

Unicode™ and East Asian Ideographs

John H. Jenkins

井作恆

International and Text Group
Apple Computer, Inc.

Technical Director
Unicode, Inc.

`jenkins@apple.com`

OCLC CJK Users Group Annual Meeting
Chicago, Illinois, March 24, 2001

Topics of Discussion

- Unicode in General
 - What is Unicode?
 - History of Unicode
 - Encoding Forms
 - Unicode's Ten Design Principles
 - Benefits of Unicode
- Unicode and Ideographs
 - Ideographs and the IRG
 - Ideographs in Unicode (Unihan)
 - Future Compatibility Ideographs
 - Ancillary Characters
 - Solving the Variants Problem

Unicode in General

- What is Unicode?
- History of Unicode
- Encoding Forms
- Unicode's Ten Design Principles
- Benefits of Unicode

What is Unicode?

- Unicode™
 - A trademark owned by Unicode, Inc.
- Unicode Consortium (Unicode, Inc.)
 - Owns and promotes the standard
 - <http://www.unicode.org/>
- Unicode Technical Committee
 - Designs and modifies the standard
- Unicode Standard (in books and on-line)
 - *Is* the standard
 - Unicode 3.0 available in bookstores
 - Unicode 3.1 now available on the Web
- ISO/IEC 10646
 - International standard closely related to Unicode
 - Developed by JTC1/SC2/WG2

OCCLC CJK Users Group Annual Meeting
Chicago, Illinois, March 24, 2001

History of Unicode

- 1989: Drafts of 10646 and Unicode
- 1991: Unicode 1.0
- 1993: ISO/IEC 10646-1:1993 & Unicode 1.1
 - Merged standard
- 1996: Unicode 2.0
 - 21,204 ideograph, including 302 compatibility ideographs
- 1998: Unicode 2.1 (added)
- 2000: Unicode 3.0
 - 6582 new ideographs
- 2001: Unicode 3.1
 - 43,253 new ideographs, including 530 new compatibility ideographs

Unicode Encoding Forms

- Unicode “scalar values” are numbers between 0 and 1,114,111
 - Each Unicode character corresponds to a single scalar value
 - Scalar values are referred to using “U+” and four to six hexadecimal digits
- Encoding forms
 - UTF-8 (for byte- oriented data and storage)
 - Each Unicode scalar value requires between 1 and 4 bytes
 - Prior to Unicode 3.1, encoded scalar values required 1 to 3 bytes
 - UTF-16 (for word-oriented data and storage)
 - Each Unicode scalar value requires 2 or 4 bytes
 - Prior to Unicode 3.1, encoded scalar values required 2 bytes
 - UTF-32 (you get the picture)
 - Each Unicode scalar value requires 4 bytes

Unicode is Efficient

- Unicode text is simple to parse and process
 - Limited forward- or back-scanning is needed in some rare circumstances
- Unicode text is not stateful
 - Proper *rendering* may require some state data
 - If part of the data is lost, the rest can be interpreted

Unicode Encodes Characters, Not Glyphs

- Visual variation is not important
 - An *A* is an **A** is an *A* is an *A*
- Fundamental identity *is* important
- Character to glyph mapping is *m* to *n*
 - E.g., ligatures (fi, fl)
 - E.g., accented letters (é, å)

Unicode Defines Plain Text

- Defaults of text formatting left to higher-level protocols
 - Typeface, point-size, italicization, boldfacing, ligature control, embedding graphics or other data, language tags
- Minimum amount of information for legible display

Unicode Uses Logical Order

- The default for memory representation is logical order
- Scripts are written on paper in different directions
 - Most go left-to-right across a page
 - Arabic, Hebrew, Syriac, and Thaana go right-to-left
 - Some reordering can also happen

Unicode Unifies Characters From Different Scripts

- Where the *identity* of the character is not in question, even if the *shape* varies from place to place
- *Italic*, Roman, **Blackletter** letter-forms in Latin
- English and Polish accents
- Chinese, Japanese, Korean, and Vietnamese ideographs
- Greek and Coptic
 - ΑΒΓΔΕΖΗΘΙΚΛΜΝΞΟΠΡΣΤΥΦΧ ΨΩ
 - ⲀⲂⲄⲆⲈⲊⲌⲎⲏⲑⲓⲕⲗⲙⲛⲝⲟⲡⲣⲤⲥⲧⲩⲫⲭⲮⲰⲲⲳⲴⲶⲷⲸⲹⲺⲻⲼⲽⲾⲿ

