

LINGUISTICS COURSE

Writing Systems of the World

A tour of the world's writing systems: logograms, syllabaries, abjads, alphabets, abugidas and invented scripts.

10 Chapters

Philip Osztromok

Generated with Claude

Contents

Chapter 1 What Is Writing? 2

Chapter 2 The First Scripts 7

Chapter 3 Chinese Characters 12

Chapter 4 Syllabaries 17

Chapter 5 Abjads 22

Chapter 6 Alphabets 27

Chapter 7 Abugidas 32

Chapter 8 Invented Scripts 38

Chapter 9 Spelling Rules 44

Chapter 10 Lost and Found 50

Writing Systems of the World

Course 1 · Chapter 1 · What Is Writing?

You can read this page because a set of symbols has been tied to a language by convention. Every script on Earth does the same job, but the ties differ: some symbols stand for meanings, some for syllables, some for single sounds. This chapter sets out the vocabulary for the whole course and asks where writing came from.

> **THE IDEA IN A BOX**
>
> Wikipedia defines a writing system as "any conventional system for representing a particular language using a set of symbols". The useful question to ask of any script is: *what does one symbol stand for?* The answer sorts almost every script in the world into a type.

Writing Is Not Language

- **Speech comes first.** The Writing system article says that while many spoken or signed languages are not written, there are no written languages without a spoken counterpart.
- **One language, several scripts.** A script is a set of symbols, and the same language can be written in more than one. (Serbian written in both Latin and Cyrillic letters is my own example, not the article's.)
- **One script, many languages.** The article calls the Latin alphabet by far the most common script used by writing systems.
- **Drawing is not writing.** On the narrow definition the article uses, drawings and maps are excluded. *Proto-writing* (ideographic and mnemonic symbols such as the Inca quipu) lacks the capability to fully encode language, and most proto-writing systems did not become writing systems.

The Vocabulary

The article's typology divides systems first by what the symbols represent: **morphographic** (meaning-based) or **phonographic** (sound-based). Phonographic systems split into syllabic and alphabetic ones, and alphabets into abjads, pure alphabets and abugidas. These are the six terms the rest of the course relies on.

TYPE	ONE SYMBOL STANDS FOR	EXAMPLE (CHAPTER)
Logographic	a unit of meaning, such as a word or morpheme	Chinese characters (3)
Syllabary	a syllable or mora	Japanese kana (4)
Abjad	a consonant only	Hebrew, Arabic (5)
Alphabet	a phoneme, roughly (consonants and vowels)	Greek, Latin (6)
Abugida	a consonant plus vowel, with the vowel marked by a predictable change to the symbol	Devanagari (7)
Featural	a sub-phonetic element (a feature of how a sound is made)	Hangul (8)

WATCH OUT: "ALPHABET" IS NOT A SYNONYM FOR "SCRIPT"

People say "the Chinese alphabet" or "the Japanese alphabet", but neither is an alphabet in this typology. Also, the article says most systems are mixed, combining sound-based and meaning-based elements, and that the featural class, proposed by Geoffrey Sampson, is disputed by several scholars, including John DeFrancis. The table is a map of tendencies, not sealed boxes.

How many symbols?

The type largely predicts the size of the symbol set. According to the article, alphabets typically use fewer than 100 symbols, syllabaries may use hundreds, and logographies thousands. It gives Japanese kana as covering roughly 100 moras and the Yi script as having 756 symbols. The World's Writing Systems project, which the article links to, lists 294 known writing systems.

Where Writing Began

The article describes writing as invented independently in several places. Dates are as the article gives them.

SYSTEM	PLACE	DATE (PER THE ARTICLE)
Cuneiform	Southern Mesopotamia (Sumerian)	Akkadian writing appears in significant quantities c. 2350 BC

SYSTEM	PLACE	DATE (PER THE ARTICLE)
Egyptian hieroglyphs	Egypt	Evolved from proto-writing between 3400 and 3100 BC; earliest coherent texts c. 2600 BC; generally regarded as independent of cuneiform
Chinese characters	Yellow River valley	c. 1200 BC, with no evidence of contact with the Near East
Mesoamerican scripts	Including Olmec and Maya	Also considered independent inventions
Indus script	Indus valley	c. 2600 to c. 2000 BC; undeciphered, and whether it was true writing is disputed
First alphabetic writing	Sinai Peninsula (a Semitic language)	c. 1800 BC; most alphabets descend from or were inspired by Proto-Sinaitic

The last row matters for the rest of the course: the great majority of alphabetic scripts, and several abjads and abugidas, trace back to one source. Chapters 5 to 7 follow those lines.

Seeing Script Types in Python

Unicode (met in Computational Linguistics 1) gives every character a name that includes its script. The example is my own; I ran it and copied the output.

```
import unicodedata
for ch in ["A", "Ж", "ع" ,"א", "क", "あ", "한", "語"]:
    print(ch, unicodedata.name(ch))
```

```
A LATIN CAPITAL LETTER A
Ж CYRILLIC CAPITAL LETTER ZHE
א HEBREW LETTER ALEF
ع ARABIC LETTER AIN
क DEVANAGARI LETTER KA
あ HIRAGANA LETTER A
한 HANGUL SYLLABLE HAN
語 CJK UNIFIED IDEOGRAPH-8A9E
```

Eight characters, eight names, and every one of the six types above is represented by at least one. The same word "water" in different scripts shows how the symbol count varies (again my own example, run and copied):

```
words = {"English": "water", "Greek": "νερό", "Hebrew": "מים",
         "Hindi": "पानी", "Japanese": "水", "Korean": "물"}
for lang, w in words.items():
    print(lang, w, len(w))
```

```
English water 5
Greek νερό 4
Hebrew 3 מים
Hindi पानी 4
Japanese 水 1
Korean 물 1
```

WATCH OUT: THIS COUNTS CODE POINTS, NOT SYMBOLS A READER SEES

Python's `len` counts Unicode code points. The Hindi word shows 4 because vowel signs are separate code points; how a reader groups them is a different question, taken up in Chapter 7. The Korean syllable block 물 counts as 1 here because the string holds the precomposed form. The numbers illustrate that "length" depends on what you count, not that one script is shorter than another.

NOTE ON SOURCES

This chapter uses Wikipedia's Writing system article, of which I read the first 100,000 of 154,316 characters; the rest was not read. Definitions, symbol counts and invention dates are the article's, and I have not checked them against the primary scholarship. The Serbian example, the chapter-to-type mapping and the Python examples are mine.

Try It Yourself

EXERCISE 1

Classify each as logographic, syllabary, abjad, alphabet, abugida or featural, using the table above: Greek, Hebrew, Japanese kana, Devanagari, Chinese characters, Hangul. Which classification does the article say is disputed?

View solution

EXERCISE 2

Run the first Python example with three characters of your own from other scripts (for example Ω, ა and ス). Which scripts do the names report, and does any character surprise you?

View solution

EXERCISE 3

In two or three sentences, explain why a drawing of a deer is not writing but a symbol that is always read as the word "deer" might be. Use the article's idea of proto-writing in your answer.

View solution

CHAPTER 1 QUICK REFERENCE

- Writing system: a conventional set of symbols for representing a particular language
- Types by what one symbol stands for: logographic (meaning), syllabary, abjad, alphabet, abugida, featural
- Most real systems are mixed; the featural class is disputed
- Independent inventions per the article: Mesopotamia, Egypt, China, Mesoamerica (Indus disputed); most alphabets descend from Proto-Sinaitic

COMING NEXT

Writing Systems of the World 2 looks at the first scripts, cuneiform and hieroglyphs, and the trick that turned pictures into sounds.

Writing Systems of the World

Course 1 · Chapter 2 · The First Scripts

Chapter 1 listed Mesopotamia and Egypt among the places where writing was invented. This chapter looks at both systems, cuneiform and Egyptian hieroglyphs, because they show the same problem being solved in two ways: a picture can stand for a thing, but only a sound can stand for a name, a grammatical ending or an idea with no picture.

> **THE IDEA IN A BOX**
>
> Both early scripts mix signs for words (logograms) with signs for sounds, and add silent signs (determinatives) that tell the reader which kind of word is meant. Writing took off when pictures began to be read for their *sound* as well as their meaning.

Cuneiform

Facts below are from Wikipedia's Cuneiform article, of which I read the first part.

- **Origin.** It grew out of pictographic proto-writing in the late 4th millennium BC, drawing on a Near Eastern clay-token accounting system. Writing is first recorded in Uruk, at the end of the 4th millennium BC. The article says the meaning and usage of the tokens is still a matter of debate.
- **From pictures to wedges.** Early signs were drawn with a sharpened reed stylus and lacked wedge shapes. In the mid-3rd millennium BC a wedge-tipped stylus pressed into clay produced the strokes that give the script its name.
- **Shrinking inventory.** Around 2800 BC syllabic use began, and the sign inventory fell from some 1,500 signs to some 600.
- **Medium.** Signs were mainly made on clay tablets. Most surviving tablets were baked by accidental fires.
- **Logo-syllabic.** Signs could stand for whole words (logograms) or for syllables, and many served both roles. Determinatives, such as signs marking gods, countries or cities, guided the reader to a word's category.
- **Many languages.** It began with Sumerian and was adapted for Akkadian, Elamite, Hittite and others. Old Persian used a simplified, separate set of signs.

- **Decipherment.** Old Persian was the first deciphered, beginning with Georg Friedrich Grotefend in 1802. Henry Rawlinson and Edward Hincks worked on the trilingual Behistun inscription, and the article says cuneiform was successfully deciphered by 1857.
- **The end.** The last known cuneiform inscription, an astronomical text, was written in 79/80 AD.

Egyptian Hieroglyphs

Facts below are from Wikipedia's Egyptian hieroglyphs article, of which I read the first part.

