How Many Languages Are There in the World?

Stephen R. Anderson

Download this document as a pdf.

The object of inquiry in linguistics is human language, in particular the extent and limits of diversity in the world's languages. One might suppose, therefore, that linguists would have a clear and reasonably precise notion of how many languages there are in the world. It turns out, however, that there is no such definite count—or at least, no such count that has any status as a scientific finding of modern linguistics. The reason for this lack is not (just) that parts of the world such as highland New Guinea or the forests of the Amazon have not been explored in enough detail to ascertain the range of people who live there. Rather, the problem is that the very notion of enumerating languages is a lot more complicated than it might seem.

There are a number of coherent (but quite different) answers that linguists might give to this apparently simple question.

More than you might have thought.

When people are asked how many languages they think there are in the world, the answers vary quite a bit. One random sampling of New Yorkers, for instance, resulted in answers like "probably several hundred." However we choose to count them, though, this is not close. When we look at reference works, we find estimates that have escalated over time. The 1911 (11th) edition of the *Encyclopedia Britannica*, for example, implies a figure somewhere around 1,000, a number that climbs steadily over the course of the twentieth century. That is not due to any increase in the number of languages, but rather to our increased understanding of how many languages are actually spoken in areas that had previously been underdescribed. Much pioneering work in documenting the languages of the world has been done by missionary organizations (such as the Summer Institute of Linguistics, now known as SIL International) with an interest in translating the Christian Bible. As of 2009, at least a portion of the bible had been translated into 2,508 different languages, still a long way short of full coverage. The most

extensive catalog of the world's languages, generally taken to be as authoritative as any, is that of Ethnologue (published by SIL International), whose detailed classified list as of 2009 included 6,909 distinct languages.

A family is a group of languages that can be shown to be genetically related to one another. The best known languages are those of the Indo-European family, to which English belongs. Considering how widely the Indo-European languages are distributed geographically, and their influence in world affairs, one might assume that a good proportion of the world's languages belong to this family. That is not the case, however: there are about 200 Indo-European languages, but even ignoring the many cases in which a language's genetic affiliation cannot be clearly determined, there are undoubtedly more families of languages (about 250) than there are members of the Indo-European family.

Languages are not at all uniformly distributed around the world. Just as some places are more diverse than others in terms of plant and animal species, the same goes for the distribution of languages. Out of Ethnologue's 6,909, for instance, only 230 are spoken in Europe, while 2,197 are spoken in Asia. One area of particularly high linguistic diversity is Papua-New Guinea, where there are an estimated 832 languages spoken by a population of around 3.9 million. That makes the average number of speakers around 4,500, possibly the lowest of any area of the world. These languages belong to between 40 and 50 distinct families. Of course, the number of families may change as scholarship improves, but there is little reason to believe that these figures are radically off the mark.

We do not find linguistic diversity only in out of the way places. Centuries of French governments have striven to make that country linguistically uniform, but (even disregarding Breton, a Celtic language; Allemannisch, the Germanic language spoken in Alsace; and Basque), Ethnologue shows at least ten distinct Romance languages spoken in France, including Picard, Gascon, Provençal, and several others in addition to "French."

Multilingualism in North America is usually discussed (apart from the status of French in Canada) in terms of English vs. Spanish, or the languages of immigrant populations such as Cantonese or Khmer, but we should remember that the Americas were a region with many languages well before modern Europeans or Asians arrived. In pre-contact times, over 300 languages were spoken in North America. Of these, about half have died out completely. All we know of them comes from early word lists or limited grammatical and textual records. But that still leaves about 165 of North America's indigenous languages spoken at least to some extent today.

Once we go beyond the major languages of economic and political power, such as English, Mandarin Chinese, Spanish, and a few more with millions of speakers each, everywhere we look in the world we find a vast number of others, belonging to many genetically distinct families. But whatever the degree of that diversity (and we discuss below the problem of how to quantify it), one thing that is fairly certain is that a surprising proportion of the world's languages are in fact disappearing—even as we speak.

Fewer than there were last month.