Unicode Uses Dynamic Composition

- Not everything has to be explicitly in the standard to be used
- Accented forms can be freely created
- Note the presence of Ideographic Description Sequences, new in 3.0
 - Stop-gap to represent unencoded ideographs

Unicode Uses Equivalent Sequences

- Many forms can be represented in more than one way
 - Many accents
 - E.g., e + ´ is equivalent to é
 - Some ligatures
 - E.g., fi is equivalent to f + i in most instances
 - Other characters
 - E.g., Å may be Angstrom sign or A + ring

Unicode is Convertible

- Superset of most current character set standards
 - Macintosh and Windows code pages
 - National standards
 - GB, JIS, CNS, KS X
 - Industrial standards
 - Big Five, Xerox
- Use as a hub in character set conversion, or use to avoid having to implement multiple character sets

Benefits of Unicode

- Write your code only once
- Can be used to provide virtual support for other character sets
- Contains a lot of characters not readily available anywhere else
- Rapidly becoming the lingua franca of the Web
- Future script development is all being done in Unicode or with Unicode in mind

Unicode and Ideographs

- Ideographs and the IRG
- Ideographs in Unicode
- Future Compatibility Ideographs
- Ancillary Characters
- Solving the Variants Problem

How Many Ideographs Are There?

- Short answer: "Nobody knows", but up to 100,000
- *KangXi* dictionary contains some 50,000
 - All of KangXi is now in Unicode
- The line between character & glyph is often fuzzy
 - Sometimes language-specific
 - Sometimes person-specific
- Rare, obscure, obsolete, & nonce forms abound
- How many English words are there?

井井
說說

The Ideographic Rapporteur Group

- Successor to CJK-JRG
- Organized by WG2
- Main officers are rapporteur and editor
- Members include mainland China, Taiwan, Hong Kong SAR, Japan, North and South Korea, Vietnam, Singapore, and the US/Unicode
- Web site: <http://www.cs.cuhk.edu.hk/~irg>

How Ideographs are Defined by the IRG

- IRG *defines* ideographs by their mappings
 - Official national standards
 - Copies submitted by member countries
 - National character collections
 - Submitted to IRG by member countries
- Ideographs also *identified* by dictionary positions
 - KangXi
 - Dai Kan-Wa Jiten
 - Hanyu Da Zidian
 - Dae Jaweon
 - Virtual positions are assigned for characters missing from a given dictionary
- Unicode provides additional data
 - Additional mappings
 - Readings, definitions
 - Send corrections to errata@unicode.org

Ideographs in Unicode 3.1 (Unihan)

- Sources were major national standards for mainland China, Taiwan, Japan, South Korea, Hong Kong SAR, Vietnam, and additional word lists
 - Chinese Dictionaries: KangXi, Hanyu Da Zidian, Hanyu Da Cidian, Ci Hai (PRC edition), Ci Yuan (Shuowen already covered)
- Compatibility ideographs added to Unicode for round-trip compatibility with KS X 1001, Big Five, CNS 11643-1992, and major industrial standards
- Total ideograph count: 71,089 (76%) of Unicode
 - Larger than most other ideographic coded character sets in common use
 - 70,195 main ideographs and 844 compatibility ideographs

Ideograph Blocks

- CJK Unified Ideographs
 - Encoded at U+4E00 through U+9FA5
 - Encoded in Unicode since 1.1
- CJK Unified Ideographs Extension A
 - Encoded at U+3400 through U+4DB5
 - Encoded in Unicode since 3.0
- CJK Unified Ideographs Extension B
 - Encoded at U+20000 through U+2A6D6
 - Encoded in Unicode in 3.1
- Each block ordered by “four-dictionary algorithm”

Compatibility Ideograph Blocks

- CJK Compatibility Ideographs
 - Encoded at U+F900 through U+FA2D
 - Encoded in Unicode since 1.1
 - Includes
 - Pronunciation variants for KS X 1001
 - Duplicate ideographs from Big Five
 - 32 Ideographs from IBM-J
 - Includes 12 “unique” ideographs
- CJK Compatibility Ideographs Supplement
 - