



# Typing Different Languages into Computer Systems: A Brief Tutorial

Internationalized Naming Systems  
VeriSign Global Registry Services  
November 2002

# IDNs in any language

## Internationalized Domain Names (IDNs)



д о м е й н . c o m

환경보호국민운동.com

オルソ.com

澳洲唐人街.com

pasaz.com

россия.org

deharomañas.com

This tutorial has been developed to help explain how speakers of different languages input those languages on computer keyboards. This tutorial will demonstrate Chinese, Japanese and Korean, but other languages follow similar principles.



# Different Scripts = Different Input Methods

▶ **IDNs are available in all character sets or scripts identified in Unicode 3.1**

Arabic	Mongolian	Gurmukhi
Armenian	Myanmar	Han (Chinese, Japanese, Korean ideographs)
Bengali	Oriya	Hangul
Bopomofo (Zhuyin)	Sinhala	Hebrew
Cherokee	Syriac	Hiragana
Cyrillic	Tamil	Kannada
Devanagari	Telegu	Katakana
Ethiopic	Thaana	Khmer
Georgian	Thai	Lao
Greek	Tibetan	Latin
	Yi	Malayalam

▶ **These scripts are used in more than 350 languages**

▶ **Users around the world have adapted to input their language's scripts on computer keyboards**

- Different scripts use different keyboards or a standard (Latin-alphabet) keyboard with a soft keyboard
- Computer operating systems have Input Method Editors (IME) that facilitate the input of the scripts



# Helpful Terms

Relationship between Script, Character and Language				
Script	Latin	Arabic	Han	Greek
Character	L	س	光	Ω
Language	English	Farsi	Chinese	Greek

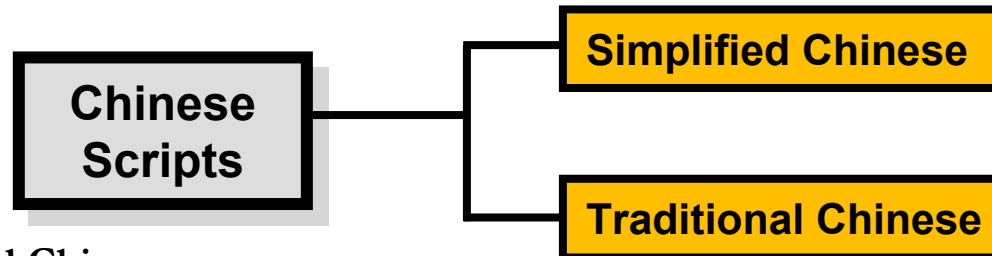
- ▶ **Script**
  - A script is a collection of symbols used to represent textual information in a language. Examples: Latin, Cyrillic, Greek,
- ▶ **Character**
  - A character, in an abstract sense, is an element of writing that is the smallest quantity having semantic value. A character is the basic building block of any script, and thus any written language. In some languages, a single character can represent an idea (Chinese for “light”: 光) while in other languages, several characters are needed to form a word that represents an idea (English: “light”).
- ▶ **Written Language**
  - A written language is a writing system made of characters from one or more scripts. Written languages are, in many cases, made up of multiple scripts. Examples of languages: English, French, Japanese, Russian, Urdu
- ▶ **Syllable**
  - A unit of spoken language consisting of a single uninterrupted sound formed by a vowel, diphthong, or syllabic consonant alone, or by any of these sounds preceded, followed, or surrounded by one or more consonants.
  - **Syllabary:** A list of syllables. A set of written characters for a language, each representing a syllable.
- ▶ **Input Method Editor (IME)**
  - A program that enables users to enter complex characters and symbols using a standard (Latin-alphabet) computer keyboard. IME programs are designed to enable users to communicate in different languages, such as Chinese, Japanese or Korean, without having to run a separate Chinese, Japanese or Korean version of the operating system.





