Implementing Cross-Locale CJKV Code Conversion

Ken Lunde

CJKV Type Development
Adobe Systems Incorporated

ftp://ftp.oreilly.com/pub/examples/nutshell/ujip/unicode/iuc13-c2-paper.pdf
ftp://ftp.oreilly.com/pub/examples/nutshell/ujip/unicode/iuc13-c2-slides.pdf

Code Conversion Basics

Adobe

- Algorithmic code conversion
  - Within a single locale: Shift-JIS, EUC-JP, and ISO-2022-JP
  - A purely mathematical process
- Table-driven code conversion
  - Required across locales: Chinese ↔ Japanese
  - Required when dealing with Unicode
  - Mapping tables are required
  - Can sometimes be faster than algorithmic code conversion—depends on the implementation

Copyright © 1998 Adobe Systems Incorporated

Code Conversion Basics (Cont'd)

- CJKV character set differences
  - Different number of characters
  - Different ordering of characters
  - Different characters

Copyright © 1998 Adobe Systems Incorporated

Character Sets Versus Encodings

- **Common CJKV character set standards**
  - China: GB 1988-89, GB 2312-80; GB 1988-89, GBK
  - Taiwan: ASCII, Big Five; CNS 5205-1989, CNS 11643-1992
  - Hong Kong: ASCII, Big Five with Hong Kong extension
  - Japan: JIS X 0201-1997, JIS X 0208:1997, JIS X 0212-1990
  - South Korea: KS X 1003:1993, KS X 1001:1992, KS X 1002:1991
  - North Korea: ASCII (?), KPS 9566-97
  - Vietnam: TCVN 5712:1993, TCVN 5773:1993, TCVN 6056:1995
- **Common CJKV encodings**
  - Locale-independent: EUC-*, ISO-2022-*
  - Locale-specific: GBK, Big Five, Big Five Plus, Shift-JIS, Johab, Unified Hangul Code
  - Other: UCS-2, UCS-4, UTF-7, UTF-8, UTF-16

Copyright © 1998 Adobe Systems Incorporated

Chinese Character Relationships

- Simplified and Traditional forms
  - Traditional forms still used in Taiwan and Korea
  - Simplified forms used in Japan
  - Greatly simplified forms used in China
  - Simplified/Traditional relationship not always one-to-one
  - Simplified/Traditional relationship is locale-specific
- Variant forms
  - Alternate character forms with same semantics
- Common forms
  - Luckily, many forms are still common across all CJKV locales:

    一 山 字 人 正 大 田 白 血

Copyright © 1998 Adobe Systems Incorporated

Advantages of Unicode

Adobe

- A common representation for all characters
  - Only $2n$ mapping tables required (where n equals the number of supported character set plus encoding combinations)
  - Otherwise, $n \times (n - 1)$ mapping tables would be required.
- Mapping tables are readily available at the following URL:

  ftp://ftp.unicode.org/
- Unicode is used as information interchange code

Copyright © 1998 Adobe Systems Incorporated

Handling Common Characters

- Trivial effort that goes through Unicode
- Consider KS X 1001:1992 to JIS X 0208:1997 conversion:

Glyph	KS X 1001:1992	⇒	Unicode	⇒	JIS X 0208:1997	Glyph
一	76-73		4E00		16-76	一
山	63-03		5C71		27-19	山
田	79-03		7530		37-36	田
血	90-76		8840		23-76	血

Copyright © 1998 Adobe Systems Incorporated

Handling Simplified/Traditional

Adobe

- More problematic because simplified/traditional databases are required
  - A Unicode code point does not reflect such relationships!
- Consider GB 2312-80 to JIS X 0208:1997 conversion:

Glyph	GB 2312-80	⇒	Unicode	⇒	Unicode'	⇒	JIS X 0208:1997	Glyph
黑	26-58		9ED1		**9ED2**		25-85	黑
汉	26-26		6C49		**6F22**		20-33	漢

- Note how the Unicode representation is altered to acheive a successful conversion
- Japan and China share many simplified forms (such as 国), which can result in more direct mappings