Whatever the world's linguistic diversity at the present, it is steadily declining, as local forms of speech increasingly become moribund before the advance of the major languages of world civilization. When a language ceases to be learned by young children, its days are clearly numbered, and we can predict with near certainty that it will not survive the death of the current native speakers.

The situation in North America is typical. Of about 165 indigenous languages, only eight are spoken by as many as 10,000 people. About 75 are spoken only by a handful of older people, and can be

assumed to be on their way to extinction. While we might think this is an unusual fact about North America, due to the overwhelming pressure of European settlement over the past 500 years, it is actually close to the norm. Around a quarter of the world's languages have fewer than a thousand remaining speakers, and linguists generally agree in estimating that the extinction within the next century of at least 3,000 of the 6,909 languages listed by Ethnologue, or nearly half, is virtually guaranteed under present circumstances. The threat of extinction thus affects a vastly greater proportion of the world's languages than its biological species.

Some would say that the death of a language is much less worrisome than that of a species. After all, are there not instances of languages that died and were reborn, like Hebrew? And in any case, when a group abandons its native language, it is generally for another that is more economically advantageous to them: why should we question the wisdom of that choice?

But the case of Hebrew is quite misleading, since the language was not in fact abandoned over the many years when it was no longer the principal language of the Jewish people. During this time, it remained an object of intense study and analysis by scholars. And there are few if any comparable cases to support the notion that language death is reversible.

The economic argument does not really supply a reason for speakers of a "small" and perhaps unwritten language to abandon that language simply because they also need to learn a widely used language such as English or Mandarin Chinese. Where there is no one dominant local language, and groups with diverse linguistic heritages come into regular contact with one another, multilingualism is a perfectly natural condition. When a language dies, a world dies with it, in the sense that a community's connection with its past, its traditions and its base of specific knowledge are all typically lost as the vehicle linking people to that knowledge is abandoned. This is not a necessary step, however, for them to become participants in a larger economic or political order.

For further information about the issues involved in language endangerment, see the LSA's FAQ "What is an endangered language?"

Count the flags!

To this point, we have assumed that we know how to count the world's languages. It might seem that any remaining imprecision is similar to what we might find in any other census-like operation: perhaps some of the languages were not home when the Ethnologue counter came calling, or perhaps some of them have similar names that make it hard to know when we are dealing with one language and when with several; but these are problems that could be solved in principle, and the fuzziness of our numbers should thus be quite small. But in fact, what makes languages distinct from one another turns out to be much more a social and political issue than a linguistic one, and most of the cited numbers are matters of opinion rather than science. The late Max Weinreich used to say that "A language is a dialect with an army and a navy." He was talking about the status of Yiddish, long considered a "dialect" because it was not identified with any politically significant entity. The distinction is still often implicit in talk about European "languages" vs. African "dialects." What counts as a language rather than a "mere" dialect typically involves issues of statehood, economics, literary traditions and writing systems, and other trappings of power, authority and culture — with purely linguistic considerations playing a less significant role.

For instance, Chinese "dialects" such as Cantonese, Hakka, Shanghainese, etc. are just as different from one another (and from the dominant Mandarin) as Romance languages such as French, Spanish, Italian and Romanian. They are not mutually intelligible, but their status derives from their association

with a single nation and a shared writing system, as well as from explicit government policy. In contrast, Hindi and Urdu are essentially the same system (referred to in earlier times as “Hindustani”), but associated with different countries (India and Pakistan), different writing systems, and different religious orientations. Although varieties in use in India and Pakistan by well-educated speakers are somewhat more distinct than the local vernaculars, the differences are still minimal—far less significant than those separating Mandarin from Cantonese, for example. For an extreme example of this phenomenon, consider the language formerly known as Serbo-Croatian, spoken over much of the territory of the former Yugoslavia and generally considered a single language with different local dialects and writing systems. Within this territory, Serbs (who are largely Orthodox) use a Cyrillic alphabet, while Croats (largely Roman Catholic) use the Latin alphabet. Within a period of only a few years after the breakup of Yugoslavia as a political entity, at least three new languages (Serbian, Croatian and Bosnian) had emerged, although the actual linguistic facts had not changed a bit.

