Or consider a different kind of statement: "I'm going to the bank with a fishing pole." Most likely, this means that the speaker is going to a river bank and is carrying a fishing pole. But it could also mean that the speaker is going to a financial institution, carrying a fishing pole, or it could mean that the speaker is going to a financial institution known for its fishing pole – or even that the river bank the speaker is going to has some sort of notable fishing pole on it. We reason out a preferred meaning based on what we know about the world, but a computer does not know much about the world. How, then, can it process natural language?

From the other side of things, let us think for a moment about what you may have observed a computer doing with natural language. When you get a spam message, your email client often is intelligent enough to mark it as spam. Search for a page in a foreign language on the internet, and you can get an automatic translation, which usually does a decent job informing you as to what the site is about. Your grammar checker, although not unproblematic, is correct a surprising amount of the time. Look at a book's listing on a site that sells books, like Amazon, and you may find automatically generated lists of keywords; amazingly, many of these words and phrases seem to give a good indication of what the book is about.

If language is so difficult, how is it that a computer can "understand" what spam is, or how could it possibly translate between two languages, for example from Chinese to English? A computer does not have understanding, at least in the sense that humans do, so we have to wonder what technology underlies these applications. It is these very issues that we delve into in this book.

1.1.1 Encoding language

There is a fundamental issue that must be addressed here before we can move on to talking about various applications. When a computer looks at language, what is it looking at? Is it simply a variety of strokes on a piece of paper, or something else? If we want to do anything with language, we need a way to represent it.

This chapter outlines the ways in which language is represented on a computer; that is, how language is encoded. It thus provides a starting point for understanding the material in the rest of the chapters.

If we think about language, there are two main ways in which we communicate – and this is true of our interactions with a computer, too. We can interact with the computer by writing or reading **text** or by speaking or listening to **speech**. In this chapter, we focus on the representations for text and speech, while throughout the rest of the book we focus mainly on processing text.

text

speech

1.2 Writing systems used for human languages

If we only wanted to represent the 26 letters of the English alphabet, our task would be fairly straightforward. But we want to be able to represent any language in any writing system, where a **writing system** is "a system of more or less permanent

writing system

marks used to represent an utterance in such a way that it can be recovered more or less exactly without the intervention of the utterer" (Daniels and Bright, 1996).

And those permanent marks can vary quite a bit in what they represent. We will look at a basic classification of writing systems into three types: alphabetic, syllabic, and logographic systems. There are other ways to categorize the world's writing systems, but this classification is useful in that it will allow us to look at how writing systems represent different types of properties of a language by means of a set of characters. Seeing these differences should illustrate how distinct a language is from its written representation and how the written representation is then distinct from the computer's internal representation (see Section 1.3).

For writing English, the idea is that each letter should correspond to an individual sound, more or less, but this need not be so (and it is not entirely true in English). Each character could correspond to a series of sounds (e.g., a single character for *str*), but we could also go in a different direction and have characters refer to meanings. Thus, we could have a character that stands for the meaning of "dog". Types of writing systems vary in how many sounds a character represents or to what extent a meaning is captured by a character. Furthermore, writing systems differ in whether they even indicate what a word is, as English mostly does by including spaces; we will return to this issue of distinguishing words in Section 3.4.

One important point to remember is that these are systems for writing down a language; they are not the language itself. The same writing system can be used for different languages, and the same language in principle could be written down in different writing systems (as is the case with Japanese, for example).

1.2.1 Alphabetic systems

We start our tour of writing systems with what should be familiar to any reader of English: alphabets. In **alphabetic systems**, a single character refers to a single sound. As any English reader knows, this is not entirely true, but it gives a good working definition. **alphabetic system**

We will look at two types of alphabetic systems. First, there are the **alphabets**, or phonemic alphabets, which represent all sounds with their characters; that is, both consonants and vowels are represented. Many common writing systems are alphabets: Etruscan, Latin, Cyrillic, Runic, and so forth. Note that English is standardly written in the Latin, or Roman, alphabet, although we do not use the entire repertoire of available characters, such as those with accents (e.g., *è*) or **ligatures**, combinations of two or more characters, such as the German ß, which was formed from two previous versions of *s*. **alphabet** **ligature**

As an example of an alphabet other than Latin, we can look at Cyrillic, shown in Figure 1.1. This version of the alphabet is used to write Russian, and slight variants are used for other languages (e.g., Serbo-Croatian). Although some characters correspond well to English letters, others do not (e.g., the letter for [n]). The characters within brackets specify how each letter is said – that is, pronounced; we will return to these in the discussion of phonetic alphabets later on.

