1 Introduction

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 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 sound different: for example, the word for 'sun' is either *nepass* or *keshowghes*, and the word for 'copper' is either *matassen* or *osawas*. And it is not just the words that are different, but 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 pattern would not surprise 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 attributed to the Eastern Algonquian branch of the Algonquian language family (see Chapter 11.1) – derives from the descriptions of seventeenth- and eighteenth-century writers, as the language died out 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. Twenty-five years 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 land on the shores of the Lena River and, on September 25, 1632, found 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*, 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; and sentences have the Subject-Object-Verb (SOV) order that would not surprise Powhatan speakers, but is peculiar for the Russians.

Did you know? Unlike the English, who had never encountered anything like the Powhatan language, Russians had had some exposure to languages related to Yakut (that is, Turkic languages; see Chapter 6). However, their exposure to other Turkic languages was so limited that Yakut sounded quite exotic to the Russian ear.

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 Tatars"), published by the Dutch cartographer Nicolaes Witsen in 1692 in Amsterdam. Today, Yakut fairs much better than Powhatan: it is spoken by approximately 350,000–450,000 speakers.

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 in 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 numerous others – belong to the same language family, the Austronesian family (discussed in more detail in Chapter 9). Today, Maori is spoken by 60,000 people in New Zealand, Tongan by 96,300 people in Tonga and Fijian by 360,000 people in Fiji.

Nor were these encounters between the English and the Powhatan Indians, the Russians and the Yakut, the Dutch and the Austronesians isolated phenomena. In fact, such encounters between speakers of very different languages have happened over and over again in the course of human history, whenever one group has moved to a new territory and encountered another group; after all, no reports have ever been made of any human group that did not have 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, changes in the languages of both groups. This book is about diverse human languages and the peoples who speak them, how these languages came to be spoken where they are now spoken, and how they interacted with and changed each other.

While people are typically first struck by the differences between their own language and another language they come across, as were the English, the Russians and the Dutch in the encounters described above, it is also the similarities between languages that are 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".

An investigation into the world's languages can also shed new light on the question of the relatedness 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 Lapps, the Gypsies and many other groups. In recent years, the toolkit of a historian of human populations – which already contained tools from archeology and linguistics – has been enriched by the addition of new genetic methodologies. Sometimes, the new evidence from genetic studies provides additional support for the conclusions of linguists, and in other cases genetic studies contradict linguistic ones – in this book, we will review examples of both. Thus, 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 around different parts of the world, defined mostly not by the familiar division into continents like Eurasia or even geopolitical regions like Europe, but based on geolinguistic factors. (For a critique of the geographical division of the world into continents, see for instance Lewis and 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 (section 3.2). Conversely, languages of Africa become the subject of two different chapters: languages of North Africa are introduced in Chapter 6 together with languages of the Middle East and Central Asia, while languages of sub-Saharan Africa are the topic of Chapter 7. The last chapter (Chapter 12) is dedicated to the issue of macro families. Chapters 2 through 11 also contain "Focus on" sections concerned either with general issues, such as field linguistics or language change, or controversies surrounding specific languages, such as Dyirbal or Pirahã. But before we can examine languages in various parts of the world, a more general question of what language is and how many languages exist must be addressed.

1.1 Languages, Dialects and Accents

Ferdinand de Saussure, a Swiss linguist considered to be one of the fathers of twentieth-century 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 does not help much in drawing the boundary between one language and another: after all, it is not clear 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 the Yiddish linguist 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 was considered at the time a mere dialect of German – because it 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 national state. For example, a language that was known up to the beginning of the 1990s as Serbo-Croatian has recently "broken" into not just two but four languages, each claiming distinctness from the others and attempting 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.

Conversely, many countries are multilingual. For example, Belgium has three distinct linguistic zones: the Flemish (sometimes called "Dutch") zone in the north, the Walloon (or French) zone in the south and the German-speaking zone in the east. Brussels, 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, such as Nigeria, Indonesia and Papua New Guinea, have hundreds of languages spoken there.

