April 26, 1607. After a long 144-day voyage, three ships belonging to the Virginia Company of London and led by Captain John Smith make landfall at the southern edge of the mouth of the Chesapeake Bay, which they name Cape Henry. Shortly thereafter, they are forced to move their camp along the estuary, to a new location eventually known as “James Towne” or Jamestown, Virginia. Almost immediately, they encounter a group of “American Indians” who communicate with each other in what surely sounds like language, only it is quite different from the English of the Virginia Company settlers. Sure enough, there are words in this language, and just like back home, people from different places pronounce the same words somewhat differently. But to an English ear, these words are unrecognizable: not only are there words for things unfamiliar to the English settlers (some of which will later be taken into the English language, like raccoon, moccasin, opossum, and others), but even words for familiar objects and concepts are different: for example, the word for ‘sun’ is nepass and the word for ‘good’ is wingan. And it is not just the words that are different, so is the way the words are put together: for example, grammatical objects in this language typically precede rather than follow the verb. (The terms in small caps here and throughout the text are explained in the glossary.) This grammatical pattern would not have surprised the settlers, had they come from the Basque Country or Turkey or Japan, or even had they arrived 700 years earlier, but for the Virginia Company men it must have been a striking pattern. The differences between the language of these “American Indians” and English are so ear-grating that the English settlers start compiling lists of “American Indian”
words: Captain John Smith himself compiles a list of about fifty words, and William Strachey publishes a “dictionary” of the language containing about a thousand words. Today, most of what we know about this language – called Powhatan and belonging to the Eastern Algonquian branch of the Algonquian language family (see Chapter 11) – derives from the descriptions of seventeenth- and eighteenth-century writers, as the language died in the 1790s when its last speakers switched to English.

At about the same time as the English are colonizing the eastern seaboard of what will become the United States, the Russians are pushing into Siberia. Only a quarter of a century after the first encounter between the Virginia Company men and the Powhatan Indians, a Russian company of twenty or so men led by Pyotr Beketov lands on the shores of the Lena River and, on September 25, 1632, founds the fortified town of Yakutsk. As they settle the frozen expanse of northeastern Siberia, the Russians too come into contact with people who speak a language quite distinct from their own. These people call themselves Sakha (with the stress on the last syllable) and the language – Sakha Tyla (also with the stress on the last syllable in both words), but today the better-known name for this group and their language is Yakut. As with Powhatan and English, Yakut is quite novel for the Russian speakers. It has some sounds that are unfamiliar to the Russian ear, such as the front rounded vowels [y] and [ø], as in the French words chute ‘fall’ and peu ‘few’, respectively. Words are completely unrecognizable and can often be quite long. Sentences have the Subject-Object-Verb (SOV) order that would not surprise Powhatan speakers but is peculiar for the Russians. So like the English settlers in North America, the Russians start compiling word lists and recording texts in Yakut.

Yet, curiously, the first printed text in Yakut was not in a Russian book, but in a treatise titled “Noord en Oost Tartarye” (“North and East Tataria”), published by the Dutch cartographer Nicolaes Witsen in 1692 in Amsterdam. Today, Yakut fares much better than Powhatan: it is spoken by approximately 450,000 speakers (2010 Russian population census).

A decade after the Russians founded Yakutsk, another Dutchman, Abel Tasman – a seafarer, explorer, and merchant in the services of the Dutch East India Company (VOC in Dutch) – sails to New Zealand, Tonga, and Fiji. There, he and his men encounter people who speak languages quite distinct from Dutch but similar to each other: Maori, Tongan, and Fijian. Once again, the words in these
languages Strike the Dutch explorers as different, and so do the grammatical patterns. For example, sentences in these languages typically start not with the grammatical subject, as do sentences in Dutch, as well as English, Russian, Yakut, and Powhatan, but with the verb. And like the English settlers and the Russians, the Dutch are so staggered by the dissimilarities between their own language and the newly discovered ones that they start compiling word lists and grammars, which laid the foundation for the later realization that all of these languages—and over a thousand others—belong to the same language family, the Austronesian family (discussed in more detail in Chapter 9). Today, Maori is reported to have 150,000 speakers (at different degrees of proficiency) in New Zealand, Tongan is spoken by 96,000 people in Tonga and an additional 73,000 speakers elsewhere, and Fijian by approximately 340,000 people in Fiji.

