Lionbridge

FAQ

What are Double-Byte, Single-Byte, and Multi-Byte Encodings?

You may have heard some Asian languages described as being double-byte. What is this about?

First, some background. The characters that comprise text must be represented as numbers so that computers can deal with them. Languages with many characters require more numbers. The industry term for these numbers is “code points.”

Single-Byte

Most languages use an alphabet with a limited set of text symbols, punctuation marks, and special characters, and one byte per character suffices. One byte is enough to distinguish every possible character in such a language. One byte gives us the ability to represent 256 characters — which is enough for the combined alphabets of English, French, Italian, German, and Spanish; or, enough individually, for each of the alphabets used for Russian, Greek, Turkish, Arabic or Hebrew. These languages are sometimes called “single-byte.”

Multi-Byte

The Asian languages — Chinese, Japanese, and Korean (CJK) — are intrinsically different. Their character sets (meaning all the symbols needed to express the language) contain a subset that is less complex, including ASCII characters and punctuation marks. The subset requires one byte only. However, Asian languages also have a larger set of ideographic characters of Chinese origin — literally thousands of them. We need two or more bytes for representing such a great number of these complex characters. The term for mixing single-byte characters alongside two-or-more-byte characters is “multi-byte.”

What you Should Know

+ The fundamental issue is whether the number of characters assigned to a language exceeds 256 characters, which is the limit for a “single byte language.”

+ Western or Eastern European languages based on Latin characters do not exceed the 256 character limit.

+ Most Central European languages and Turkish use Latin-based extension characters, and these fit in the 256 character limit.

+ Russian (and related Cyrillic scripts), Greek, Thai, Arabic, and Hebrew (which use non-Latin-based characters) also do not exceed the 256 character limit.

+ Chinese, Japanese, and Korean each far exceed the 256 character limit, and therefore require multi-byte encoding to distinguish all of the characters in any of those languages.

Double-Byte

So, what are “double-byte” languages?
Double byte implies that, for every character, a fixed width sequence of two bytes is used, distinguishing about 65,000 characters. Even in early computing, however, this number was already recognized to be insufficient. This was the case with a primitive type of Unicode encoding, called UCS-2, used on older Microsoft platforms. Actually, though still widely used, the term double-byte is obsolete. Today, the term multi-byte is more properly used.

How Encodings Work

The association of languages with encodings (single-byte, double-byte, or multi-byte) has changed with the advent of modern Unicode. Unicode consists of useful things, such as:

+ A catalog of more than a million characters covering 90 scripts
+ A set of code charts for visual reference
+ An encoding methodology and set of standard character encodings
+ A bidirectional display order that handles the correct display of text containing both right-to-left scripts, such as Arabic and Hebrew, and left-to-right scripts, like English

Now, if you want to get technical, the truth is there are myriad encodings. Unicode may be the most well-known, but it co-exists with other encodings that originate from various standards organizations like ISO, ANSI, and KSC. Some of these encodings use more than one byte, even for the so-called “single-byte languages.”

For instance, a special character in French that is encoded in UTF-8 (Unicode Transformation Format with 8 bits) can be more than one byte. But don’t let that confuse you — French is still classified as a “single-byte language,” even though the encoding that may be selected for it in a specific case can be a multi-byte encoding.

About Lionbridge

Lionbridge enables more than 800 world-leading brands to increase international market share, speed adoption of products and effectively engage their customers in local markets worldwide. Using our proprietary cloud technology platforms and our global crowd of more than 100,000 professional cloud workers, we provide translation, online marketing, content management and application testing solutions that ensure global consistency and local relevance across all touch points of the customer lifecycle. Based in Waltham, Mass., Lionbridge maintains solution centers in 26 countries. To Learn more visit www.lionbridge.com.