а	б	в	г	д	е	ё	ж	з	и	й
[a]	[b]	[v]	[g]	[d]	[je]	[jo]	[ʒ]	[z]	[i]	[j]
к	л	м	н	о	п	р	с	т	у	ф
[k]	[l]	[m]	[n]	[o]	[p]	[r]	[s]	[t]	[u]	[f]
х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
[x]	[ts]	[tɕ]	[ʂ]	[ɕɕ]	[-]	[ɨ]	[ʲ]	[e]	[ju]	[ja]

Figure 1.1 The Cyrillic alphabet used for Russian

Some alphabets, such as the Fraser alphabet used for the Lisu language spoken in Myanmar, China, and India, also include **diacritics** to indicate properties such as a word's tone (how high or low pitched a sound is). A diacritic is added to a regular character, for example a vowel, indicating in more detail how that sound is supposed to be realized. In the case of Fraser, for example, *M*: refers to an [m] sound (written as *M*), which has a low tone (written as :).

diacritic

abjad

Our second type of alphabetic system also often employs diacritics. **Abjads**, or consonant alphabets, represent consonants only; some prime examples are Arabic, Aramaic, and Hebrew. In abjads, vowels generally need to be deduced from context, as is illustrated by the Hebrew word for "computer", shown on the left-hand side of Figure 1.2.

מחשב	מחשב	מחשב
b š x m	*b š x m*	*b š x m*
[maxʃev]	[mexuʃav]	[mexaʃav]
'computer'	'is digitized'	'with + he thought'

Figure 1.2 Example of Hebrew (abjad) text

The Hebrew word in its character-by-character transliteration *bšxm* contains no vowels, but context may indicate the [a] and [e] sounds shown in the pronunciation of the word [maxʃev]. (Note that Hebrew is written right to left, so the *m* as the rightmost character of the written word is the first letter pronounced.) As shown in the middle and right-hand side of Figure 1.2, the context could also indicate different pronunciations with different meanings.

The situation with abjads often is a little more complicated than the one we just described, in that characters sometimes represent selected vowels, and often vowel diacritics are available.

A note on letter–sound correspondence As we have discussed, alphabets use letters to encode sounds. However, there is not always a simple correspondence between a word's spelling and its pronunciation. To see this, we need look no further than English.

English has a variety of non-letter–sound correspondences, which you probably labored through in first grade. First of all, there are words with the same spellings representing different sounds. The string *ough*, for instance, can be pronounced at least five different ways: "cough", "tough", "through", "though", and "hiccough". Letters are not consistently pronounced, and, in fact, sometimes they are not pronounced at all; this is the phenomenon of silent letters. We can readily see these in "knee", "debt", "psychology", and "mortgage", among others. There are historical reasons for these silent letters, which were by and large pronounced at one time, but the effect is that we now have letters we do not speak.

Aside from inconsistencies of pronunciation, another barrier to the letter–sound correspondence is that English has certain conventions where one letter and one sound do not cleanly map to one another. In this case, the mapping is consistent across words; it just uses more or less letters to represent sounds. Single letters can represent multiple sounds, such as the *x* in "tax", which corresponds to a *k* sound followed by an *s* sound. And multiple letters can consistently be used to represent one sound, as in the *th* in "the" or the *ti* in "revolution".

Finally, we can alternate spellings for the same word, such as "doughnut" and "donut", and **homophones** show us different words that are spelled differently but spoken the same, such as "colonel" and "kernel". **homophone**

Of course, English is not the only language with quirks in the letter–sound correspondences in its writing system. Looking at the examples in Figure 1.3 for Irish, we can easily see that each letter does not have an exact correspondent in the pronunciation.

Spelling	Pronunciation	Meaning
samhradh	[sauruh]	'summer'
scri'obhaim	[shgri:m]	'I write'

Figure 1.3 Some Irish expressions

The issue we are dealing with here is that of **ambiguity** in natural language, in this case a letter potentially representing multiple possible sounds. Ambiguity is a recurring issue in dealing with human language that you will see throughout this book. For example, words can have multiple meanings (see Chapter 2); search queries can have different, often unintended meanings (see Chapter 4); and questions take on different interpretations in different contexts (see Chapter 6). In this case, writing systems can be designed that are unambiguous; phonetic alphabets, described next, have precisely this property. **ambiguity**

Phonetic alphabets You have hopefully noticed the notation used within the brackets ([]). The characters used there are a part of the International Phonetic Alphabet (IPA). Several special alphabets for representing sounds have been developed, and probably the best known among linguists is the IPA. We have been discussing problems with letter–sound correspondences, and phonetic alphabets help us discuss these problems, as they allow for a way to represent all languages unambiguously using the same alphabet.