- **Origin.** They grew out of proto-literate symbol systems in the Early Bronze Age, around the 33rd century BC. The article dates the first decipherable sentence in the Egyptian language to the 28th century BC.
- **Size.** More than 1,000 distinct characters overall. About 900 were in use in the Middle Kingdom, and by the Greco-Roman period the known hieroglyphs exceeded 5,000.
- **Three kinds of sign.** Phonetic signs (24 single-consonant signs, plus signs for two or three consonants), logograms (the sun sign for "sun"), and determinatives (silent signs that clarify a word's meaning, since many words are homophones).
- **Vowels.** Vowels were normally not written, so some scholars call the script an abjad. The article says it is debated whether vowels were written at all.
- **Direction.** Usually read right to left, though left to right and alternating lines also appear; figures facing left or right show which way to read.
- **Everyday forms.** Hieratic and demotic are simplified cursive scripts developed from hieroglyphs and better suited to papyrus; hieroglyphs stayed in formal and monumental use.
- **The end.** The latest known inscription is the Graffito of Esmet-Akhom at Philae, dated 394 AD.

The Rebus Principle

The hieroglyphs article explains the key trick with an example: an eye picture could represent the word "eye" or the sound of the pronoun "I". Using a picture for its sound, ignoring its meaning, is the **rebus principle**. Once a script has it, any word can be written, including names and grammar. Here is the same idea in English, made up by me for illustration:

PICTURE	READ AS	WORD WRITTEN
bee + leaf	"bee" + "leaf"	belief
eye	"I"	I (the pronoun)
sun + flower	"sun" + "flower"	sunflower (a meaning-compound, not a rebus)

WATCH OUT: HIEROGLYPHS ARE NOT "PICTURE WRITING" ONLY

Popular accounts treat hieroglyphs as symbols for ideas. The article describes a mix of phonetic signs, logograms and determinatives, and quotes Champollion's 1822 description of the system as "figurative, symbolic, and phonetic all at once". The Rosetta Stone (found in 1799, with hieroglyphic, demotic and Greek texts) was the key to the decipherment that Champollion completed in the 1820s.

Comparison

FEATURE	CUNEIFORM	EGYPTIAN HIEROGLYPHS
Earliest	Uruk, end of 4th millennium BC	33rd century BC proto-literate symbols; 28th century BC first decipherable sentence
Type	Logo-syllabic	Logograms, consonant signs and determinatives; vowels normally unwritten
Medium	Clay tablets, reed stylus	Monuments; cursive forms on papyrus
Languages	Sumerian, Akkadian, Elamite, Hittite and more	Egyptian
Last known use	79/80 AD	394 AD

Both Scripts in Unicode

Both scripts have Unicode blocks (the cuneiform block is U+12000 to U+123FF, the hieroglyph block U+13000 to U+1342F, per the articles). The example is my own; I ran it and copied the output.

```
import unicodedata
for cp in [0x12000, 0x1202D, 0x13000, 0x13080]:
    ch = chr(cp)
    print(ch, "U+%05X" % cp, unicodedata.name(ch), len(ch.encode("utf-8")))
```

```
𒀀 U+12000 CUNEIFORM SIGN A 4
𒀭 U+1202D CUNEIFORM SIGN AN 4
𓀀 U+13000 EGYPTIAN HIEROGLYPH A001 4
𓂀 U+13080 EGYPTIAN HIEROGLYPH D010 4
```

Each sign takes 4 bytes in UTF-8 because both blocks are beyond the Basic Multilingual Plane. The hieroglyph names are catalogue numbers (A001, D010), a sign list used by Egyptologists; the cuneiform names use the sign's conventional reading (A, AN). That naming difference is my observation from the output.

NOTE ON SOURCES

This chapter uses Wikipedia's Cuneiform article (read 100,000 of 194,774 characters) and Egyptian hieroglyphs article (read 100,000 of 208,138 characters); the remaining parts were not read, and dates and counts are the articles' and unchecked against primary scholarship. The English rebus table, the comparison table and the Python example are mine.

Try It Yourself

EXERCISE 1

Classify cuneiform and Egyptian hieroglyphs using the six types from Chapter 1. Why does neither fit one box neatly?

View solution

EXERCISE 2

Invent two English rebus spellings of your own (a picture or two read for sound), then say what a determinative would add to make each unambiguous.

View solution

EXERCISE 3

Use Python to print the name of U+12001 and U+13001. What does each name tell you about how the two blocks are organised?

View solution

CHAPTER 2 QUICK REFERENCE

- Cuneiform: Uruk, end of 4th millennium BC; logo-syllabic; wedge strokes on clay; used for many languages; last known text 79/80 AD
- Hieroglyphs: from the 33rd century BC; logograms, consonant signs and determinatives; vowels normally unwritten; last inscription 394 AD
- Rebus principle: a picture read for its sound, which lets any word be written
- Determinative: a silent sign marking a word's category

COMING NEXT

Writing Systems of the World 3 turns to Chinese characters, the one early logographic system still in everyday use.

Writing Systems of the World

Course 1 · Chapter 3 · Chinese Characters

Cuneiform and hieroglyphs are no longer written. Chinese characters are, by a very large number of people, and they are the main surviving logographic system. This chapter looks at what a character stands for, how characters are built, and which other languages have borrowed them.

> **THE IDEA IN A BOX**
>
> Calling Chinese writing "pictures for ideas" is a myth. Per Wikipedia's Chinese characters article, characters generally represent *morphemes*, the smallest units of meaning, and as these are nearly always one syllable, written Chinese is often called *morphosyllabic*. Most characters also contain a part that hints at the sound.

Origins

- **Earliest evidence.** The first attested characters are oracle bone inscriptions from the 13th century BCE, found at Anyang in Henan and tied to Shang royal divinations. The article dates the oldest attested writing to Late Shang (c. 1250 to 1050 BCE), with the very earliest examples from the reign of Wu Ding, 1250 to 1200 BCE.
- **Quantity.** Over 150,000 oracle bone fragments have been found to date.
- **Standardisation.** After Qin's conquests led to the Qin dynasty in 221 BCE, the small seal script was standardised for use throughout the country under Chancellor Li Si.

Chapter 1 gave c. 1200 BC for the Yellow River valley; the figures are consistent with the Shang dates here.

How Characters Are Built

The traditional analysis comes from the *Shuowen Jiezi* (c. 100 CE, by Xu Shen) and was the main framework for nearly two thousand years. It has six categories, called *liushu*.

CATEGORY	IDEA
Pictographs	a stylised picture of the thing
Indicatives	a sign pointing to an abstract idea
Compound ideographs	meanings of two parts combined
Phono-semantic compounds	one part for meaning, one part for sound
Loangraphs	a character borrowed for a similar-sounding word
Zhuanzhu	definition unclear; modern scholars generally disregard it

(The category names are the article's; my one-line glosses are my own paraphrase.) The article's proportions for about 3,500 frequently used characters in Standard Chinese are the surprise: phono-semantic compounds make up roughly 58%, pure signs about 18%, semantic-form and phonetic-form compounds together 19%, and pure semantographs about 5%. So the majority of common characters have a sound clue, which is the rebus principle from Chapter 2 working inside a logographic system.

A worked example (mine)

Two standard textbook cases, which I give as my own examples and have not checked against the article: 媽 *mā* "mother" combines 女 (woman) for meaning with 馬 *mǎ* (horse) for sound, and 河 *hé* "river" combines the water element with 可 *kě*. The sound clue is only approximate (here the tone and vowel differ), which is why readers must still learn each character.

```
import unicodedata
for ch in "水河可媽馬女":
    cp = ord(ch)
    print(ch, "U+%04X" % cp, unicodedata.name(ch), 0x4E00 <= cp <= 0x9FFF)
```

```
水 U+6C34 CJK UNIFIED IDEOGRAPH-6C34 True
河 U+6CB3 CJK UNIFIED IDEOGRAPH-6CB3 True
可 U+53EF CJK UNIFIED IDEOGRAPH-53EF True
媽 U+5ABD CJK UNIFIED IDEOGRAPH-5ABD True
馬 U+99AC CJK UNIFIED IDEOGRAPH-99AC True
女 U+5973 CJK UNIFIED IDEOGRAPH-5973 True
```

All six fall in the CJK Unified Ideographs block, U+4E00 to U+9FFF, which the article gives. Notice the names: Unicode names these by code point, not by meaning or reading, because one block serves Chinese, Japanese and Korean.

How Many Characters?

- Writing all frequently used vocabulary requires roughly 3,500 to 5,000 characters, per the article.
- More than 100,000 characters have been identified and included in the Unicode Standard (as of 2025, per the article).

Compare Chapter 1: alphabets use fewer than 100 symbols. A logographic system trades a large inventory for a direct link between symbol and meaning.

Simplified and Traditional

Simplified characters are broadly used in mainland China, Singapore and Malaysia; traditional characters in Taiwan, Hong Kong and Macau. The article, in the part I read, does not say when or how the divergence began.

Characters Beyond Chinese

- **Japanese** uses kanji, the subject of the site's kanji pages. The article calls Japanese "the only major non-Chinese language still written using them". Chapter 4 looks at the kana that Japanese adds.
- **Korean** uses hanja, and **Vietnamese** used chữ Hán, with chữ Nôm for writing Vietnamese itself. Both have largely switched to alphabets (Hangul is the subject of Chapter 8).
- **Zhuang** also had its own script, sawndip.

> **WATCH OUT: A CHARACTER IS NOT A WORD, AND CHINESE WRITING IS NOT "THE LANGUAGE"**
>
> Characters represent morphemes, and many Chinese words are two or more characters long. And because characters stand for meaning, readers who speak different varieties of Chinese can share a written form while speaking differently; that claim is a common one, and I have not checked it in the part of the article I read.

NOTE ON SOURCES

This chapter uses Wikipedia's Chinese characters article (read 100,000 of 240,614 characters; the rest was not read). Dates, percentages and counts are the article's and unchecked against primary scholarship. The article text I read does not cover radical counts or dictionary sizes, so none appear here. The glosses of the six categories, the 媽 and 河 examples and the Python code are mine.

Try It Yourself

EXERCISE 1

Using the article's figures, roughly how many of 3,500 common characters are phono-semantic compounds? Why does that undermine the idea that Chinese is a pure "picture writing"?