Because of these discrepancies between linguistic varieties and geopolitical divisions, 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 vs. 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. Furthermore, 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 rather 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, the speaker of one language may have an easier time understanding a speaker of another language than the other way around. For example, most Ukrainians have no problem understanding Russian, but the average Russian – who has not been exposed to much Ukrainian – might understand only bits and pieces of his interlocutor's 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:

(1-1) English: *We shall sing tomorrow.*
Norwegian: *Vi skal synge i morgen.*

Word for word, these two sentences are very much parallel: *we/vi*, *shall/skal*, *sing/synge*, *tomorrow/i morgen*. Note also that the word order is the same in both English and Norwegian. But now let's rephrase those sentences to start with 'tomorrow':

(1-2) English: *Tomorrow we shall sing.*
Norwegian: *I morgen skal vi synge.*

The words, of course, remain the same, but the order of the words now differs: in English, *tomorrow* is followed by the grammatical subject *we*, which is in turn followed by the AUXILIARY verb *shall*, while in Norwegian *i morgen* 'tomorrow' is followed by the auxiliary verb *skal* 'shall', which is in turn followed by the subject *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 from English to 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 in Norwegian 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 of DIALECT CONTINUUM, when a range of dialects is 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" (*aka* "Chinese whispers"), where 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, in Austria and in the 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, for example, from Bern through Munich, Frankfurt and 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).

Note also that 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" (as in Weinreich's definition) that makes most people consider them dialects of the same language at all. For instance, where both Low German and Dutch have STOP CONSONANTS such as [k, t, p], as in *ma**k**en* 'to make', *da**t*** 'that' and *dor**p*** 'village', High and Middle German have FRICATIVE CONSONANTS such as [x, s, f], as in *ma**ch**en* 'to make', *da**s*** 'that' and *dor**f*** 'village' (compare also the German/Dutch *dorp/dorf* with the English *thorp*). The imaginary line between the Low German/Dutch varieties with [k], on the one hand, and the 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.

Did you know? The fricative consonant [x] is not commonly found in English. It is the final sound in the German pronunciation of *Bach*.

Generally, such geographical boundaries of a certain linguistic feature – be it the pronunciation of a consonant or a vowel, a certain lexical choice or the use of some syntactic construction – are known as ISOGLOSSES. Typically, major dialects or even groups of dialects are demarcated by a bundle of such isoglosses. Note that 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 is known as the Western Romance dialect continuum.

Dialect continua are found not only in Western Europe, but in many other parts of the world. Elsewhere in Europe, we find dialectal continua among varieties of East Slavic languages (Russian, Byelorussian, Rusyn and Ukrainian) and among varieties of South Slavic languages (Slovene, Croatian, Bosnian, Serbian, Macedonian and Bulgarian). Outside of Europe, dialect continua have been described in the Turkic-speaking world (see Chapter 6), the Arabic-speaking parts of North Africa and the Middle East (see Chapter 6), and the Persian-speaking area (see Chapter 3), 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), to mention just a few examples.

While the distinction between dialects and languages is drawn on the basis of 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 United States). 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*.

Did you know? 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), with characteristic pronunciation patterns such as using the aforementioned vowel of *full* in words like *love*, as well as with 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 participial forms (as in *I've took it* or *You done it, did you?*; see Trudgill 1999: 13).

Did you know? The lack of distinction between past tense and PARTICIPLE forms of irregular verbs is not as strange as it might sound: think about the so-called regular verbs (such as *play*, *walk*) – in Modern English, they 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

We have seen in the previous section that 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, Swiss German, Bavarian German and Plattdeutsch (Low German) can be grouped under the heading of the German language. Similarly, dialects (or possibly even dialect groups) like 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 bunched 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, Frisian and English are all members of the West Germanic language family. As with human biological families, a language family is a phylogenetic unit: a classification of languages into a language family implies that they descend from a common parent language, known as a PROTO-LANGUAGE. Thus, German, Dutch, Frisian, English and other members of the West Germanic family all descended from a common parent language, known as Proto-West-Germanic.