Each phonetic symbol in a phonetic alphabet is unambiguous: the alphabet is designed so that each speech sound (from any language) has its own symbol. This eliminates the need for multiple symbols being used to represent simple sounds and one symbol being used for multiple sounds. The problem for English is that the Latin alphabet, as we use it, only has 26 letters, but English has more sounds than that. So, it is no surprise that we find multiple letters like *th* or *sh* being used for individual sounds.

The IPA, like most phonetic alphabets, is organized according to the articulatory properties of each sound, an issue to which we return in Section 1.4.2. As an example of the IPA in use, we list some words in Figure 1.4 that illustrate the different vowels in English.

bead: [bid]		boot: [but]
bid: [bɪd]		book: [bʊk]
bade: [be(i)d]	bud: [bʌd]	bode: [bo(ʊ)d]
bed: [bɛd]		bought: [bɔt]
bad: [bæd]		body: [bɑdi]

Figure 1.4 Example words for English vowels (varies by dialect)

At http://purl.org/lang-and-comp/ipa you can view an interactive IPA chart, provided by the University of Victoria's Department of Linguistics. Most of the English consonants are easy to figure out, e.g., [b] in "boy", but some are not obvious. For example, [θ] stands for the *th* in "thigh"; [ð] for the *th* in "thy"; and [ʃ] for the *sh* in "shy".

1.2.2 Syllabic systems

syllabic system **Syllabic systems** are like alphabetic systems in that they involve a mapping between characters and sounds, but the units of sound are larger. The unit in question is **syllable** called the **syllable**. All human languages have syllables as basic building blocks of speech, but the rules for forming syllables differ from language to language. For example, in Japanese a syllable consists of a single vowel, optionally preceded by at most one consonant, and optionally followed by [m], [n], or [ŋ]. Most of the world's languages, like Japanese, have relatively simple syllables. This means that the total number of possible syllables in the language is quite small, and that syllabic writing systems work well. But in English, the vowel can also be preceded by a sequence of **consonant cluster** several consonants (a so-called **consonant cluster**), and there can also be a consonant cluster after the vowel. This greatly expands the number of possible syllables. You could design a syllabic writing system for English, but it would be unwieldy and difficult to learn, because there are so many different possible syllables.

abugida There are two main variants of syllabic systems, the first being **abugidas** (or **alphasyllabary** **alphasyllabaries**). In these writing systems, the symbols are organized into families.

All the members of a family represent the same consonant, but they correspond to different vowels. The members of a family also look similar, but have extra components that are added in order to represent the different vowels. What is distinctive about an abugida is that this process is systematic, with more or less the same vowel components being used in each family.

To write a syllable consisting of a consonant and a vowel, you go to the family for the relevant consonant, then select the family member corresponding to the vowel that you want. This works best for languages in which almost all syllables consist of exactly one consonant and exactly one vowel. Of course, since writing is a powerful technology, this has not stopped abugidas from being used, with modifications, to represent languages that do not fall into this pattern. One of the earliest abugidas was the Brahmi script, which was in wide use in the third century BCE and which forms the basis of many writing systems used on the Indian subcontinent and its vicinity.

As an example, let us look at the writing system for Burmese (or Myanmar), a Sino-Tibetan language spoken in Burma (or Myanmar). In Figure 1.5, we see a table displaying the base syllables.

က [kə̰]	ခ [kʰə̰]	ဂ [gə̰]	ဃ [gə̰]	င [ŋə̰]
စ [sə̰]	ဆ [sʰə̰]	ဇ [zə̰]	ဈ [zə̰]	ည [ɲə̰]
ဋ [tə̰]	ဌ [tʰə̰]	ဍ [də̰]	ဎ [də̰]	ဏ [nə̰]
တ [tə̰]	ထ [tʰə̰]	ဒ [də̰]	ဓ [də̰]	န [nə̰]
ပ [pə̰]	ဖ [pʰə̰]	ဗ [bə̰]	ဘ [bə̰]	မ [mə̰]
ယ [yə̰]	ရ [yə̰] ([rə̰])	လ [lə̰]	ဝ [wə̰]	သ [θə̰]
	ဟ [hə̰]	ဠ [lə̰]	အ [ə̰]	

Figure 1.5 Base syllables of the Burmese abugida

As you can see in the table, every syllable has a default vowel of [ə̰]. This default vowel can be changed by adding diacritics, as shown in Figure 1.6, for a syllables that start with [k]. We can see that the base character remains the same in all cases, while diacritics indicate the vowel change. Even though there is some regularity, the combination of the base character plus a diacritic results in a single character, which distinguishes abugidas from the alphabets in Section 1.2.1. Characters are written from left to right in Burmese, but the diacritics appear on any side of the base character.