View solution

EXERCISE 2

Run the Python example on three characters you find on a Chinese or Japanese web page. Do they fall in the block U+4E00 to U+9FFF? What does the name tell you?

View solution

EXERCISE 3

Why does a logographic system need thousands of characters while an alphabet needs fewer than a hundred? Answer in terms of what one symbol stands for (Chapter 1).

View solution

CHAPTER 3 QUICK REFERENCE

- Earliest Chinese writing: Shang oracle bones, c. 1250 to 1050 BCE
- Characters represent morphemes (usually one syllable): morphosyllabic

- About 58% of 3,500 common characters are phono-semantic compounds (meaning part plus sound part)
- 3,500 to 5,000 characters cover frequent vocabulary; one CJK block (U+4E00 to U+9FFF) serves Chinese, Japanese and Korean

COMING NEXT

Writing Systems of the World 4 turns to syllabaries: Japanese kana, Cherokee and Linear B.

Writing Systems of the World

Course 1 · Chapter 4 · Syllabaries

Chapter 3 ended with Japanese borrowing Chinese characters. Japanese then added something Chinese does not have: a set of symbols for syllables. This chapter looks at the syllabary type through three examples from three continents and three periods: Japanese kana, the Cherokee syllabary and Linear B.

> **THE IDEA IN A BOX**
>
> Wikipedia defines a syllabary as "a set of written symbols that represent the syllables or (more frequently) morae which make up words". Most symbols stand for an optional consonant followed by a vowel (CV). A syllabary works best for a language with simple syllables.

Why Syllabaries Fit Some Languages

- **Simple syllables.** The Syllabary article says that because Japanese uses mainly CV syllables, a syllabary is well suited to write it. Modern syllabary languages tend to have simple phonotactics, with a predominance of single-mora CV syllables.
- **Complex syllables are a problem.** The article says English would need over 10,000 separate glyphs as a pure syllabary, so alphabets handle complex syllables better. Syllabaries often use "dummy vowels" for coda consonants instead.
- **Often from logograms.** The article says syllabaries often begin as simplified logograms, with katakana as its example.
- **Not an abugida.** The article separates syllabaries from abugidas (Chapter 7): in an abugida related syllables share graphical elements in a regular way, while in a syllabary the similarities are not systematic.

Japanese Kana

From Wikipedia's Kana article (first 100,000 of 150,752 characters read).

- **Two sets.** Hiragana and katakana, both developed around 800 CE per the Syllabary article. The Kana article says they are syllabaries used to write Japanese morae.

- **Origins.** Early Japanese was written in kanji. Man'yōgana, kanji used for their sound, represented grammatical particles and morphemes. Hiragana developed from cursive man'yōgana; katakana from abbreviated parts of regular-script man'yōgana, as a glossing system for Buddhist sutras. Kūkai is traditionally said to have invented kana, which the article says is not fully established.
- **Division of labour.** Hiragana is the default syllabary: native words without kanji, particles and okurigana (inflection endings). Katakana covers words of foreign origin, foreign names, onomatopoeia, emphasis, technical terms and some branding. Furigana are small kana beside kanji to show pronunciation.
- **Moraic, not strictly syllabic.** A CVC syllable needs two kana; the only coda kana is ん (n). The Syllabary article says kana are more correctly called a moraic writing system.
- **Voicing marks.** Voiced consonants (g, z, d, b) reuse the unvoiced kana plus a voicing mark (dakuten); p uses the h kana with a small circle (handakuten).
- **Not fully phonetic.** Kana do not mark pitch accent, so homophones like はし ("chopsticks" 箸 and "bridge" 橋) need context, which kanji help.

For learning the kana themselves, the site has the Hiragana and Katakana courses; this course asks only what kind of system they are.

The Cherokee Syllabary

From Wikipedia's Cherokee syllabary article (first 100,000 of 205,413 characters read).

- **Inventor.** Sequoyah, who was illiterate until he created it, began work around 1809 and finished in the early 1820s after about twelve years, dropping or changing most of his original characters.
- **Size.** The modern system has 85 characters. Sequoyah's original had 86; the 86th was dropped in 1828 when Elias Boudinot adapted the chart for printing.
- **Borrowed look, different sounds.** Some characters resemble Latin, Greek, Cyrillic or Glagolitic letters but do not represent the same sounds.
- **Fast uptake.** The article says it became popular almost immediately; by 1825 most Cherokee people could read and write in it, and the *Cherokee Phoenix* adopted it after the 1828 changes.
- **True syllabary.** "Each symbol represents a syllable rather than a single phoneme", though some distinctions (voicing, vowel length, tone) go unmarked.

Linear B

From Wikipedia's Linear B article (first 100,000 of 223,272 characters read).

- **Language.** Mycenaean Greek, the earliest attested form of Greek; c. 1450 to 1200 BC, from palace archives at Knossos, Kydonia, Pylos, Thebes and Mycenae. It disappeared when Mycenaean civilisation fell in the Late Bronze Age collapse.
- **Signs.** Around 87 syllabic signs and over 100 ideographic signs; the ideograms have no phonetic value.
- **Uses.** Mostly confined to administrative contexts at palace sites.
- **Decipherment.** Michael Ventris deciphered it in 1952, building on Alice Kober's work; John Chadwick helped, and the two did the bulk of it between 1951 and 1953. It was adapted from Linear A, an undeciphered script perhaps used for Minoan.
- **A mismatch.** Greek has complex syllables, but Linear B is a syllabary of simple signs. The Syllabary article says complex onsets were written with two signs or simplified, and codas generally ignored: *ko-no-so* for Knōsos.

SCRIPT	LANGUAGE	SIGNS	FIT
Kana	Japanese	No total given in the article's text; two sets	Good: mostly CV
Cherokee	Cherokee	85 (modern)	Dummy vowels for codas; separate sign for /s/
Linear B	Mycenaean Greek	around 87 syllabic, over 100 ideographic	Poor: complex syllables simplified

Kana and Cherokee in Unicode

My own example; I ran it and copied the output. The first lines show a regularity: each katakana sits exactly 0x60 above its hiragana twin. I found this by experiment and have not seen the article state it.

```
import unicodedata
for h in "かきく":
    k = chr(ord(h) + 0x60)
    print(h, k, "U+%04X" % ord(h), "U+%04X" % ord(k), unicodedata.name(h), "|", unicoded
```

```
g = unicodedata.normalize("NFD", "が")
print(len("が"), len(g), [unicodedata.name(c) for c in g])
for cp in [0x13A0, 0x13A1, 0x10000]:
    print(chr(cp), "U+%04X" % cp, unicodedata.name(chr(cp)))
```

```
か カ U+304B U+30AB HIRAGANA LETTER KA | KATAKANA LETTER KA
き キ U+304D U+30AD HIRAGANA LETTER KI | KATAKANA LETTER KI
く ク U+304F U+30AF HIRAGANA LETTER KU | KATAKANA LETTER KU
1 2 ['HIRAGANA LETTER KA', 'COMBINING KATAKANA-HIRAGANA VOICED SOUND MARK']
Ꭰ U+13A0 CHEROKEE LETTER A
Ꭱ U+13A1 CHEROKEE LETTER E
𐀀 U+10000 LINEAR B SYLLABLE B008 A
```

The fourth line shows that が (ga) can be split into か plus a combining voicing mark: the dakuten from the Kana article, visible as a separate code point. The Cherokee block (U+13A0 to U+13FF) and Linear B syllabary block (U+10000 to U+1007F) are the articles' figures.

WATCH OUT: "SYLLABARY" IS NOT "SIMPLE" OR "PRIMITIVE"

A syllabary suits the sound structure of its language. Cherokee literacy spread widely within a few years of the script's completion, according to the article. Also note the Syllabary article carries accuracy-dispute and citation-needed tags, so some of its claims, including those I quote above, are less firm than the others used in this course.

NOTE ON SOURCES

This chapter uses Wikipedia's Syllabary article (whole page as returned), Kana, Cherokee syllabary and Linear B (each read to 100,000 characters; the rest not read). Dates, counts and names are the articles' and unchecked against primary scholarship. The Syllabary article's tags are noted above. The comparison table's "Fit" column wording and the Python example are mine.

Try It Yourself

EXERCISE 1

Why is a syllabary a good fit for Japanese but a poor one for English? Use the article's figure for English and Chapter 1's idea of what one symbol stands for.

View solution

EXERCISE 2

Run the Python example on the hiragana row さしす (sa, shi, su). Does the 0x60 offset still give the katakana twins? Print their names.

View solution

EXERCISE 3

The article writes Knōsos in Linear B as *ko-no-so*. What was lost, and what does that tell you about using a CV syllabary for a language with codas?

View solution

CHAPTER 4 QUICK REFERENCE

- Syllabary: one symbol per syllable or mora, usually CV; best for languages with simple syllables
- Kana: hiragana and katakana, developed around 800 CE from man'yōgana; moraic; voicing marks for g, z, d, b, p
- Cherokee: Sequoyah, c. 1809 to early 1820s; 85 characters in the modern system
- Linear B: Mycenaean Greek, c. 1450 to 1200 BC; deciphered by Ventris in 1952; codas dropped

COMING NEXT

Writing Systems of the World 5 moves to abjads: Phoenician, Hebrew and Arabic, where the vowels are left out.

Writing Systems of the World

Course 1 · Chapter 5 · Abjads

Chapter 1 said most alphabets descend from one source, Proto-Sinaitic. This chapter follows the first branch: the Phoenician script and the two great scripts that grew from it, Hebrew and Arabic. All three write consonants and leave the reader to supply the vowels.

> **THE IDEA IN A BOX**
>
> Wikipedia defines an abjad as a writing system in which "only consonants are represented by letter signs". The vowels are not missing by accident: the language's structure lets the reader fill them in.

The Word "Abjad"

- The term was introduced in 1990 by Peter T. Daniels, per the Abjad article.
- The name comes from the first four letters of the Arabic alphabet in their traditional order, which correspond to *ʾa, b, j, d*.