Remarkably, these encounters between the English and the Powhatan Indians, the Russians and the Yakut, the Dutch and the Austronesians were not isolated phenomena. In fact, such encounters between speakers of very different languages happened over and over again in the course of human history, whenever one group moved to a new territory and encountered another group; after all, no reports have ever been made of any human tribe, ethnicity, or people that did not have a language. Whether these encounters between different linguistic groups were peaceful or otherwise, they naturally led to linguistic curiosity on both sides, linguistic interaction, and ultimately to changes in the languages of both groups. This book is about diverse human languages and the peoples who speak them, in what ways languages differ from one another and in what ways they are similar, how these languages came to be spoken where they are now spoken, and how they interacted and changed each other.

When they encounter a group speaking a language different from their own, people are typically first struck by the differences between their own language and the new language. That was the case with the English, the Russians, and the Dutch, as noted above—and doubtlessly, that is how the Powhatan Indians, the Yakut, and the Polynesians felt too. Yet the similarities between languages are also interesting. Although it is tempting to focus on the differences between languages, their peculiarities and the “exotic” elements found in some languages but not in most, in this book we will also examine patterns of commonality across languages. After all, the “exotic” can only be understood in contrast to the “mundane.” It is important to note that sometimes what is perceived as “exotic” versus “mundane” depends on our frame of reference: English—which seems pretty “normal” and “typical” to its speakers—turns out to be quite exotic in certain ways when compared to other languages around the world. For example, its “th” sounds are fairly unusual, as is its large vowel inventory.
An investigation of the world’s languages can also shed new light on the question of the history and prehistory of the peoples who speak these languages. As we will see throughout this book, linguistic studies have been instrumental in figuring out the past of Native Americans, the Yakut, and the inhabitants of the South Sea Islands, as well as the Hungarians, the Saamis, the Gypsies, and many other groups. One of the goals of this book is to show that a study of human languages, enhanced by evidence from other disciplines, such as anthropology, archeology, history, and genetics, leads us to a better understanding of the human condition.

Most of this book (Chapters 2–11) is organized by part of the world, as shown in Figure 1.1. This division used here is based not on the familiar division into continents like Eurasia or even geopolitical regions like Europe, but on geolinguistic factors. (For a critique of geographical division of the world into continents, see e.g. Lewis & Wigen 1997.) Thus, parts of the Middle East and South Asia are considered in the same chapter (Chapter 3), which concerns both the Indo-European languages (Section 3.1) and the non-Indo-European languages of this region (Sections 3.2 and 3.3). Conversely, languages of Africa are discussed in two different chapters: languages of North Africa are introduced in Chapter 6 together with languages of the Greater Middle East, while languages of sub-Saharan Africa are the topic of Chapter 7. Chapters 2 through 11 also contain “Focus on” sections concerned with either general issues, such as language contact or language change, or controversies surrounding specific languages, such as Pirahã. The last chapter (Chapter 12) deals with three issues that did not fit comfortably

---

**Fascinating Language: Scots**

Scots is the closest relative of English, spoken in southern and central (lowlands) parts of Scotland. Until the Renaissance period, English and Scots were the same language, gradually growing more and more distinct until they became no longer mutually intelligible (at least, not easily so). Although some people still dispute the language status of Scots and consider it a dialect of English, it is hardly more similar to English than Danish is to Norwegian, which are generally considered distinct languages. One familiar example of Scots is the song “Auld Lang Syne,” written by Robert Burns in 1788:

An sheerly yil bee yur pynt-staup!
An sheerly al bee myn!
An will tak a cup of kyndnes yet,
Fir auld lang syne.

How much of it do you understand without a translation? What is *sheerly*? Or *bee*? (It’s not a familiar insect!) Another poet who wrote in Scots is Robert Louis Stevenson. While he wrote most of his prose in English, much of his poetry is in Scots, including a poem called “The Maker to Posterity” (Maker is pronounced “mah-ker” and means ‘poet’). In it, Stevenson describes Scots as a lofty language alongside Greek and Latin (Lallans – from “lowlands” – is one of the several names for Scots, and Tantallon is a mid-fourteenth-century castle overlooking the Firth of Forth):

No bein fit to write in Greek,
I wrote in Lallans,
Dear to my hert as the peat reek,
Auld as Tantallon.