က [kə̰]	ကု [kṵ]	ကေး [kéi]	ကို့ [kò]
ကာ [kà]	ကူ [kù]	ကယ့် [kḛ]	ကိုး [kó]
ကား [ká]	ကူး [kú]	ကယ် [kè]	ကော့ [kɔ̰]
ကိ [ki̱]	ကေ့ [ke̱i]	ကဲ [ké]	ကော် [kɔ̀]
ကီ [kì]	ကေ [kèi]	ကို့ [ko̱]	ကော [kɔ́]
ကီး [kí]			

Figure 1.6 Vowel diacritics of the Burmese abugida

syllabary The second kind of syllabic system is the **syllabary**. These systems use distinct symbols for each syllable of a language. An example syllabary for Vai, a Niger-Congo language spoken in Liberia, is given in Figure 1.7 (http://commons.wikimedia.org/wiki/Category:Vai_script).

An abugida is a kind of syllabary, but what is distinctive about a general syllabary is that the syllables need not be organized in any systematic way. For example, in Vai, it is hard to see a connection between the symbols for [pi] and [pa], or any connection between the symbols for [pi] and [di].

1.2.3 Logographic writing systems

logograph The final kind of writing system to examine involves **logographs**, or logograms. A logograph is a symbol that represents a unit of meaning, as opposed to a unit of sound. It is hard to speak of a true logographic writing system because, as we will see, a language like Chinese that uses logographs often also includes phonetic information in the writing system.

To start, we can consider some non-linguistic symbols that you may have encountered before. Figure 1.8, for example, shows symbols found on US National Park Service signs (http://commons.wikimedia.org/wiki/File:National_Park_Service_sample_pictographs.svg). These are referred to as **pictographs**, or pictograms, because they essentially are pictures of the items to which they refer. In some sense, this is the simplest way of encoding semantic meaning in a symbol. The upper left symbol, for instance, refers to camping by means of displaying a tent.

pictograph

Some modern systems evolved from a more pictographic representation into a more abstract symbol. To see an example of such character change, we can look at the development of the Chinese character for "horse", as in Figure 1.9 (http://commons.wikimedia.org/wiki/Category:Ancient_Chinese_characters).

Figure 1.7 The Vai syllabary

Originally, the character very much resembled a horse, but after evolving over the centuries, the character we see now only bears a faint resemblance to anything horse-like.

There are characters in Chinese that prevent us from calling the writing system a fully meaning-based system. **Semantic-phonetic compounds** are symbols with a meaning element and a phonetic element. An example is given in Figure 1.10, where we can see that, although both words are pronounced the

semantic-phonetic compound

Figure 1.8 US National Park Service symbols (pictographs)

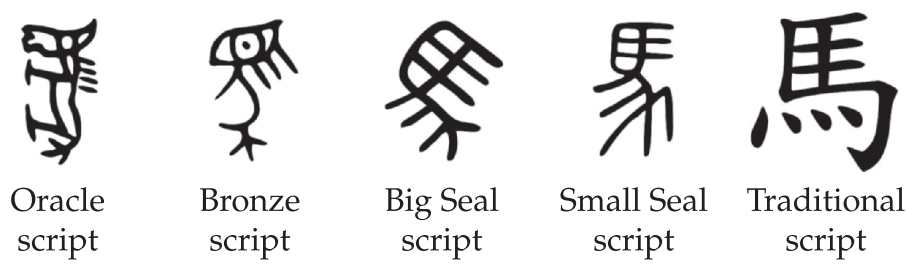

Figure 1.9 The Chinese character for "horse"

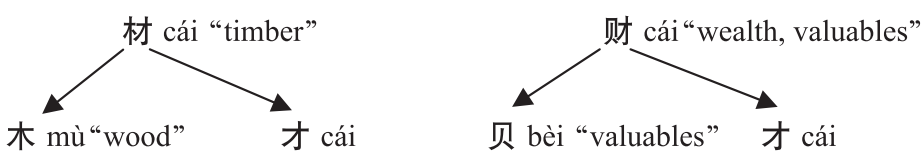

Figure 1.10 Semantic-Phonetic Compounds used in writing Chinese

same, they have different meanings depending on the semantic component. Of course, it is not a simple matter of adding the phonetic and semantic components together: knowing that the meaning component of a semantic-phonetic compound is "wood" by itself does not tell you that the meaning of the compound is "timber".