Why Consonants Are Enough

The article says Semitic words are built from a root of usually three consonants, with vowels marking inflected or derived forms. Its example is the Arabic root K-T-B, which yields *kataba* "he wrote", *yaktubu* "he writes" and *maktabah* "library". The three consonants carry the core meaning, so a reader who knows the language can usually infer the word. The article suggests this helps readers recognise common roots, but marks that claim as lacking citations.

> **WATCH OUT: AN ABJAD IS NOT AN "INCOMPLETE ALPHABET"**
>
> The omission suits the structure of the language. Treating it as a defect is a view from languages where vowels carry the main meaning. Also, the hieroglyphs article from Chapter 2 mentioned that some scholars call Egyptian an abjad, so the type is not limited to Semitic languages, though most examples are.

Phoenician

From Wikipedia's Phoenician alphabet article (first 100,000 of 190,094 characters read).

- **Size and direction.** 22 consonant letters, written right to left. Late varieties sometimes used *matres lectionis* to mark some vowels.
- **Dates.** c. 1050 to 146 BC. The article notes 1050 BC is a conventional start date, since no Phoenician inscriptions are securely dated to the 11th century; the oldest known are from the 10th century.
- **Origin.** It developed directly from the Proto-Sinaitic script of the Late Bronze Age, which was derived from Egyptian hieroglyphs.
- **Spread.** Phoenician maritime merchants carried it across the Mediterranean, and adaptations appeared from the 9th century BC. The article credits its simple design, "one sound was represented by one symbol", as a key reason for its success.
- **Letter names.** Each letter has a name and a meaning: *aleph* means "ox" and *bet* means "house". Some meanings are uncertain, such as *gimel* ("throwing stick" or "camel").

The family tree

DESCENDANT	VIA	COVERED IN
Aramaic, Samaritan	Direct descendants	this chapter
Hebrew, Syriac, Arabic	Derived from Aramaic (Arabic via Nabataean, per the Abjad article)	this chapter
Greek, then Latin (via Cumae and Italic scripts)	Direct from Phoenician	Chapter 6
Cyrillic	From medieval Greek	Chapter 6
Ge'ez	From South Arabian	Chapter 7

Hebrew and Arabic

FEATURE	HEBREW	ARABIC
Letters	22 consonants plus 5 final forms; 4 matres lectionis (Abjad article)	28; 3 matres lectionis (Abjad article)
Direction	Right to left	Right to left, cursive; letters mostly connect
Letter shapes	Fixed, with final forms	Up to four shapes: initial, medial, final, isolated
Vowel marks	Niqqud, optional	Harakat, optional
Origin	Via Aramaic	Nabataean; earliest known text a late fourth-century inscription at Jabal Ram, Jordan

- **"Impure" abjads.** The Abjad article calls Arabic, Hebrew, Aramaic and Pahlavi impure because some consonant letters double as vowel letters (*matres lectionis*). This began rarely and became more common over time.
- **When vowels are written.** Vowel marks are optional and "not the dominant (or literate) form". The Arabic article says everyday writing, publications and street signs usually omit them, while the Qur'ān, children's books and grammar texts usually include them.
- **Other languages.** The Arabic script is also used for Persian and Urdu, with extra letters such as پ added in adaptations.

Seeing It in Python

My own example; I ran it and copied the output. It writes the root K-T-B bare and then with a short vowel mark after each consonant.

```
import unicodedata
bare = "كتب"
voweled = "كَتَبَ"
print(len(bare), len(voweled))
print([unicodedata.name(c) for c in voweled])
print([unicodedata.bidirectional(c) for c in "אבג"], unicodedata.bidirectional("a"))
print(len("שלום"), len("שָׁלוֹם"))
print(chr(0x10900), unicodedata.name(chr(0x10900)))
```

```
3 6
['ARABIC LETTER KAF', 'ARABIC FATHA', 'ARABIC LETTER TEH', 'ARABIC FATHA', 'ARABIC LETTE
['R', 'R', 'R'] L
4 7
𐤀 PHOENICIAN LETTER ALF
```

The bare root is 3 code points; with a short-vowel mark after each consonant it is 6. In Hebrew, "shalom" is 4 code points bare and 7 with niqqud. The vowel marks are separate code points, which fits their being optional additions. The `bidirectional` property `R` marks the Hebrew letters as right-to-left, against `L` for the Latin letter. Phoenician has its own Unicode block (the article gives U+10900 to U+1091F), and its first letter is named ALF.

NOTE ON SOURCES

This chapter uses Wikipedia's Abjad (read 100,000 of 131,879 characters), Phoenician alphabet (100,000 of 190,094) and Arabic alphabet (100,000 of 231,051) articles; the rest was not read. Dates and counts are the articles' and unchecked against primary scholarship; two Abjad claims are marked there as needing citations, and I say so where they appear. The Abjad article gives no contextual-form or Unicode detail, so the Arabic shapes row and the Arabic Unicode range come from the Arabic article. The Python example and the "family tree" table layout are mine.

Try It Yourself

EXERCISE 1

Using the root K-T-B, explain in two sentences how an abjad can work without writing vowels, and why that would be a worse fit for English.

View solution

EXERCISE 2

Run the Python example on a short Arabic or Hebrew word of your own, with and without vowel marks (copy one from a dictionary site). How many code points does each version have?

View solution

EXERCISE 3

Trace Hebrew and Arabic back to their source using the family tree table. Which script do they share as an ancestor, and which earlier script did that one come from?

View solution

CHAPTER 5 QUICK REFERENCE

- Abjad: only consonants written; term coined by Peter T. Daniels in 1990, from Arabic a-b-j-d
- Semitic roots of three consonants (K-T-B) make this workable
- Phoenician: 22 letters, right to left, from Proto-Sinaitic; ancestor of Aramaic, Greek and others
- Hebrew and Arabic: "impure" abjads with some vowel letters and optional vowel marks

COMING NEXT

Writing Systems of the World 6 turns to alphabets proper: how the Greeks added vowels, and how Latin and Cyrillic spread.

Writing Systems of the World

Course 1 · Chapter 6 · Alphabets

Chapter 5 ended with Phoenician giving rise to both the Aramaic line (Hebrew, Arabic) and the Greek line. This chapter follows the Greek line, the one that gave the world the alphabet in the narrow sense of the word, and then the two scripts that grew from it: Latin and Cyrillic.

> **THE IDEA IN A BOX**
>
> Wikipedia's Alphabet article says "the Greek alphabet was the first in which vowels had independent letterforms separate from those of consonants." The broad sense of "alphabet" includes Phoenician; in the narrow sense, where letters represent both consonants and vowels, the article says the Greek alphabet was the first true alphabet.

Greek: Adding the Vowels

From Wikipedia's Greek alphabet article (first 100,000 of 226,322 characters read).

- **Date.** Commonly dated to the late 9th or early 8th century BC, conventionally around 800 BC. By the end of the 4th century BC the 24-letter Ionic-based alphabet had become standard.
- **Method.** The Greeks took the Phoenician alphabet, which had only consonants, and reassigned letters for sounds Greek needed as vowels.
- **Accents.** Polytonic orthography (Ancient Greek) has three accent marks and breathing marks; monotonic orthography, adopted officially in 1982 for Modern Greek, uses a single acute accent.

PHOENICIAN LETTER	BECAME GREEK	VOWEL
aleph	A alpha	/a/
he	E epsilon	/e/
yod	I iota	/i/
ayin	O omicron	/o/
waw	Y upsilon	/u/

This is the whole trick. Aleph, he, yod, ayin and waw stood for consonants in Phoenician (including sounds Greek lacked), and the Greeks used the spare signs to write vowels. Unlike an abjad, nothing is left for the reader to infer.

Latin

From Wikipedia's Latin script article (first 100,000 of 149,903 characters read).

- **Origin.** A form of the Greek alphabet used in Cumae, a Greek city in Magna Graecia; the Etruscans altered it and the Romans altered it again. The infobox gives c. 700 BC to the present.
- **Later letters.** W originated as a doubled V and came into common use in the later 11th century; J was introduced into English for the consonant in the 17th century; a rounded capital U was derived for the vowel in the 16th century. The ISO basic Latin alphabet has 26 letters.
- **Reach.** The article calls it the most widely adopted writing system in the world and says about 3 billion people (almost 40% of the world population) use it.

Cyrillic

From Wikipedia's Cyrillic script article (first 100,000 of 245,613 characters read) and the Alphabet article.

- **Origin.** Developed in the 9th century at the Preslav Literary School in the First Bulgarian Empire, with earliest variants dated c. 893 to c. 940. It is named after Saint Cyril, but the article says the name "denotes homage rather than authorship": his and Methodius's followers developed it.
- **Basis.** Adapted from Greek uncial letters, with ligatures and consonants from Glagolitic added for sounds Greek lacked. The Alphabet article says Glagolitic is believed to have been created by Cyril and Methodius.
- **Use.** About 250 million people in Eurasia use Cyrillic as the official script of their national languages (as of 2019), about half in Russia. It is also used for non-Slavic languages such as Chechen, Tatar and Mongolian (in Mongolia). Kazakhstan has moved from Cyrillic to Latin.

Same shape, different letter

```
import unicodedata
for ch in ["A", "A", "A"]:
    print(ch, "U+%04X" % ord(ch), unicodedata.name(ch))
print("A" == "A", "A" == "A")
```

```
A U+0041 LATIN CAPITAL LETTER A
A U+0391 GREEK CAPITAL LETTER ALPHA
A U+0410 CYRILLIC CAPITAL LETTER A
False False
```

The three look the same but are three different code points in three scripts, which the common ancestry explains. My own example, run and copied. A program comparing them treats them as different letters, which is a known source of confusion in text processing.