One common-sense notion of when we are dealing with different languages, as opposed to different forms of the same language, is the criterion of mutual intelligibility: if the speakers of A can understand the speakers of B without difficulty, A and B must be the same language. But this notion fails in practice to cut the world up into clearly distinct language units. In some instances, speakers of A can understand B, but not vice versa, or at least speakers of B will insist that they cannot. Bulgarians, for instance, consider Macedonian a dialect of Bulgarian, but Macedonians insist that it is a distinct language. When Macedonia’s president Gligorov visited Bulgaria’s president Zhelev in 1995, he brought an interpreter, although Zhelev claimed he could understand everything Gligorov said. Somewhat less fancifully, Kalabari and Nembe are two linguistic varieties spoken in Nigeria. The Nembe claim to be able to understand Kalabari with no difficulty, but the rather more prosperous Kalabari regard the Nembe as poor country cousins whose speech is unintelligible. Another reason why the criterion of mutual intelligibility fails to tell us how many distinct languages there are in the world is the existence of dialect continua. To illustrate, suppose you were to start from Berlin and walk to Amsterdam, covering about ten miles every day. You can be sure that the people who provided your breakfast each morning could understand (and be understood by) the people who served you supper that evening. Nonetheless, the German speakers at the beginning of your trip and the Dutch speakers at its end would have much more trouble, and certainly think of themselves as speaking two quite distinct (if related) languages. In some parts of the world, such as the Western Desert in Australia, such a continuum can stretch well over a thousand miles, with the speakers in each local region able to understand one another while the ends of the continuum are clearly not mutually intelligible at all. How many languages are represented in such a case?

Related to this is the fact that we refer to the language of, say, Chaucer (1400), Shakespeare (1600), Thomas Jefferson (1800) and George W. Bush (2000) all as “English,” but it is safe to say these are not all mutually intelligible. Shakespeare might have been able, with some difficulty, to converse with Chaucer or with Jefferson, but Jefferson (and certainly Bush) would need an interpreter for Chaucer. Languages change gradually over time, maintaining intelligibility across adjacent generations, but eventually yielding very different systems.

The notion of distinctness among languages, then, is much harder to resolve than it seems at first sight. Political and social considerations trump purely linguistic reality, and the criterion of mutual intelligibility is ultimately inadequate.

At least 500 (But that’s just in Northern Italy).

So does the science of Linguistics provide a better basis for measuring the number of different languages spoken in the world? When we address the question of just when forms of speech differ systematically from a linguistic point of view, we get answers that are potentially crisp and clear, but

rather surprising.

If we try to distinguish languages from one another simply in terms of their words and the patterns we can observe in sentences, problems arise. Very different languages can share words (through borrowing) while different speakers of the "same" language may vary widely in their vocabulary due to factors of education or speaking style. Different languages may display the same sentence patterns, while a single language may display a great variety of patterns. In general, linguists have found that the analysis of the external facts of language use gives us at best a slippery object of study. Rather more coherent, it seems, is the study of the abstract knowledge speakers have which allows them to produce and understand what they say or hear or read: their internalized knowledge of the grammar of their language.

We might propose, then, that instead of counting languages in terms of external forms, we might try to count the range of distinct grammars in the world. How might we do this? What differentiates one grammar from another? Some aspects of grammatical knowledge, like the way pronouns are interpreted with respect to another expression in the same sentence, seem to be common across languages. In She thinks that Mary is smart, the pronoun she can refer to any female in the universe with one exception: she here cannot refer to the same individual as Mary. This seems to be a fact not about English, but about language in general, because the same facts recur in every language when the structural relations are the same. On the other hand, the fact that adjectives precede their nouns in English (we say a red balloon, not a balloon red) is a fact about English, since the opposite is true, for instance, in French. If we had a complete inventory of the set of parameters that can serve in this way, we could then say that each particular collection of values for those parameters that we could identify in the knowledge of some set of speakers should count as a distinct language.

