhayes, G. R. et al. (2010) 'Human Adaptation and Plant Use in Highland New Guinea 49,000 to 44,000 Years Ago'. *Science*, 330/6000: 78–81.

Suzuki, E. M. (1997) *Fonética e Fonologia do Suruwahá*. Universidade Estadual de Campinas, MA thesis.

Tadmor, U. et al. (2010) 'Borrowability and the Notion of Basic Vocabulary'. *Diachronica*, 27/2: 226–46.

Thierry, É. (2002) *Les Langages Sifflés*. Paris: Ecole Pratique des Hautes Études doctoral dissertation.

Waters, M. R. and Stafford T. W., Jr, (2007) 'Redefining the Age of Clovis: Implications for the Peopling of the Americas'. *Science*, 315/5815: 1122–6.

Wichmann, S. and Holman E. W. (2009) 'Population Size and Rates of Language Change'. *Human Biology*, 81/2–3: 259–74.

Woodbury, A. (2011) 'Language Documentation'. In: Peter K. Austin and Julia Sallabank (eds.) *Handbook of Endangered Languages* (The Cambridge Handbook of Endangered Languages), pp. 159–86. Cambridge: Cambridge University Press.

Wurm, S. A. and Laycock D. C. (1961–62) 'The Question of Language and Dialect in New Guinea'. *Oceania*, 32: 128–43.