Adapting an Alphabet to a New Language

When an alphabet is borrowed for a language with sounds it lacks letters for, the Alphabet article lists three devices:

DEVICE	IDEA	EXAMPLES
Digraph	two letters, one sound	Hungarian cs, sz, zs (Alphabet article); English ch, sh, th, ng (Latin script article)
Diacritic	a mark added to a letter	German ä, ö, ü; Romanian ă, â, î, ș, ț (Latin script article); Vietnamese tones
Ligature	two letters joined as one	æ in Danish and Icelandic (Alphabet article)

```
print(unicodedata.name("æ"), len("cs"), len("ő"))
```

```
LATIN SMALL LETTER AE 2 1
```

The digraph "cs" is two code points for one sound, while the letter ő (Hungarian, which the Hungarian courses on this site use) is a single code point. Same job, different device.

WATCH OUT: LETTER COUNT IS NOT A PROPERTY OF "THE ALPHABET"

The ISO basic Latin alphabet has 26 letters, but each language using Latin adds or drops letters, digraphs and diacritics. The Cyrillic article I read did not state a letter count for Russian, so none is given here.

NOTE ON SOURCES

This chapter uses Wikipedia's Alphabet, Greek alphabet, Latin script and Cyrillic script articles (each read to 100,000 characters; the rest not read). Dates and counts are the articles' and unchecked against primary scholarship. The Alphabet article does not name the Phoenician letters that became vowels; that table comes from the Greek alphabet article. The Greek article's direction-change claim (an earlier boustrophedon stage) was not in the text I read, so it is left out. The Python examples are mine.

Try It Yourself

EXERCISE 1

Explain in your own words how the Greeks made a consonant-only script into one that writes vowels. Use the five letters from the table.

View solution

EXERCISE 2

Find two more capital letters that look the same in Latin and Cyrillic (look at the Cyrillic keyboard layout or any Russian text). Use Python to print their names and code points and test them for equality.

View solution

EXERCISE 3

Name the device (digraph, diacritic or ligature) in each: Hungarian "sz", German "ü", Danish "æ". Why might a language prefer one device to another?

View solution

CHAPTER 6 QUICK REFERENCE

- Greek (c. 800 BC): Phoenician letters aleph, he, yod, ayin, waw reassigned as vowels /a e i o u/
- Latin: from the Greek of Cumae through Etruscan, c. 700 BC onward; ISO basic alphabet has 26 letters
- Cyrillic: Preslav Literary School, 9th century; Greek uncial plus Glagolitic additions
- Adapting an alphabet: digraphs, diacritics, ligatures

COMING NEXT

Writing Systems of the World 7 looks at abugidas: Devanagari, Ethiopic and Thai, where each consonant carries a built-in vowel.

Writing Systems of the World

Course 1 · Chapter 7 · Abugidas

Chapter 4 said a syllabary gives every syllable its own unrelated symbol. Chapter 6 said an alphabet gives vowels equal status with consonants. The abugida sits between them: each symbol is a consonant with a vowel already attached, and the other vowels are marked by changing it. This chapter covers the largest family, the Brahmic scripts of South and Southeast Asia, and Ethiopia's Ge'ez script, which reached the same idea separately.

> **THE IDEA IN A BOX**
>
> Wikipedia defines an abugida as "a type of writing system whose basic characters denote consonants followed by a particular vowel". The default vowel is the *inherent vowel*. To write any other vowel, a mark is added to the consonant.

The Word and the Idea

- **The term.** From Ge'ez, where it names the Ethiopic letters; it derives from the letters "ä, bu, gi, and da". Peter T. Daniels proposed it as a linguistic label in 1990, the same year as "abjad" (Chapter 5).
- **Against a syllabary.** In a syllabary, a single symbol denotes the consonant-vowel pair, and letters sharing a consonant or vowel usually look unrelated. In an abugida the related syllables share a visible element.
- **Against an alphabet.** In an alphabet vowels have status equal to consonants; in an abugida vowels are secondary marks on the consonant.
- **Silencing the vowel.** Many abugidas have a diacritic that removes the inherent vowel to leave a bare consonant; in Sanskrit it is called the *virāma* or *halantam*.
- **Clusters.** Consonant clusters may be written by merging letters (conjuncts); some scripts stack the consonants vertically and others use special half-forms.

The Brahmic Family

The Abugida article says Indic scripts originated in India and spread to Southeast Asia, and reached Bangladesh, Sri Lanka, Nepal, Tibet, Mongolia and Russia. It says all surviving Indic scripts descend from the Brahmi alphabet. The article's examples include Devanagari (Hindi, Sanskrit, Marathi, Nepali), Bengali, Tamil, Telugu, Kannada, Malayalam, Thai, Lao, Burmese, Khmer, Tibetan, Javanese, Balinese and Baybayin.

Devanagari

From Wikipedia's Devanagari article (first 100,000 of 151,714 characters read).

- **Use.** Hindi, Sanskrit, Marathi and Nepali among over 120 languages; the article calls it the fourth most widely adopted writing system in the world.
- **Origin.** From the 3rd-century BCE Brāhmī script, through Nāgarī.
- **Inventory.** 48 primary characters: 14 vowels and 34 consonants. Written left to right, with a horizontal line (*shirorekha*) running along the top of full letters.
- **Vowels.** The consonant क carries the inherent /a/. Other vowels are dependent signs (matras) attached to it: का for ā, कि for i. A virama, as in क्, removes the vowel.

Ge'ez (Ethiopic)

From Wikipedia's Ge'ez script article (first 100,000 of 163,469 characters read) and the Abugida article.

- **Not Brahmic.** The Abugida article says it derives from a different abjad, the Sabean script of Yemen; the Ge'ez article says from Ancient South Arabian, though its origins are debated.
- **History.** First an abjad (the Ge'ez article dates this to about the 1st century CE; the Abugida article says until about AD 350), then vowel marking was added in the 4th century.
- **How vowels are shown.** By modifying the shape of the consonant itself, in seven basic vowel orders (ä, u, i, a, e, ə, o), not by separate diacritics. The article says the vowel marks were fused with the consonants, "so the system is laid out as a syllabary".
- **Use and size.** Ge'ez, Amharic, Tigrinya and other languages of Ethiopia and Eritrea; 26 consonantal letters. Written left to right today, right to left as an abjad.
- **Naming.** Unicode and ISO 15924 call it "Ethiopic". Amharic and Tigrinya speakers often call it *fidäl* (ፊደል).

Thai

From Wikipedia's Thai script article (first 100,000 of 219,119 characters read).

- **Origin.** From Old Khmer script, rooted in the South Indian Pallava alphabet and a southern branch of the Brahmi tradition. The infobox gives 1283 to present, with Ramkhamhaeng as creator; the article dates the Ram Khamhaeng Inscription to 1292 and says many scholars question its authenticity.
- **Letters.** 44 consonant symbols representing 21 distinct consonant sounds.
- **Vowels.** Vowel signs can go above, below, left of or right of the consonant, or in combinations of these places.
- **Tones.** Four tone diacritics; tone depends on the consonant's class, vowel length, the closing consonant and the mark, if present.
- **No spaces.** Words are not separated by spaces, except in certain linguistically motivated cases.

FEATURE	DEVANAGARI	GE'EZ	THAI
Family	Brahmic	South Arabian lineage	Brahmic via Old Khmer
Vowels shown by	Signs attached around the consonant	Altering the consonant's shape	Signs above, below, left or right
Consonant symbols	34 (of 48 primary characters)	26 consonantal letters	44 (for 21 sounds)

> **WATCH OUT: WHERE THE LINES BLUR**
>
> The Ge'ez article itself says its system is "laid out as a syllabary", yet the Abugida article classes it as an abugida because its signs share a consonant shape. The line between syllabary and abugida is a matter of how regular the vowel changes are, so scholars can draw it differently.

Seeing the Marks in Python

My own example; I ran it and copied the output. It builds the Devanagari consonant क (ka) and adds the vowel signs from the article.

```
import unicodedata
ka = "क"
for s in [ka, ka + "ा", ka + "ि", ka + "ू"]:
    print(s, len(s), [unicodedata.name(c) for c in s])
print("क्ष", len("क्ष"), [unicodedata.name(c) for c in "क्ष"])
```

```
for cp in range(0x1200, 0x1207):
    print(chr(cp), unicodedata.name(chr(cp)))
t = "เก"
print(t, [unicodedata.name(c) for c in t])
```

```
क 1 ['DEVANAGARI LETTER KA']
का 2 ['DEVANAGARI LETTER KA', 'DEVANAGARI VOWEL SIGN AA']
कि 2 ['DEVANAGARI LETTER KA', 'DEVANAGARI VOWEL SIGN I']
क् 2 ['DEVANAGARI LETTER KA', 'DEVANAGARI SIGN VIRAMA']
क्ष 3 ['DEVANAGARI LETTER KA', 'DEVANAGARI SIGN VIRAMA', 'DEVANAGARI LETTER SSA']
U ETHIOPIC SYLLABLE HA
Uዉ ETHIOPIC SYLLABLE HU
Uዊ ETHIOPIC SYLLABLE HI
Uዋ ETHIOPIC SYLLABLE HAA
Uዌ ETHIOPIC SYLLABLE HEE
Uው ETHIOPIC SYLLABLE HE
Uዎ ETHIOPIC SYLLABLE HO
เก ['THAI CHARACTER SARA E', 'THAI CHARACTER KO KAI']
```

Three observations, all mine from the output. First, in Devanagari the vowel sign and the virama are separate code points stored *after* the consonant, and the conjunct क्ष is stored as ka + virama + ssa, so the merged shape is something the font draws. (That also explains why the Hindi word in Chapter 1 counted as four code points, not as the number of shapes a reader sees.) Second, Unicode stores Ge'ez as whole syllables, one code point per consonant-vowel pair, treating it like a syllabary even though its shapes are related; seven orders of the same consonant (HA, HU, HI, HAA, HEE, HE, HO) appear in a row. Third, in Thai the vowel sign SARA E is stored *before* the consonant, matching its position on the left in writing.