But let us see what happens when we apply this approach to a single linguistic area, say Northern Italy. Consider the facts of negative sentences, for example. Standard Italian uses a negative marker which precedes the verb (*Maria non mangia la carne*'Maria not eats the meat'), while the language spoken in Piémonte (Piedmontese) uses a negative marker that follows the verb (*Maria a mangia nen la carn* 'Maria she eats not the meat'). Other differences correlate with this: standard Italian cannot have a negative with an imperative verb, but uses the infinitive instead, while Piedmontese allows negative imperatives; standard Italian requires a 'double negative' in sentences like *Non ho visto nessuno* 'not have I seen nobody' while Piedmontese does not use the extra negative marker, and so on. The functioning of negation here establishes a parameter that distinguishes these (and other) grammars. This is only the beginning, though. When we look more closely at the speech of various areas in Northern Italy, we find several other parameters that distinguish one grammar from another within this area, such that each of them can vary from place to place in ways that are independent of all of the others. Still staying within Northern Italy, let us suppose that there are, say, ten such parameters that distinguish one grammar from another. This is really quite a conservative estimate, in light of the variation that has in fact been found there. But if each of these can vary independently of the others, collectively they define a set of two to the tenth, or 1,024 distinct grammars, and indeed scholars have estimated that somewhere between 300 and 500 of these distinct possibilities are actually instantiated in the region!

Of course, the implications of this result for the world as a whole must be based on a thorough study of the range and limits of possible grammatical variation. But all of these forms of "Italian" have a great deal in common, and there are many ways in which they are all distinct as a group from many other languages in many other parts of the world. Since the number of possible grammatical systems expands exponentially as the number of parameters grows, if we have only about 25 or 30 of these, the number of possible languages in this sense becomes huge: well over a billion, on the assumption of

thirty distinct parameters. Obviously not all of these possibilities will be actualized, but if the space of possible grammars is covered uniformly to something like the extent we find in Northern Italy for the limited set of parameters in play there, the number of languages in the world must be much greater than the Ethnologue's 6,909.

Only one (A biologist looks at human language).

When we look at the languages of the world, they may seem bewilderingly diverse. From the point of view of communication systems more generally, however, they are remarkably similar to one another. Human language differs from the communicative behavior of every other known organism in a number of fundamental ways, all shared across languages. By comparison with the communicative devices of herring gulls, honey bees, dolphins or any other non-human animal, language provides us with a system that is not stimulus bound and ranges over an infinity of possible distinct messages. It achieves this with a limited, finite system of units that combine hierarchically and recursively into larger units. The words themselves are structured from a small inventory of sounds basic to the language, individually meaningless elements combined according to a system completely independent of the way words combine into phrases and sentences.

The particular linguistic system that each individual controls goes far beyond the direct experience from which knowledge of it arose. And the principles governing these systems of sounds, words and meanings are largely common across languages, with only limited possibilities for difference (the parameters described above). In all these ways, human language is so different from any other known system in the natural world that the narrowly constrained ways in which one grammar can differ from another fade into insignificance. For a native of Milan, the differences between the speech of that city and that of Turin may loom large, but for a visitor from Kuala Lumpur both are "Italian." Similarly, the differences we find across the world in grammars seem very important, but for an outside observer—say, a biologist studying communication among living beings in general—all are relatively minor variations on the single theme of Human language.

As the 11th edition of the *Encyclopedia Britannica* put it, "[...] all existing human speech is one in the essential characteristics which we have thus far noted or shall hereafter have to consider, even as humanity is one in its distinction from the lower animals; the differences are in nonessentials."

For Further Reading

Anderson, Stephen R. & David W. Lightfoot. 2002. *The Language Organ: Linguistics as Cognitive Physiology.* Cambridge; Cambridge University Press.

Baker, Mark. 2001. *The Atoms of Language.* New York: Basic Books.

Chambers, J. K. & Peter Trudgill. 1998. *Dialectology*. 2nd edn. Cambridge: Cambridge University Press.

Romaine, Suzanne. 2000. *Language in Society.* 2nd edn. Oxford: Oxford University Press.

The University of Maryland's Langscape project, available free online, provides interactive maps and linguistic data for 7,000 languages around the world.

*With contributions from David Harrison, Laurence Horn, Rafaella Zanuttini and David Lightfoot.