Did you know? It is common to name the parent language of a family X as "Proto-X" although, as we will see below, this is not always the case.

Just as with human biological families, a given language family can be viewed as part of a larger, more extended family (for example, your nuclear family is part of a larger family including grandparents, aunts, uncles and cousins). For example, the West Germanic family is part of a larger Germanic family. The common parent language of the Germanic family is – you guessed it! – Proto-Germanic. The Germanic family includes two other branches or subfamilies: North Germanic and East Germanic. (Unfortunately, unlike in the biological species classification, in linguistics the term "family" is not reserved for a particular level in the classification tree, which can lead to some degree of confusion.) The ancestral language of the East Germanic family is known as Proto-East-Germanic, whereas the ancestor of the North Germanic languages could be called Proto-North-Germanic; however, the latter is more commonly known from historical records as Old Norse.

The Germanic family, in turn, is part of an even larger language family called the Indo-European family – and, yes, its common ancestral language is known as Proto-Indo-European, or PIE for short (we will discuss the Indo-European language family in more detail in Chapter 2). This family relationship is schematized by the family tree in Figure 1.1.

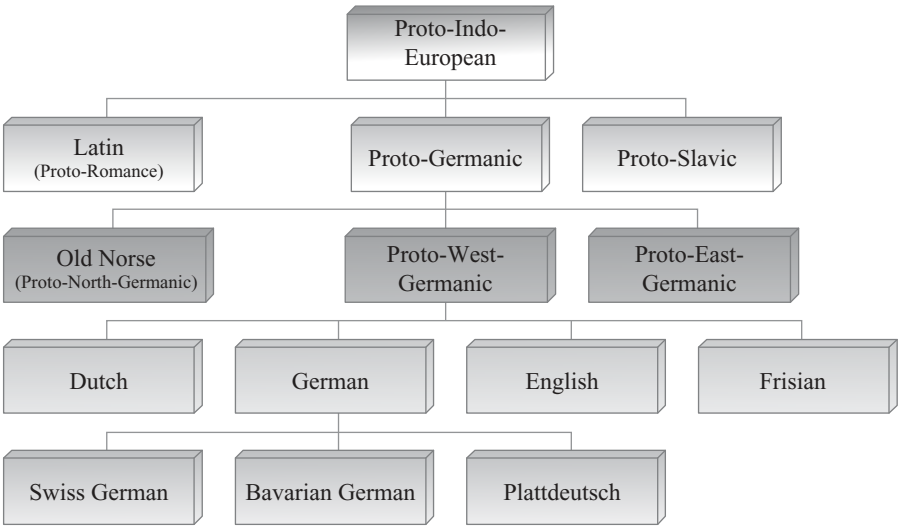

Figure 1.1 Partial Tree of Family Relationships among Indo-European Languages

Just like members of human biological families, which typically share observable characteristics, such as facial features, skin color, a predisposition to certain medical conditions and so on, so do languages in a given language family share certain observable linguistic characteristics, such as words, sounds and grammatical patterns. For example, almost all the languages in the Germanic family share the common Verb-Second pattern mentioned in the previous section: whether the sentence starts with a subject or an adverb such as 'yesterday' (or even a grammatical object), the verb must come immediately after, in the second position in the sentence. This pattern is found, for example, in Dutch, German and Swedish (see an additional example of the Verb-Second pattern in Norwegian in (1-2) above).

(1-3) a. Dutch: *Gisteren* ***las*** *ik dit boek.*
yesterday read I the book
b. German: *Gestern* ***las*** *ich dieses Buch.*
yesterday read I the book
c. Swedish: *Igår* ***läste*** *jag denna bok.*
yesterday read I the book

In fact, English is just about the only Germanic language that does not rely on the Verb-Second pattern: in English, if the sentence starts with anything but the grammatical subject, the verb must follow the subject, thus coming in the third, rather than second, position in the sentence:

(1-4) English: Yesterday I **read** the book.