NOTE ON SOURCES

This chapter uses Wikipedia's Abugida, Devanagari, Ge'ez script and Thai script articles (each read to 100,000 characters; the rest not read). Dates and counts are the articles' and unchecked against primary scholarship; the two articles disagree on when Ge'ez was an abjad (about the 1st century CE versus until about AD 350), and both dates are given. The Devanagari article gave no user count, so none appears. The comparison table's layout and the Python example and its observations are mine.

Try It Yourself

EXERCISE 1

In your own words, what is the difference between a syllabary, an abugida and an alphabet? Use the Abugida article's three-way contrast and give one script as an example of each.

View solution

EXERCISE 2

In Python, print the code points (by name) of the Devanagari syllables कु, कू and के. What is the same in each, and what changes?

View solution

EXERCISE 3

Thai has 44 consonant symbols for 21 consonant sounds. Suggest, from this chapter, one reason a script might keep more letters than sounds. Then say what the Ge'ez and Devanagari scripts do differently to show a vowel.

View solution

CHAPTER 7 QUICK REFERENCE

- Abugida: consonant plus inherent vowel; other vowels marked on the consonant; term from Ge'ez, coined for linguistics by Daniels in 1990
- Virama removes the inherent vowel; conjuncts merge consonants in clusters
- Devanagari: from Brahmi via Nagari; 14 vowels and 34 consonants; Ge'ez: South Arabian lineage, vowels by changing consonant shape; Thai: from Old Khmer, 44 consonants, vowel signs on all four sides
- Unicode stores the Brahmic vowel signs as separate code points, Ge'ez as whole syllables

COMING NEXT

Writing Systems of the World 8 looks at featural and invented scripts: Hangul and other scripts designed on purpose.

Writing Systems of the World

Course 1 · Chapter 8 · Invented Scripts

Most scripts grew slowly through borrowing and change. A few were designed by a person on a particular occasion. This chapter looks at four of them: the Korean Hangul, the Canadian syllabics, and two scripts built for fiction, Tengwar and Klingon pIqaD. Cherokee (Chapter 4) is the other famous example.

> **THE IDEA IN A BOX**
>
> Wikipedia's Writing system article says a *featural* system "uses symbols representing sub-phonetic elements", with Hangul as the main example. If a designer builds the shapes of letters to show how sounds are made, the script can encode phonetic theory as well as sound.

Hangul

From Wikipedia's Hangul article (first 100,000 of 212,572 characters read).

- **Creation.** King Sejong the Great (r. 1418 to 1450) was responsible for its creation. It was announced around late 1443 to early 1444 and officially published in 1446 in the *Hunminjeongeum*, which was also the script's original name.
- **Inventory.** Mainstream orthographies use 24 basic letters (jamo): 14 consonants and 10 vowels. Modifications and combinations bring the total to 51.
- **Featural design.** According to the *Hunminjeongeum Haerye*, the shapes of five basic consonants are based on human speech organs, and the other consonants are derived by adding strokes for "harsher" sounds. Vowels are linked to Chinese philosophy: ㆍ represents yin and earth, ㅡ yang and the sun, and ㅣ man (neutral).
- **The dispute.** The "featural script" label was coined by Sampson in 1985, questioned by DeFrancis in 1989 and rebutted by Chin-Woo Kim in 1997. Pae, Winskel and Kim argue the designation is difficult to falsify and uncertain.
- **Syllable blocks.** Letters are grouped into blocks of an initial consonant, a vowel and an optional final consonant, with a silent ㅇ when no initial sound is needed. Vowels with a long vertical stroke go to the right of the initial; those with a long horizontal stroke go below it; a final consonant goes at the bottom.

- **Count.** 19 initial consonants, 21 vowels and 28 final consonants (including none) yield 11,172 syllables, though scholars estimate only about 2,000 see regular use.
- **Literacy.** The article says hanja's difficulty contributed to high illiteracy, and that within a decade of the post-1945 push, a significant majority of Koreans in both Koreas became literate in at least Hangul.

So Hangul is a script designed to look like an alphabet and write like a syllabary: letters are phonemic, but they are stacked into syllable blocks. For learning it, the site has the Korean Alphabet 1 course.

Canadian Aboriginal Syllabics

From Wikipedia's Canadian Aboriginal syllabics article (first 100,000 of 198,188 characters read).

- **Creator.** James Evans, a missionary, developed the script by 1840, probably with Indigenous language experts, working with Cree and Ojibwe speakers. Cree oral traditions give a different origin, crediting a Cree man named Calling Badger.
- **Principle.** Each consonant has a fixed base shape, and the vowel is shown by rotating that shape. The consonant p is a chevron that points up, down, left or right to mark different vowels.
- **Classification.** Despite its name, the article says it is an abugida rather than a true syllabary.
- **Use.** All Cree languages and Naskapi, regionally Ojibwe and Blackfoot, and Inuktitut, which is co-official with Latin script in Nunavut.
- **Reception.** Resisted by colonial and missionary authorities (the Hudson's Bay Company declined to import a printing press for Evans), but the Cree community quickly took to it; a Bible was published in Cree syllabics by 1861, and the article says Cree literacy rates were possibly among the highest in the world by the late 19th century.

Tengwar

From Wikipedia's Tengwar article (first 100,000 of 119,846 characters read).

- **Creator.** J. R. R. Tolkien; probably developed in the late 1920s or early 1930s, with the first published sample in 1937. In his fiction the elf Fëanor invented it.

- **Featural design.** "The shapes of the letters correspond to the distinctive features of the sounds they represent." Most letters combine a stem with one or two bows: doubling the bow voices a stop, raising the stem makes a fricative, and shortening it makes a nasal.
- **Grid.** Four series (*témar*) by place of articulation: dental, labial, velar and labialized velar; six rows of manners of articulation.
- **Modes.** The orthographies are called modes. Vowel-diacritic modes can be treated as abugidas, while the "full writing" mode writes vowels as letters; other modes exist for English, Spanish, German and more.
- **Encoding.** Not officially encoded in Unicode; Michael Everson proposed it in 1997; the unofficial ConScript Unicode Registry maps it to U+E000 to U+E07F in the Private Use Area.

Klingon pIqaD

From Wikipedia's Klingon scripts article (whole answer as returned).

- **Origin.** The Astra Image Corporation designed the letters for *Star Trek: The Motion Picture*, often wrongly credited to Michael Okuda. They were based on Klingon battlecruiser hull markings by Matt Jeffries and on Tibetan writing, with sharp forms evoking bladed weapons.
- **Size.** 26 letters with a one-to-one correspondence between letters and sounds, plus ten digits.
- **In practice.** The vast majority of Klingonists prefer Latin-based romanisation. The script occupies U+F8D0 to U+F8FF in the Private Use Area; Everson's 1997 proposal to encode it was rejected by the Unicode Technical Committee in May 2001, reasoning that the script saw almost no communicative use.

The site's Klingon courses use the romanisation, which fits the article's description of common practice.

SCRIPT	DESIGNER	DATE	TYPE	DESIGNED TO
Hangul	King Sejong	published 1446	Alphabet in syllable blocks; featural (disputed)	Let ordinary Koreans read and write
Canadian syllabics	James Evans	by 1840	Abugida	Write Cree and Ojibwe

SCRIPT	DESIGNER	DATE	TYPE	DESIGNED TO
Tengwar	J. R. R. Tolkien	late 1920s or early 1930s	Featural; modes vary (some abugida-like)	Fiction (and later fans)
pIqaD	Astra Image Corporation	for the 1979 film (date from my own memory, not the article)	Alphabet-like, one letter per sound	Film props

The "Designed to" column is my own summary.

Seeing the Design in Python

My own example; I ran it and copied the output. It splits a Hangul syllable into its letters and reverses the arithmetic by which Unicode lays out all the syllables, then looks at Canadian syllabics and the Private Use Area.

```
import unicodedata
s = "한"
d = unicodedata.normalize("NFD", s)
print(s, len(s), len(d), [unicodedata.name(c) for c in d])
n = ord(s) - 0xAC00
print(n // (21 * 28), (n % (21 * 28)) // 28, n % 28)
print(0xD7A3 - 0xAC00 + 1, 19 * 21 * 28)
for cp in [0x1431, 0x1432, 0x1433, 0x1434]:
    print(chr(cp), unicodedata.name(chr(cp)))
print(unicodedata.category(chr(0xE000)), unicodedata.category(chr(0xF8D0)))
```

```
한 1 3 ['HANGUL CHOSEONG HIEUH', 'HANGUL JUNGSEONG A', 'HANGUL JONGSEONG NIEUN']
18 0 4
11172 11172
ᐱ CANADIAN SYLLABICS PI
ᐠ CANADIAN SYLLABICS PII
ᐳ CANADIAN SYLLABICS PO
ᐵ CANADIAN SYLLABICS POO
Co Co
```

What the output shows (my reading): 한 is one precomposed code point that splits into three letters (initial ㅎ, vowel ㅏ, final ㄴ), and the block U+AC00 to U+D7A3 holds exactly 19 x 21 x 28 = 11,172 syllables, so the position of a syllable in Unicode is computed from its letters, which

confirms the article's syllable count. The syllabics example shows the PI/PII/PO/POO series sitting side by side, consistent with one base shape in different orientations. The category `Co` means "private use": Tengwar and pIqaD have no standard characters, so a font must supply them.

WATCH OUT: DESIGNED DOES NOT MEAN ARBITRARY

A designed script can borrow from existing ones (pIqaD from Tibetan writing, per the article) and works only if a community adopts it. Cree took to syllabics quickly; pIqaD saw almost no communicative use, according to the Unicode committee's reasoning in the article. And the featural claim for Hangul is debated by scholars, so treat sweeping praise for it with care.

NOTE ON SOURCES

This chapter uses Wikipedia's Hangul, Canadian Aboriginal syllabics, Tengwar and Klingon scripts articles (the first three read to 100,000 characters, the Klingon scripts one as returned). Dates and counts are the articles' and unchecked against primary scholarship. The Klingon article did not date the film or link the design to a later film, so the table notes the 1979 date as my own memory; the Hangul article's text I read did not say who coined the name "Hangul", so I give none. The comparison table's last column, the Python and its reading are mine.