1.2.4 Systems with unusual realization

In addition to writing systems making use of characters differentiated by the shape and size of different marks, there are other writing systems in existence that exploit different sensory characteristics.

Perhaps best known is the tactile system of Braille. Braille is a writing system that makes it possible to read and write through touch, and as such it is primarily used by the blind or partially blind. We can see the basic alphabet in Figure 1.11 (http://commons.wikimedia.org/wiki/File:Braille_alfabet.jpg). The Braille system works by using patterns of raised dots arranged in cells of up to six dots, in a 3 x 2 configuration. Each pattern represents a character, but some frequent words and letter combinations have their own pattern. For instance, the pattern for *f* also indicates the number 6 and the word "from". So, even though it is at core an alphabet, it has some logographic properties.

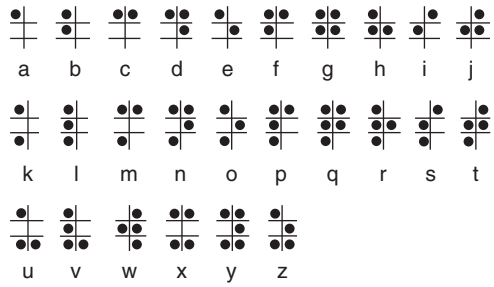

Figure 1.11 The Braille alphabet

An interesting case is the **chromatographic** writing system supposedly used by the Benin and Edo people in southern Nigeria (http://purl.org/lang-and-comp/chroma). This system is based on different color combinations and symbols. We have some reservations in mentioning this system, as details are difficult to obtain, but in principle both color and shape can encode pronunciation. **chromatographic**

1.2.5 Relation to language

As we mentioned before, there is no simple correspondence between a writing system and a language. We will look at two examples, Korean and Azeri, which will highlight different aspects of the unique ways languages are written.

Korean The writing system for Korean is a hybrid system, employing both alphabetic and syllabic concepts. The writing system is actually referred to as *Hangul* (or *Hangeul*) and was developed in 1444 during the reign of King Sejong. The Hangul system contains 24 letter characters, 14 consonants and 10 vowels. But when the language is written down, the letters are grouped together into syllables to form new characters. The letters in a syllable are not written separately as in the English system, but together form a single character. We can see an example in Figure 1.12 (http://commons.wikimedia.org/wiki/File:Hangeul.png), which shows how individual alphabetic characters together form the syllabic characters for "han" and "geul". The letters are not in a strictly left-to-right or top-to-bottom pattern, but together form a unique syllabic character. Additionally, in South Korea, *hanja* (logographic Chinese characters) are also used.

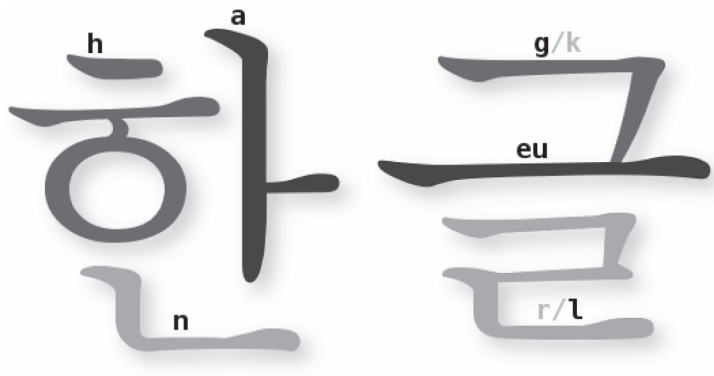

Figure 1.12 Composition of the characters for "Hangeul"

Azeri Azeri is a language whose history illustrates the distinction between a language and its written encoding. Azeri is spoken in Azerbaijan, northwest Iran, and Georgia, and up until the 1920s it was written in different Arabic scripts. In 1929, however, speakers were forced to switch to the Latin alphabet for political reasons. In 1939, it was decided to change to the Cyrillic alphabet, to bring Azeri more in line with the rest of the Soviet Union. After the fall of the USSR in 1991, speakers went back to the Latin alphabet, although with some minor differences from when they had used it before. Azeri is thus a single language that has been written in many ways.

1.3 Encoding written language

1.3.1 Storing information on a computer

Given the range of writing systems, we now turn to the question of how to encode them on a computer. But to address that, we have a more fundamental question: How do we encode anything on a computer?