Copyright © 1998 Adobe Systems Incorporated

Handling Simplified/Traditional (Cont'd)

- Word-level disambiguation is required
- Consider the traditional representations of 霉:
  - 霉 itself
  - 黴
- Disambiguation can be resolved through context

  霉雨 ⇒ 霉雨 (*méiyǔ*)

  霉菌 ⇒ 黴菌 (*méijūn*)

Copyright © 1998 Adobe Systems Incorporated

Handling Variants Forms

Adobe

- There are often more than one variant form for a given Chinese character
- Consider the following Japanese-specific character relationships (all in JIS X 0208:1997):

Standard Form	Traditional Forms	General Variants
学	學	斈
剣	劍	劒劍剱釼
辺	邊	邉
弁	辨瓣辯	辦

Copyright © 1998 Adobe Systems Incorporated

Handling The Compatibility Zone

Adobe

- KS X 1001:1992 includes 268 duplicate hanja (with multiple readings)—encoded in "CJK Compatibility Zone"
- Consider KS X 1001:1992 to GB 2312-80 conversion:

Glyph	KS X 1001:1992	⇒	Unicode	⇒	Unicode'	⇒	Unicode"	⇒	GB 2312-80	Glyph
樂	49-66		F914		**6A02**		**4E50**		32-54	乐
樂	53-05		F95C		**6A02**		**4E50**		32-54	乐
樂	68-37		6A02		6A02		**4E50**		32-54	乐
樂	72-89		F9BF		**6A02**		**4E50**		32-54	乐

- Note the two transformations in the Unicode representation
  - — Normalize to the same code point (0x6A02)
  - — Convert to simplified hanzi (0x4E50)

Copyright © 1998 Adobe Systems Incorporated

Handling Unmappable Characters

- Some characters are simply not available in the target character set plus encoding combination, not even as a variant form
  - Consider 弼 (JIS X 0208:1997 55-27, U+5F41)
  - Consider Korean hangul
- Output such characters as tags for round-trip code conversion purposes
  - Such as HTML/XML character references: `&#x5F41;`
- Remove such characters from the output altogether—not terribly graceful

Copyright © 1998 Adobe Systems Incorporated

Avoiding Code Conversion Pitfalls

Adobe

- Same glyph but different semantics
- Consider GB 2312-80 to JIS X 0208:1997 conversion:

Glyph	GB 2312-80	⇒	Unicode	⇒	Unicode'	⇒	Unicode"	⇒	JIS X 0208:1997	Glyph
气	38-88		6C14						61-67	气
气	38-88		6C14		**6C23**		**6C17**		21-04	気
气	38-88		6C14		**6C23**				61-70	氣

- The above illustrates three possible scenarios, depending on how the semantics of 气 (U+6C14) are interpreted
  - — As a Radical?
  - — As a Simplified (standard) form?
  - — As a Traditional form? (also in JIS X 0208:1997)

Copyright © 1998 Adobe Systems Incorporated

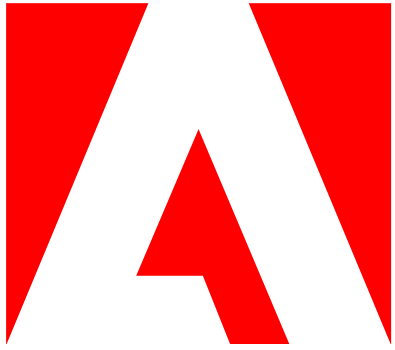
Cross-Locale Code Conversion Tools

- CJKVConv.pl
  - — Developed in Perl by Ken Lunde
  - — Test vehicle to illustrate cross-locale code conversion issues

    ftp://ftp.oreilly.com/pub/examples/nutshell/ujip/perl/cjkvconv.pl
- Uniconv
  - — Developed by Basis Technology
  - — Supports a wide variety of character sets and encodings
  - — Provides many transformations

    http://www.basistech.com/unicode/

Copyright © 1998 Adobe Systems Incorporated



Adobe