# I. Chinese Scripts

# Overview: Chinese Scripts



## ▶ Simplified Chinese

- Used mainly in China
- Due to the complexity of many Chinese characters, a simplified version has evolved and replaced the most complex characters
  - ▶ Can require more than 10 strokes to form one character; Characters must be learned and memorized one by one
  - ▶ Not all Chinese character were simplified
  - ▶ Simplified Chinese shares many of the simpler characters with Traditional Chinese
- The People’s Republic of China’s government played an important role in the development and the adoption of Simplified Chinese
  - ▶ The goal was greater efficiency and higher literacy

## ▶ Traditional Chinese

- Used mainly in Taiwan and outside of China
  - ▶ Used by Chinese-speaking population outside of mainland China: Taiwan, Singapore and SE Asian countries

Simplified Chinese	Traditional Chinese
中国	中國
Meaning: China	Meaning: China

The first character (meaning “center”) is shared by both scripts while the more complex second characters (meaning “country”) differ.



# Input Methods: Chinese Scripts

*Chinese has more than 8,000 characters – about 2,000 characters are actively used in daily life*

- ▶ **Since it is not possible to place thousands of Chinese characters on a keyboard, Chinese uses two transliteration methods to enter Chinese characters into a computer system:**
  - Pinyin
    - ▶ Used in mainland China
  - Wade-Giles (also known as Zhuyin)
    - ▶ Used in Taiwan and other South East Asian countries with Chinese population



# Input Method: Simplified Chinese

## Simplified Chinese: *Pinyin*

- ▶ Used in mainland China
- ▶ For transliterations, Pinyin uses the Latin alphabet characters on a standard (Latin-alphabet) keyboard
- ▶ Combinations of consonants and vowels from the Latin alphabet are mapped phonetically to Chinese characters

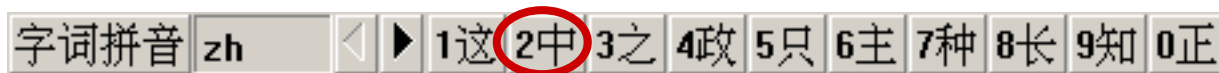


# Input Method: Simplified Chinese

## Simplified Chinese: *Pinyin*

**Example:** To enter the Simplified Chinese character for “center”, 中, a user does the following steps:

- ▶ **Step 1**
  - Using the Latin alphabet, enter the pronunciation of the first character, which sounds like “zhong”
- ▶ **Step 2**
  - As a user enters the combination of letters, the Input Method Editor (IME) of the operating system generates a list of Chinese characters that phonetically match the sound
- ▶ **Step 3**
  - Select the character meaning “center” from the list characters. This will convert the Latin alphabet characters into the intended Simplified Chinese character. In this example, the user would select #2.

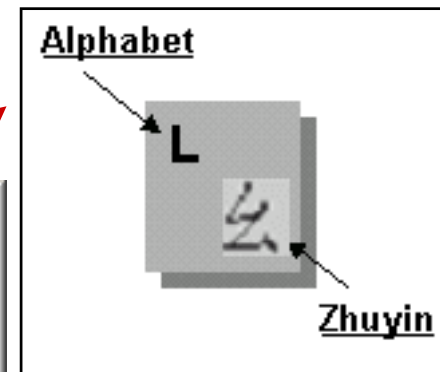


- ▶ **Repeat Steps 1 – 3 for each additional Simplified Chinese character**

# Input Method: Traditional Chinese

## Traditional Chinese: *Zhuyin*

- ▶ Used in Taiwan and other South East Asian countries with Chinese population
- ▶ In Zhuyin, combinations of consonants and vowels of Zhuyin characters are phonetically mapped to a list of Chinese characters
  - The only difference between Pinyin (used for Simplified Chinese) and Zhuyin is that Pinyin uses the English alphabet to represent phonetic sounds while Zhuyin uses Zhuyin symbols
- ▶ To enter Zhuyin characters, a user can use a special keyboard or a soft keyboard



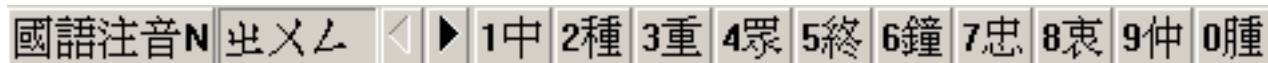
(Each key has a Zhuyin character and a Latin alphabet character)

# Input Method: Traditional Chinese

## Traditional Chinese: *Zhuyin*

**Example:** To enter the Traditional Chinese character for “center”, 中, a user does the following steps:

- ▶ **Step 1**
  - Using the Zhuyin characters to represent phonetic sounds, enter the pronunciation of the first character, which sounds like “*zhong*”
- ▶ **Step 2**
  - As a user enters the combination of Zhuyin characters, the Input Method Editor (IME) of the operating system generates a list of Chinese characters that phonetically match the sound
- ▶ **Step 3**
  - Select the character meaning “center” from the list characters. This will convert the Zhuyin characters into the intended Traditional Chinese character. In this example, the user would select #1.