This poem also illustrates one of the characteristic features of Scots pronunciation: the non-distinct pronunciation of the vowels in words like *pin* and *pen* – both are pronounced as most Americans or Brits would say *pen*. So *fit* is pronounced in Scots the same as *fête.*
elsewhere throughout the book: controversial hypotheses about large macro families, sign languages, and constructed languages. Each chapter also contains a “Fascinating Language” sidebar, such as the one on Scots in this chapter.

Before we proceed to examine languages in various parts of the world, a more general question of what language is and how many languages exist in the world must be addressed. In the remainder of this chapter, we will consider how languages diversify, how we can establish which languages are related, and how languages have been mapped.

1.1 Languages, Dialects, and Accents

Ferdinand de Saussure, a Swiss linguist considered to be one of the fathers of modern linguistics, in his *Cours de linguistique générale* (*Course in General Linguistics*, 1916) defines language as “a product of the collective mind of linguistic groups.” But this definition hardly helps us in drawing the boundary between one language and another: how can we tell who is or is not to be included in any given “linguistic group”? Take any two people, even close relatives, and they are sure to speak at least slightly differently. Yet it is not insightful to say that there are as many languages in the world as there are individual people!

Another way of defining languages is in geopolitical terms, as in the popular aphorism, commonly attributed to Max Weinreich (although there is some debate as to whether he actually coined it or just published it): “A language is a dialect with an army and navy.” Yiddish, the language Weinreich studied extensively,
was considered at the time a mere dialect of German − because it had never had an army and a navy, Weinreich contended. Indeed, it is often the case that we consider two linguistic varieties as distinct languages (rather than dialects of the same language) when they are associated with distinct flags and other trappings of a nation state. For example, a language that was known up to the beginning of 1990s as Serbo-Croatian has since then “divided” into not just two but four languages, each claiming distinctness from the others and trying as hard as they can to purge each other’s influences: Serbian, Croatian, Bosnian, and Montenegrin. Similarly, the differentiation between Danish, Norwegian, and Swedish as three separate languages might not have existed were it not for the fact that these are spoken in three different countries, because linguistically all three are so similar that linguists sometimes refer to them as one language, Mainland Scandinavian.

Conversely, many countries are multilingual. For example, Belgium has three distinct linguistic zones: the Flemish (sometimes called “Dutch”) zone in the north, the French (or “Walloon”) zone in the south, and the German-speaking zone in the east. Brussels, the capital of Belgium as well as the seat of the European Parliament, is a bilingual city in its own right and can be considered a fourth linguistic zone of Belgium. Likewise, Switzerland has four linguistic zones: French-speaking in the west, German-speaking in the north and center, Italian-speaking in the southeast, and Romansch-speaking in the east. Some countries are more multilingual than that: for example, Nigeria boasts 525 languages, Indonesia 719 languages, and Papua New Guinea a whopping 851 languages! Thus, if the geopolitical definition of language may work − albeit poorly − for Europe, it does not work at all for other parts of the world, which are far more linguistically diverse than Europe.

Consequently, linguists prefer the definition of language in terms of “mutual intelligibility”: if two linguistic varieties are mutually intelligible, they are considered dialects of the same language, and if they are not, they constitute distinct languages. However, even this definition of language-versus-dialect is not without problems. Most obviously, mutual intelligibility is a matter of degree and is relative to a text or situation: the same two speakers may have an easier or harder time understanding each other depending on the topic of conversation and even on how they phrase what they are saying. Moreover, the degree of mutual intelligibility or similarity between languages depends on who assesses it: a person who does not speak either of the languages is more likely to perceive
similarities than differences between them, while a person speaking one of the languages would focus more on the differences and would, as a result, assess the languages as more different than a non-speaker would. Finally, “mutual intelligibility” is not always mutual: depending on exposure to the other language, a speaker of one language may have an easier time understanding a speaker of another language than the other way around. For example, most speakers of Ukrainian have no problem understanding Russian, but the average Russian—who has not been exposed to much Ukrainian—might understand only bits and pieces of their interlocutor’s Ukrainian speech.