Try It Yourself

EXERCISE 1

Use the article's figures to explain why only about 2,000 of Hangul's 11,172 possible syllables are in regular use, and why 11,172 is still an easy number for a computer to cover.

View solution

EXERCISE 2

In Python, decompose the syllable 글 (as in 한글) into its letters with NFD normalisation. Name each letter and say which position (initial, vowel, final) it occupies.

View solution

EXERCISE 3

Compare Canadian syllabics and Tengwar: how does each use the shape of a symbol to show something systematic? Which is closer to the "featural" idea from Chapter 1, and why?

View solution

CHAPTER 8 QUICK REFERENCE

- Hangul: King Sejong, published 1446; 24 basic letters (14 consonants, 10 vowels); syllable blocks; 11,172 possible syllables; featural status disputed
- Canadian syllabics: James Evans, by 1840; vowels shown by rotating the consonant; classed as an abugida
- Tengwar: Tolkien; letter shapes correspond to sound features; unofficial Private Use Area encoding
- pIqaD: designed for the film; 26 letters; most Klingonists use Latin romanisation

COMING NEXT

Writing Systems of the World 9 turns from the scripts to their rules: direction, spelling and orthography.

Writing Systems of the World

Course 1 · Chapter 9 · Spelling Rules

Chapters 2 to 8 were about the symbols. A script alone does not tell you how to write a language: you also need the conventions for which letters spell which sounds, which way the lines run, and what to do when a language has to be written in a script that is not its own. This chapter covers direction, orthography, English spelling and romanisation.

THE IDEA IN A BOX

Wikipedia's Orthography article describes an orthography as "a set of conventions for writing a language", covering spelling, punctuation, word boundaries, capitalisation, hyphenation and emphasis. The same script can carry very different orthographies.

Which Way Do the Lines Run?

From Wikipedia's Writing system article (first 100,000 of 154,316 characters read).

DIRECTION	EXAMPLES (PER THE ARTICLE)
Left to right, top to bottom	The Greek alphabet and its successors settled on this pattern
Right to left	Arabic; Hebrew came to be written this way; the Phoenician direction "initially stabilized after c. 800 BC"
Vertical, top to bottom, columns left to right	Mongolian script
Vertical, columns right to left	Scripts using Chinese characters, traditionally
Bottom to top	Ogham (on the corner of a stone); Hanunoo lines move upward but are read left to right
Either way / alternating	Egyptian hieroglyphs (figures turned to face the start of the line); the early alphabet, most commonly written in alternating-direction (boustrophedon) lines

- **Why left to right?** The article gives one reason for the Greek pattern: the hand does not interfere with what is being written, because most people are right-handed.
- **It can change.** The article says horizontal left-to-right writing was widely adopted for Chinese-derived scripts only in the 20th century, due to Western influence.

Direction in Unicode

A computer storing mixed text keeps characters in *reading* order and gives each character a direction class. My own example, run and copied; the second part counts the ough-words in a sentence the English orthography article quotes.

```
import unicodedata
for ch in ["a", "ع" ,"א", "語", "1", " "]:
    print(repr(ch), unicodedata.bidirectional(ch))
s = "Though the tough cough and hiccough plough him through"
words = [w for w in s.lower().split() if "ough" in w]
print(words)
print(len(words))
```

```
'a' L
'א' R
'ع' AL
'語' L
'1' EN
' ' WS
['though', 'tough', 'cough', 'hiccough', 'plough', 'through']
6
```

L is left-to-right, R right-to-left (Hebrew), AL is Arabic letter (also right to left), EN is a European number and WS is white space. The Chinese character is classed L, which matches the modern left-to-right practice in the article. Digits stay left-to-right inside right-to-left text; that is my observation from the class, not a claim from the article.

Orthography: Shallow and Deep

From Wikipedia's Orthography article (whole answer as returned).

- **Shallow.** Spelling-to-sound correspondences are relatively simple and consistent. Spanish has a clear one-to-one correspondence between phonemes and graphemes, and Hawaiian

uses only 13 letters.

- **Deep.** Correspondences are highly complex or inconsistent; the article's example is Navajo.
- **Effect.** Children learn to read faster in shallow orthographies like Spanish than in deep ones like English.
- **Regulators.** Some nations use language academies, such as the Académie Française and the Royal Spanish Academy. English has none.
- **Reform.** Noah Webster sought easily noticeable differences between American and British spelling (honor and honour). Modern Greek simplified its historical diacritics to a single accent.

Why English Spelling Is So Irregular

From Wikipedia's English orthography article (first 100,000 of 205,599 characters read).

CAUSE	WHAT THE ARTICLE SAYS
Sound change	Spellings were largely derived from Middle English and generally do not reflect sound changes since the late 15th century, such as the Great Vowel Shift
Printing	Early compositors, who were multilingual, sometimes used patterns from other languages; "ghost" was "gost" until a Flemish pattern was unintentionally substituted
Loanwords	Foreign spellings are usually kept: Polish cz in *Czech*, Norwegian fj in *fjord*. Early Middle English respelled French imports, but this stopped after about 1400
Etymological respelling	A "b" was added to *debt* and an "s" to *island* to suggest Latin or Greek origins
No regulator	"English has never had any official regulating authority for spelling"

- **No one-to-one match.** English has 24 to 27 consonant phonemes and 13 to 20 vowels but only 26 letters.
- **One sound, many spellings.** The sound /uː/ can be spelled in up to 24 different ways, and /iː/ in at least 18 to 21. For /ʃ/ the article lists sh, ch (French loans), s (*sugar*), t (*ration*), c (*ocean*), ss (*tissue*), sch (*schedule*) and x (Chinese loans).
- **One spelling, many sounds.** The letters ou can be pronounced in as many as nine different ways. The four-letter ough can be pronounced in at least ten different ways; the article cites

Robert A. Heinlein's sentence "Though the tough cough and hiccough plough him through" to show it (the Python example above found six ough-words in it).

- **Silent letters.** Examples listed: l in *talk* and *half*, w in *two* and *sword*, gh in *night* and *brought*, and the silent e.
- **Webster.** He promoted more phonetic spellings in the United States, such as *flavor*, *fiber*, *defense*, *analyze* and *catalog*; the article says these forms already existed, but his dictionaries helped standardise them.

> **WATCH OUT: IRREGULAR IS NOT THE SAME AS RANDOM**
>
> Most oddities have a history, as the table shows, and the article says the spelling "generally do[es] not reflect" sound change, not that it ignores sound. Also, the article calls English spelling "considered irregular and complex compared to that of other languages", which is a comparison, not a claim that no pattern exists.

Romanisation

From Wikipedia's Romanization article (first 100,000 of 107,175 characters read). When a language needs writing in Latin letters for an outsider, the article distinguishes two aims:

AIM	PRINCIPLE	EXAMPLES (PER THE ARTICLE)
Transliteration	A one-to-one mapping of characters in the source into the target script; less weight on how the result sounds	Nihon-shiki for Japanese (1885); ISO 9 for Cyrillic (1995); BGN/PCGN (Russian version 1947)
Transcription	Represents speech; phonemic or phonetic	Hepburn for Japanese (1867), designed for English speakers; Hanyu Pinyin (1958); Revised Romanization of Korean (2000)

- Pinyin has been used officially in China, adopted by much of the international community, and in Taiwan since 1 January 2009. The Revised Romanization of Korean replaced a modified McCune-Reischauer system used from 1984 to 2000.
- For computers, ease of creation and storage matters: Harvard-Kyoto aims at 7-bit ASCII, ITRANS was once prevalent on Usenet, and wāpuro rōmaji is "not strictly a system, but a collection of common practices" for typing Japanese.

NOTE ON SOURCES

This chapter uses Wikipedia's Writing system (directionality section, first 100,000 characters), Orthography, English orthography (first 100,000 characters) and Romanization (first 100,000 characters) articles. The English orthography article did not mention color/colour, so I use honor and honour from the Orthography article instead; the Orthography article did not discuss the Great Vowel Shift, printing, Finnish or Italian, and the Romanization article did not mention the 1928 Turkish reform, so none appear. Counts and dates are the articles' and unchecked against primary scholarship. The table layouts, the Python example and the digit and Chinese-character observations are mine.

Try It Yourself

EXERCISE 1

Sort these into shallow or deep, using the Orthography article: Spanish, English, Hawaiian, Navajo. Then say what the article reports about learning to read in each kind.

View solution

EXERCISE 2

Use the table of causes to explain why *debt* has a silent b and *ghost* has an h. Which cause covers each?

View solution

EXERCISE 3

Run the Python direction check on three characters of your own (one each from different scripts). Which direction class does each report, and does it match the article's list of directions?

View solution

CHAPTER 9 QUICK REFERENCE

- Orthography: the conventions for writing a language (spelling, punctuation, word boundaries and more)

- Shallow orthography (Spanish): simple sound-letter matches; deep (English, Navajo): complex or inconsistent
- English irregularity: sound change, printing, loanwords, etymological respelling, no regulator
- Romanisation: transliteration (character to character) versus transcription (sound to letters)

COMING NEXT

Writing Systems of the World 10 is the last chapter: how lost scripts were deciphered, which scripts are in danger, and a recap of the whole course.

Writing Systems of the World

Course 1 · Chapter 10 · Lost and Found

The last chapter looks at what happens when a script stops being read, how some were recovered, why others still resist, and what is happening to small scripts today. It ends with a recap of the whole course.

> **THE IDEA IN A BOX**
>
> Wikipedia defines decipherment as "the discovery of the meaning of the symbols found in extinct languages and/or alphabets", and says it should not be confused with cryptanalysis. A decipherer is not breaking a code someone hid; they are recovering a convention that nobody alive remembers.

What Decipherment Needs

From Wikipedia's Decipherment article (whole answer as returned).