- ▶ **Repeat Steps 1 – 3 for each additional Traditional Chinese character**



# Input Method: ASCII

## Traditional Chinese and Simplified Chinese

Example: To enter ASCII characters (i.e. .com), a user does the following steps:

### ▶ Step 1

- Switch the input mode from Chinese to ASCII by clicking a toggle key



### ▶ Step 2

- Type ASCII characters using the same keyboard

### ▶ Step 6

- Click the toggle key to switch back to Chinese input mode



The process of entering ASCII characters is the same for Pinyin and Zhuyin



# Spoken Chinese

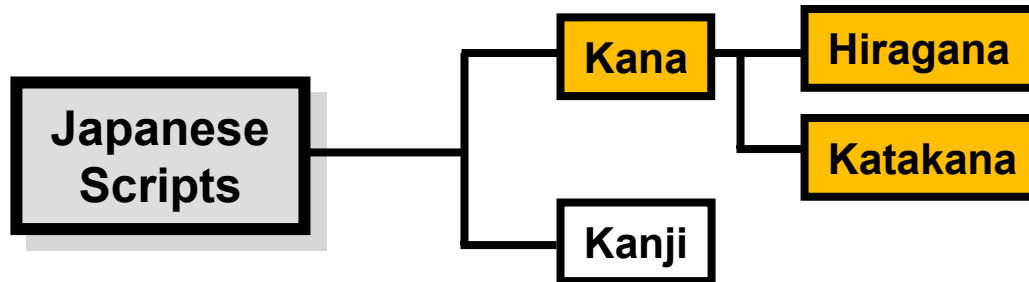
- ▶ In the spoken Chinese language, a syllable can have a number of different meanings depending on the intonation with which it is pronounced
- ▶ This tonal quality of Chinese makes transliteration from English names to Chinese confusing and potentially embarrassing
- ▶ Since IDNs are composed of local language characters, they do not rely on transliteration, therefore, any potential confusion is eliminated
- ▶ Example:

Example	Reading	Tone	Tone Name	Meaning
吗	<i>ma (1)</i>	None	<i>qingsheng</i>	question particle
妈	<i>ma (2)</i>	Flat	<i>yinping</i>	mother
麻	<i>ma (3)</i>	Rising	<i>yangping</i>	hemp, flax
马	<i>ma (4)</i>	Falling-Rising	<i>shangseung</i>	horse
骂	<i>ma (5)</i>	Falling	<i>qusheng</i>	cursing, swearing



## II. Japanese

# Overview: Japanese Scripts - Kana



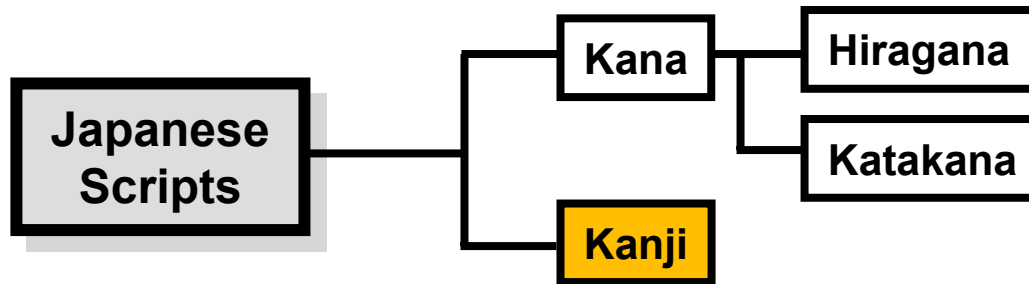
## Kana

- ▶ Kana is made up of two syllabaries (list of syllables), Hiragana and Katakana
  - Hiragana and Katakana are characters used to represent sounds

	Hiragana	Katakana
Description	Hiragana is a syllabary, not an alphabet. It consists of 48 syllables.	Katakana is a syllabary consisting of 48 syllables. They represent the same set of sounds as Hiragana.
Usage	Used to write word endings and native Japanese words for which there is no Kanji. Used in combination with Kanji, Katakana.	Used to write words of foreign origin, onomatopoeic words, foreign names and for emphasis.
Origin	Derived from Kanji. No longer carry the meaning of the Kanji from which they were derived.	Derived from Kanji. No longer carry the meaning of the Kanji from which they were derived.
Shape	Cursive appearance – a smoother, handwritten style	Squared, more rigid appearance
Examples	みりをわはたなわ	コオエケクウカタ



# Overview: Japanese Scripts - Kanji



## Kanji

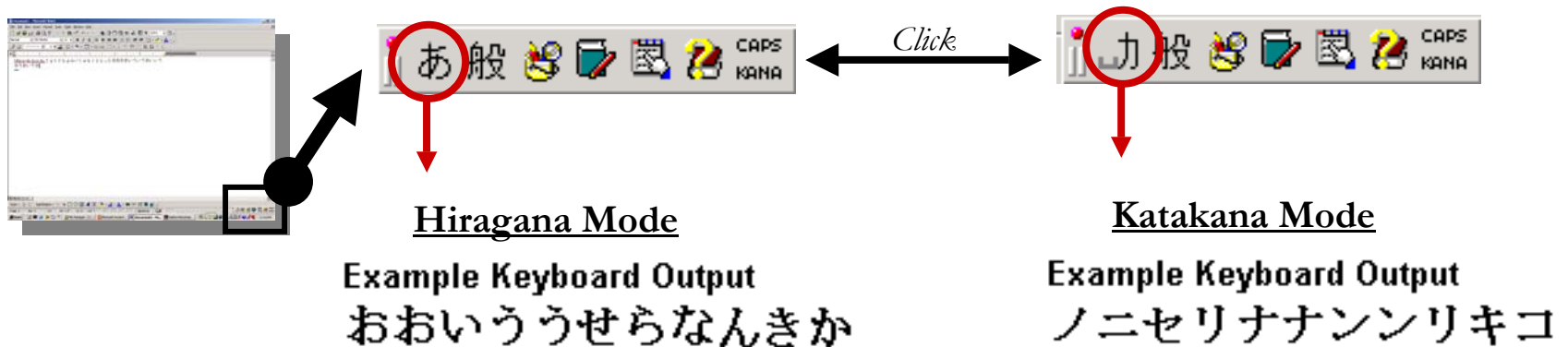
- ▶ Kanji is another name for the Chinese characters used in Japanese
- ▶ Includes characters from Traditional Chinese and Simplified Chinese as well as Japanese versions of Simplified Chinese characters. Also includes some characters invented in Japan.
  - Some characters are shared by Traditional Chinese, Simplified Chinese and Kanji
  - Some characters are shared only by Simplified Chinese and Kanji
  - Some characters are not shared at all. The same meaning is represented by different characters

Meaning	Traditional	Simplified	Kanji	Notes
<i>Center</i>	中	中	中	Character is shared by Traditional Chinese, Simplified Chinese and Kanji
<i>Country</i>	國	国	国	Character is shared only by Simplified Chinese and Kanji
<i>China</i>	中國	中国	中国	“China” is written using a combination shared characters and not-shared characters
<i>Air</i>	氣	气	気	Characters are not shared but have the same meaning

# Input Method: Kana

## Kana

- ▶ Modern Japanese is written with a mixture Kana and Kanji
  - For an average writing sample, one normally finds 60% Hiragana, 10% Katakana, and 30% Kanji. Actual percentages depend upon the nature of the text
    - ▶ Technical and formal literature may contain more Kanji
- ▶ Users click a toggle key on the keyboard to switch between the two writing systems of Kana: Hiragana and Katakana





# Input Method: Kanji

## Kanji

- ▶ As with Chinese, it is not possible to place thousands of Kanji characters on a keyboard, so Japanese uses Hiragana as a transliteration method to enter Kanji into a computer system.