But the problem actually runs deeper: when gauging the degree of mutual intelligibility, what we are comparing is (snippets of) texts, rather than languages, which are cognitive systems of rules in the minds of the speakers that allow them to produce such texts. To illustrate what I mean by this, let’s consider the following sentences in English and Norwegian:

\[
\begin{align*}
1.1 & \text{ Languages, Dialects, and Accents} \\
(1-1) & \text{ English: } \text{We shall sing tomorrow.} \\
& \text{Norwegian: } \text{Vi skal synge i morgen.}
\end{align*}
\]

Word for word, these two sentences are very much parallel: \text{we} ~ \text{vi}, \text{shall} ~ \text{skal}, \text{sing} ~ \text{singe}, \text{tomorrow} ~ \text{i morgen}. Note also that the word order is the same in both English and Norwegian sentences. But now let’s rephrase those sentences to start with ‘tomorrow’:

\[
\begin{align*}
(1-2) & \text{ English: } \text{Tomorrow we shall sing.} \\
& \text{Norwegian: } \text{I morgen skal vi synge.}
\end{align*}
\]

The words, of course, remain the same, but the order of the words is now different: in English, \text{tomorrow} is followed by the grammatical subject \text{we}, which is in turn followed by the auxiliary verb \text{shall}, while in Norwegian \text{i morgen} ‘tomorrow’ is followed by the auxiliary verb \text{skal} ‘shall’, which is in turn followed by the subject \text{vi} ‘we’. In fact, in the absence of intonation (in speech) or punctuation marks (in writing), the Norwegian sentence in (1–2) may be taken by an English speaker to be a question: ‘Tomorrow, shall we sing?’

What we see in these examples is that the degree of mutual intelligibility (or similarity) may be dependent on the actual phrasing: the sentences in (1-1) are much more similar than those in (1-2). Why such a discrepancy? To understand it, we need to examine not sentences, but the grammatical rules that underlie them. And such rules differ between English and Norwegian in a consistent way: the English word order in both (1-1) and (1-2) is achieved by placing the auxiliary verb after the subject, while the Norwegian word order in both sentences is achieved by placing the auxiliary verb in the second position,
regardless of whether the first position is occupied by the subject (vi ‘we’) or an adverb (i morgen ‘tomorrow’). Linguists refer to this rule that we see in the Norwegian examples as verb-second, or V2 for short. Two different rules may, on occasion, produce very similar outputs, as in (1-1), creating an impression of a greater similarity between two languages than really exists. Conversely, an impression of a greater dissimilarity may be created by heavily using dialectal words or dialectal pronunciation features.

The task of drawing boundaries between dialects and languages is even more difficult because of the phenomenon known as a dialect continuum, which is a range of dialects that are spoken across some geographical area, with the dialects of neighboring areas differing from each other only slightly, and the dialects from the opposite ends of the continuum being much less similar to each other and possibly not mutually intelligible at all. Think of it as a “game of telephone” when one player whispers a word to the next person in the chain, who in turn whispers it to the next person and so on: what each person whispers to the next is quite similar to what was whispered to them, but what the last person in the chain hears may be quite different from what the first person said.

One example of a dialect continuum is the so-called Continental West Germanic dialect continuum, including all varieties of High German (spoken in the German-speaking parts of Switzerland, Austria and in southern parts of Germany, around Munich and Nuremberg), Middle German (spoken around Frankfurt-am-Main, Cologne and Dresden), and Low German (spoken in the northern parts of Germany, around Bremen, Hamburg and Kiel), as well as Dutch and Flemish. If one travels through the countryside, say, from Bern through Munich, Frankfurt, Hamburg and into Antwerp or Bruges, one will encounter many local linguistic varieties, each of which is quite similar to the previous one, but a person from Bern and a person from Antwerp will not be able to understand each other if each of them speaks in their local variety. And just a few generations ago, the same was true even for two German speakers, one from Munich and one from Hamburg: the only way they were able to converse was to revert to Standard German, the variety used in education, the media and administration (today, Hamburg is mostly High German-speaking).

While German and Dutch dialects constitute a dialect continuum, the boundary between what is called German and what is called Dutch is rather arbitrary and based to a large degree on geopolitical divisions rather than on linguistic factors. Thus, Low German dialects in northern Germany are in some ways more similar to Dutch varieties across the border than to High German dialects in southern Germany; it is a shared “army and navy”

Did you know? The fricative consonant [x] is not commonly found in English. It is the final sound in the German pronunciation of Bach.
(as in Weinreich’s definition) that makes most people consider High German and Low German dialects of the same language at all. The similarity between Low German and Dutch, on the one hand, and the difference between High German and Low German, on the other hand, can be illustrated by the pronunciation of certain consonants. Where both Low German and Dutch have stop consonants such as [k, t, p], as in machen ‘to make’, dat ‘that’ and dorp ‘village’, High and Middle German have fricative consonants such as [x, s, f], as in machen ‘to make’, das ‘that’ and dorf ‘village’ (the German/Dutch word dorp/dorf has a cognate in English placenames like Mablethorpe and Scunthorpe, both in Lincolnshire). The imaginary line between the Low German and Dutch varieties with [k], on the one hand, and High and Middle German varieties with [x], on the other, is known as the Benrath line (or, more informally, machen–maken line). Similarly, the [t]/[s] and [p]/[f] lines run through Middle German dialects but they do not coincide precisely with the [k]/[x] (Benrath) line.

Generally, such geographical boundaries of a certain linguistic feature – be it the pronunciation of a consonant or vowel, the choice of a certain word, or the use of some syntactic construction – are known as isoglosses. Major dialects or even groups of dialects are demarcated by bundles of isoglosses. Importantly, dialectal divisions as defined by isoglosses need not coincide with language boundaries based on geopolitical realities: for example, Galician, spoken in the northwestern corner of Spain and considered by some to be a dialect of Spanish, is on the same side of many isoglosses as Portuguese varieties and not as other Spanish dialects. Moreover, if one travels from northern Portugal through northern Spain, (southern) France and into Italy, one encounters a series of dialects each of which is similar to the neighboring ones but quite distinct from the ones further away. Hence, this area also forms a dialect continuum, known as the Western Romance dialect continuum.

Dialect continua (plural of continuum) are found not only in Western Europe, but in many other parts of the world. Elsewhere in Europe, we find dialect continua among varieties of East Slavic languages (Russian, Byelorussian, Rusyn and Ukrainian) and among varieties of South Slavic languages (Slovenian, Croatian, Bosnian, Serbian, Macedonian and Bulgarian); see Chapter 2. Outside of Europe, dialect continua have been described in the Persian-speaking area (see Chapter 3), the Turkic-speaking world (see Chapter 4), and the Arabic-speaking parts of North Africa and the Middle East (see Chapter 6), as well as among varieties of Algonquian languages in the northeastern United States and Canada, and among varieties of Eskimo-Aleut languages in Alaska and northern Canada (see Chapter 11), and elsewhere.

While the distinction between dialects and languages is drawn based on mutual intelligibility, a finer distinction is sometimes drawn between dialects
and accents (note that this distinction is commonly used in Britain but not as commonly in the USA). Dialects can differ from each other in many ways, including pronunciation, word meaning and use, and grammatical features, while the term “accent” is reserved for varieties solely with distinct pronunciation patterns. Accents are typically very local and may be limited to a single city or a small rural area. One example of a local accent is the Liverpool accent: some of its characteristic features include using a fricative consonant in place of a stop (for example, [bajx] instead of [bajk] for “bike”), as well as using the same vowel as in the words full and put for words like love and blood. (The non-distinct pronunciation of the vowels in full and love – and in similar words – is typical for many northern English dialects and accents.)

In contrast, the local variety of English found in Newcastle is not just an accent but is a full-blown dialect (known as Geordie). Geordie has its own characteristic pronunciation patterns such as using the aforementioned vowel of full in words like love, as well as its own characteristic grammatical features such as allowing two modals in a row (as in She might could come tomorrow) and using past-tense forms of the so-called irregular verbs instead of participle forms (as in I’ve took it or You done it, did you? – see Trudgill 1999: 13). This lack of distinction between past-tense and participle forms of irregular verbs is not as strange as it might appear: think about the so-called regular verbs (such as play, jump, work) – in Modern English, which do not distinguish past-tense and participle forms (unlike in other Germanic languages and earlier forms of English; see Eide 2009).

1.2 Language Families

As we have seen in Section 1.1, several mutually intelligible dialects (or even local accents) can be viewed as constituting one bigger linguistic variety, often referred to as a language. For example, Canadian English, Scottish English and Australian English, and even more local dialects/accents like New York City English, Liverpool English and Geordie (Newcastle English) can be considered together under the heading of the English language. In a similar fashion, several related languages may be seen as constituting a language family. For example, German, Dutch, and English are all members of the West Germanic language family.

What does it mean for languages to be related? Let’s first consider what it means for two human beings to be related. You call people your relatives if they are known to have a common ancestor with you: your siblings share the same parents with you, your cousins share the same grandparents, and so on. Similarly, two languages are considered related (or “members of the same language