- **Multilingual inscriptions.** The Rosetta Stone enabled the decipherment of Egyptian hieroglyphs. But the article notes that "multilingual text may be insufficient for a decipherment", since translation is not a reversible process.
- **Known language and personal names.** Leibniz proposed in 1714 that parallel content in bilingual inscriptions could be matched by correlating where personal names occur.
- **Statistics and patterns.** Analysts use symbol frequencies, positions within words, and recurring numeric formats, such as a small number at the end of a text possibly marking a date.
- **Community effort.** Decipherments "emerge from the incremental progress brought about by a broader community of researchers."

The classic cases

SCRIPT	WHO AND WHEN (PER THE DECIPHERMENT ARTICLE)	KEY AID
Egyptian hieroglyphs	Champollion, 1822	The Rosetta Stone

SCRIPT	WHO AND WHEN (PER THE DECIPHERMENT ARTICLE)	KEY AID
Old Persian cuneiform	Grotefend, Burnouf and Rawlinson, 1823	See Chapter 2 for Grotefend's 1802 start and the Behistun work
Linear B	Ventris, Chadwick and Kober, 1952	Recognising the language as Greek (Chapter 4)
Maya script	Knorozov and Proskouriakoff, among others, in the 1950s	Not stated in the article
Linear Elamite	Desset and colleagues, 2022	Not stated in the article

The article says Maurice Pope called the most famous three the decipherment of hieroglyphs, cuneiform and Linear B. Note the 1823 date here is the article's for Old Persian; Chapter 2 gave Wikipedia's Cuneiform article's dates for the wider cuneiform work (Grotefend from 1802, deciphered by 1857). The two articles count different stages, so I give both without merging them.

The Rosetta Stone

From Wikipedia's Rosetta Stone article (first 100,000 of 198,566 characters read).

- **The stone.** A granodiorite stele carrying three versions of a decree issued by a council of priests at Memphis for King Ptolemy V Epiphanes in 196 BC: hieroglyphs at the top, Demotic in the middle, Ancient Greek at the bottom.
- **Found.** In July 1799 by French soldiers near Rashid (Rosetta) in the Nile Delta, while strengthening Fort Julien during Napoleon's invasion of Egypt. After the French surrendered in 1801 it passed to the British; it has been displayed at the British Museum since 1802.
- **Start with Greek.** Because scholars could already read Greek, the Greek text was the starting point, and the first translations were published in 1803.
- **Young.** From 1814 to 1819 Thomas Young identified phonetic signs spelling the Greek name Ptolemaios and noted similarities between the hieroglyphic and Demotic texts, showing that Demotic was partly phonetic.
- **Champollion.** Built a phonetic hieroglyphic alphabet and announced it in a September 1822 lecture to the Académie. This fits Chapter 2's point that hieroglyphs mix sound signs with logograms and determinatives.

Still Unread

From Wikipedia's Undeciphered writing systems article (first 100,000 of 124,335 characters read). It defines such a system as one for which "no decipherment has achieved broad acceptance in the specialist literature".

SCRIPT	DATE (PER THE ARTICLE)	NOTE
Proto-Elamite	c. 3100 to c. 2900 BCE	
Indus script	c. 2800 to c. 1900 BCE	Claimed decipherments remain disputed
Cretan hieroglyphs	c. 2100 BCE	
Linear A	c. 1800 to c. 1450 BCE	Sign values often transcribed by comparison with Linear B, but the language is unknown
Byblos syllabary	c. 1700 BCE	
Olmec / Isthmian	Olmec c. 1000 to c. 400 BCE; Isthmian c. 400 BCE to c. 500 CE	Partial readings proposed for Isthmian; overall decipherment disputed
Rongorongo	Attested by the 19th century	Often hypothesised to encode language
Voynich manuscript	Parchment radiocarbon-dated to the early 15th century	

- **Why they resist.** The article gives general causes: no bilingual texts or other external anchors such as securely identified names or dates; uncertain or unknown underlying languages; small corpora; loss of archaeological context; and, for some, uncertainty whether the signs are writing at all.
- **The hardest case.** The Decipherment article, citing Gelb and Whiting, says that when the writing is known but the language is not, decipherment is "extremely difficult to achieve on the basis of internal information only".
- **Count.** "At least a dozen languages remain undeciphered" (Decipherment article).

WATCH OUT: DO NOT TRUST "CRACKED!" HEADLINES

> The Undeciphered writing systems article says decipherment claims exist for the Indus script, the Phaistos Disc and the Isthmian script but "remain disputed". A decipherment counts only when it has broad acceptance among specialists, and it should let others read new texts, not just one.

A Toy Decipherment (Mine)

A real decipherment is slow community work, but the statistics mentioned above can be seen in a toy. I made up ten symbols for ten English letters and "wrote" a sentence; the code finds the most common symbol and the most common whole word. I ran it and copied the output. This is an illustration of my own, not a real decipherment.

```
from collections import Counter
key = {"a": "△", "e": "○", "t": "□", "n": "◇", "o": "☆",
       "s": "▽", "r": "◊", "i": "◁", "h": "▷", "d": "✦"}
text = "the stone then rose and the notes ran on and on to the end"
coded = "".join(key.get(c, c) for c in text)
print(coded)
c = Counter(ch for ch in coded if ch != " ")
print(c.most_common(3))
print(Counter(coded.split()).most_common(2))
```

```
□▷○ ▽□☆◇○ □▷○◇ ◊☆▽○ △◇✦ □▷○ ◇☆□○▽ ◊△◇ ☆◇ △◇✦ ☆◇ □☆ □▷○ ○◇✦
[('◇', 9), ('○', 8), ('□', 7)]
[('□▷○', 3), ('△◇✦', 2)]
```

The most common whole word is the three-symbol group □▷○ (3 times), which a reader who knows English would guess is "the"; the next, △◇✦, is "and". That guess uses a known language, which a decipherer of an unknown language does not have, and it works on such a short sentence only because the sentence is built to repeat words. It shows why Type II cases in the article (known writing, unknown language) are so hard.

Scripts in Danger

Evidence here is thinner. Wikipedia had no page for the Endangered Alphabets Project at the addresses I tried (both returned errors), so this section rests on the project's own website, which is one source and is not neutral:

- The Endangered Alphabets Project publishes the *Atlas of Endangered Alphabets*, covering "Indigenous and minority writing systems, and the people who are trying to save them".
- It states that "85% of the world's writing systems are on the verge of vanishing". I could not check this figure against another source, so treat it as the project's claim. The site says that when a culture loses its traditional script, centuries of written material are lost too. Its page names no specific scripts, and it does not say who runs it.

Scripts also change by choice or policy, as earlier chapters showed: Vietnamese and Korean largely switched from Chinese characters to alphabets (Chapter 3), Kazakhstan moved from Cyrillic to Latin (Chapter 6), and Modern Greek simplified its accents (Chapter 9). Against that, Cherokee and Canadian syllabics (Chapters 4 and 8) show a small community adopting a script quickly when it fits.

Course Complete: The Whole Map

CH	TYPE OR TOPIC	ONE SYMBOL STANDS FOR	EXAMPLES
1	What writing is	(the six types introduced)	Writing versus language and drawing
2	First scripts	Words, sounds and silent category signs (rebus principle)	Cuneiform, hieroglyphs
3	Logographic	A morpheme, usually with a sound clue	Chinese characters
4	Syllabary	A syllable or mora	Kana, Cherokee, Linear B
5	Abjad	A consonant	Phoenician, Hebrew, Arabic
6	Alphabet	A consonant or a vowel	Greek, Latin, Cyrillic
7	Abugida	A consonant plus an inherent vowel	Devanagari, Ge'ez, Thai
8	Invented scripts	Designed, sometimes by feature	Hangul, Canadian syllabics, Tengwar, pIqaD
9	Rules	Direction, orthography, romanisation	English spelling, Pinyin
10	Lost and found	Decipherment and survival	Rosetta Stone, Linear A

- **The threads.** Almost every system mixes meaning and sound (Chapters 1 to 3); the alphabet's family line runs from Proto-Sinaitic through Phoenician to Greek, Latin and Cyrillic (Chapters 5 and 6); a script's type depends on what fits its language (Chapter 4).
- **The cautions.** Type names describe tendencies; the featural class is disputed; dates and counts are the sources' and mostly from one encyclopedia.
- **Where next.** For related courses on the site, see Introduction to Linguistics and Historical Linguistics and Language Families on how languages relate, and Phonetics and the IPA on the sounds scripts record. The site's Hiragana, Katakana and Korean Alphabet 1 courses and kanji pages teach individual scripts in depth.

NOTE ON SOURCES

This chapter uses Wikipedia's Decipherment (whole answer as returned), Rosetta Stone (first 100,000 of 198,566 characters) and Undeciphered writing systems (first 100,000 of 124,335) articles, and the Endangered Alphabets Project's own site for the endangered-scripts figure, which I could not cross-check. The Decipherment article did not mention Behistun or Proto-Elamite, and Proto-Elamite's date comes from the Undeciphered writing systems article. Dates are the articles' and unchecked against primary scholarship. The toy decipherment, its reading, the recap table and the cross-chapter links are mine.

Try It Yourself

EXERCISE 1

Explain in two or three sentences why the Rosetta Stone helped decipher hieroglyphs but would not, on its own, be enough. Use the article's remark about translation and Young's and Champollion's contributions.

View solution

EXERCISE 2

Run the toy decipherment with a different sentence of your own. Does the most frequent word still turn out to be a common English word? What would change if you did not know the language?

View solution

EXERCISE 3

Pick any script from the course and write four sentences: its type, one fact about its origin, one way it fits (or does not fit) its language, and one open question you still have about it.

View solution

CHAPTER 10 QUICK REFERENCE

- Decipherment recovers a forgotten convention; it is not cryptanalysis
- Aids: bilingual inscriptions, personal names, a known language, statistics, a community of researchers
- Hieroglyphs 1822 (Champollion), Linear B 1952 (Ventris, Chadwick, Kober), Linear Elamite 2022
- Still unread: Indus, Linear A, Proto-Elamite, Rongorongo and others; "cracked" claims need broad acceptance

COURSE COMPLETE

That is *Writing Systems of the World*. A combined PDF of all ten chapters is in the course's `pdfs` folder.