**Example: To enter the Kanji character for “gate”, 間 , a user does the following steps:**

### ▶ Step 1

- Switch the input mode to “Hiragana” by clicking a toggle key



### ▶ Step 2

- Using the Hiragana characters to represent phonetic sounds, enter the pronunciation of the character, which sounds like “MA”, ま.

### ▶ Step 3

- Click the “Conv” (“conversion”) button on the IME or the space bar on the keyboard and the IME generates a list of Kanji characters that phonetically match the MA sound of the hiragana.

### ▶ Step 4

- Select the character meaning “gate” from the list of characters. This will convert the Hiragana character (ま) into the intended Kanji character (間). In this case, the user would select #1.

- ▶ Repeat Steps 2 – 4 for each additional Kanji character



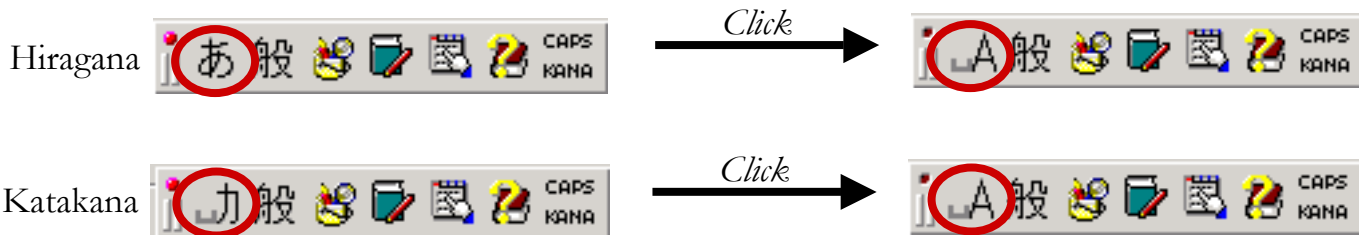
# Input Method: ASCII

## Japanese

Example: To enter ASCII characters (i.e. .com), a user does the following steps:

### ▶ Step 1

- Switch the input mode from Hiragana or Katakana to ASCII by clicking the toggle key



### ▶ Step 2

- Type ASCII characters using the same keyboard

### ▶ Step 3

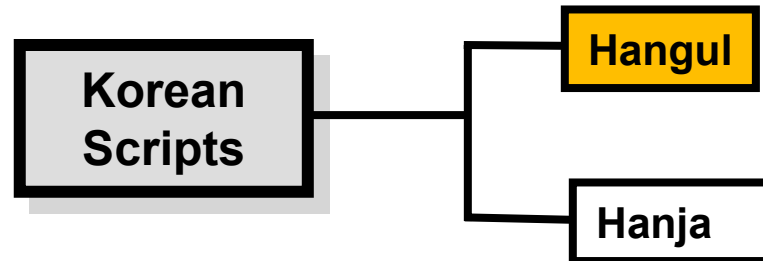
- Click the toggle key to switch back to Hiragana or Katakana input mode





## III. Korean Scripts

# Overview: Korean Scripts - Hangul

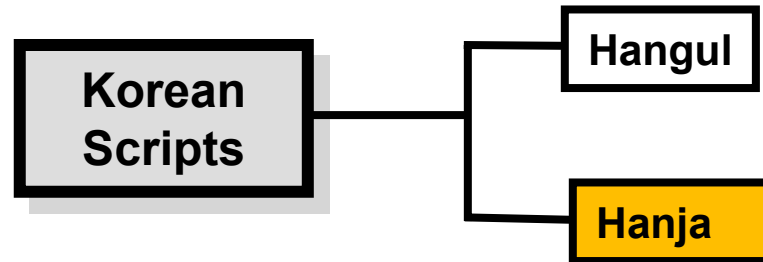


## Hangul

- ▶ Hangul is the Korean alphabet.
- ▶ Composed of 24 letters (jamo) – 14 consonants and 10 vowels
  - Examples of the letters (jamo): ㅏ ㅎ ㄱ ㅓ ㄹ ㅈ
  - The letters are combined into syllable blocks

Jamo	Pronunciation	Hangul
가	GA	ㄱ plus ㅏ
갈	GAL	ㄱ plus ㅓ plus ㄹ
갈	GALG	ㄱ plus ㅓ plus ㄹ plus ㄱ

# Overview: Korean Scripts - Hanja



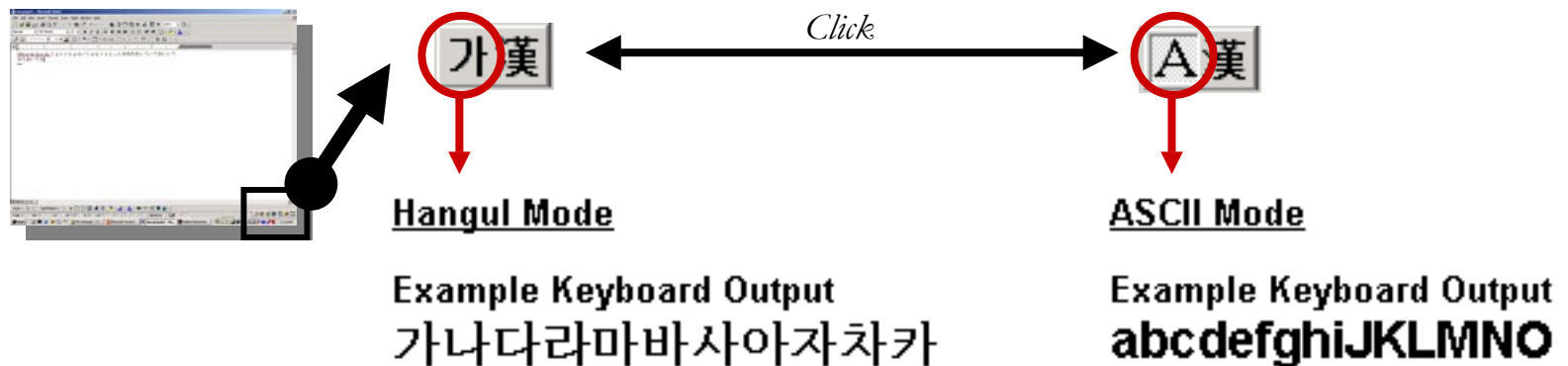
## Hanja

- ▶ Hanja is another name for the Chinese characters used in Korean
- ▶ Includes characters from Traditional Chinese only
- ▶ The use of Hanja characters mixed with Hangul characters is becoming less popular in everyday communications
  - The mixture of scripts is used primarily for formal documents (i.e. legal papers) or personal names
- ▶ The Korean language can not be written entirely in Hanja characters

# Input Method: Korean - Hangul

## Hangul

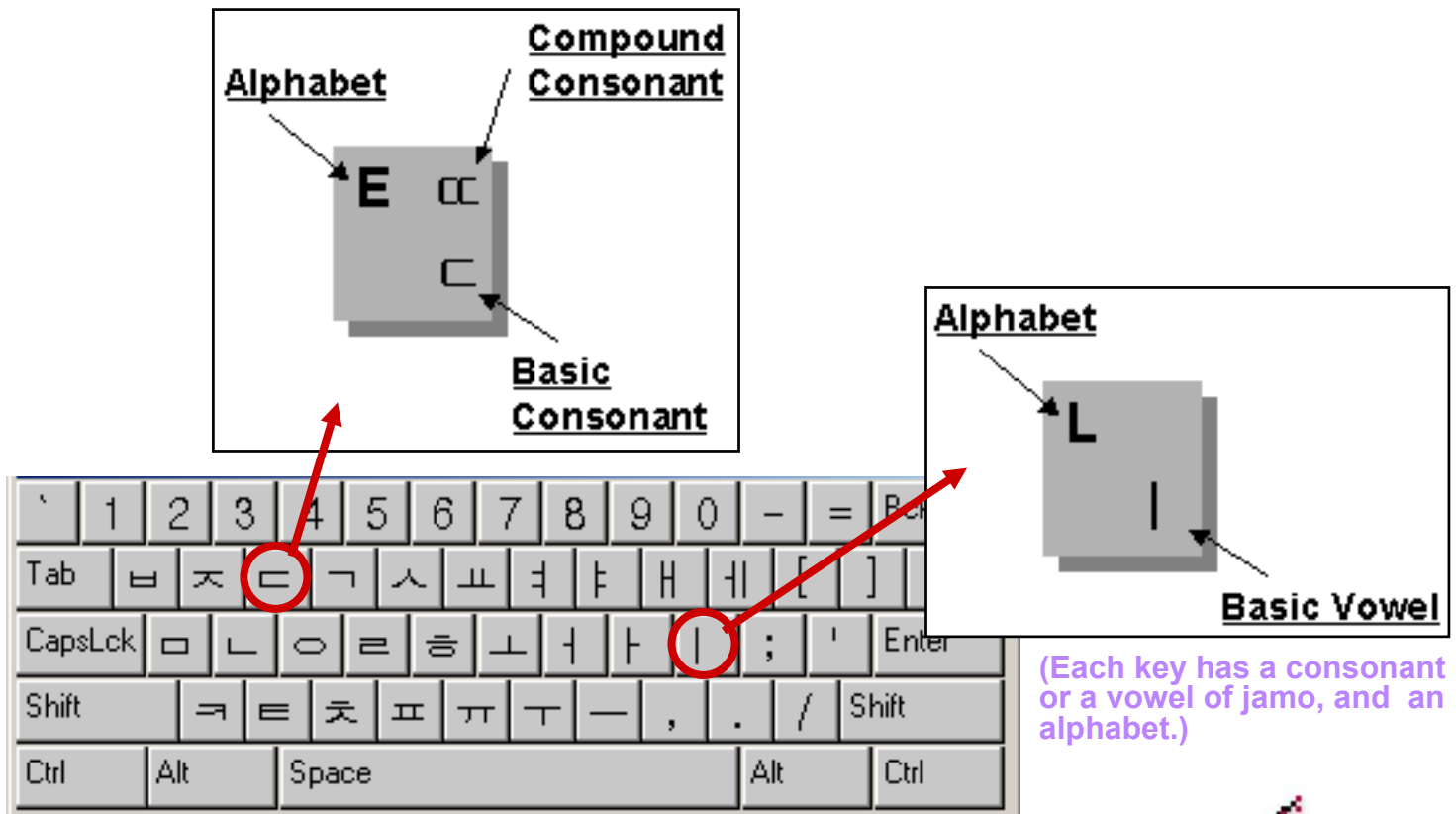
- ▶ Users click a toggle key on the keyboard to switch between Hangul mode and ASCII mode.
- ▶ In Hangul mode, users enter the letters (jamo) in a sequence (top to bottom and left to right). The IME will automatically compose and combine letters into the Hangul syllable blocks



# Input Method: Korean - Hanguk

## Hanguk

- ▶ To enter Hiragana or Katakana characters, a user can use a special keyboard or a soft keyboard



# Input Method: Korean - Hanja

## Hanja

- ▶ As with Chinese, it is not possible to place thousands of Hanja characters on a keyboard, so Korean uses Hangul as a transliteration method to enter Hanja into a computer system.

**Example: To enter the Hanja character for “gate” a user does the following steps:**

### ▶ Step 1

- Switch the input mode to “Hangul” by clicking a toggle key



### ▶ Step 2

- Using the Korean letters (jamo), enter a syllable and highlight it:



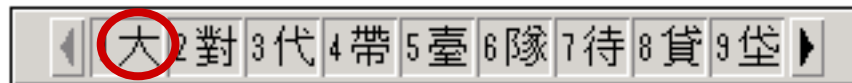
### ▶ Step 3

- Click the Conversion button and the IME generates a list of Hanja characters that phonetically match the sound of the Hangul.



### ▶ Step 4

- Select the character meaning “gate” from the list characters. This will convert the Hangul character into the intended Hanja character. In this example, the user would select #1.



- ▶ Repeat Steps 2 – 4 for each additional Hanja character



# Input Method: ASCII

## Hangul

Example: To enter ASCII characters (i.e. .com), a user does the following steps:

▶ **Step 1**

- Switch the input mode from Hangul to ASCII by clicking the toggle key



▶ **Step 2**

- Type ASCII characters using the same keyboard

▶ **Step 3**

- Click the toggle key to switch back to Hangul input mode

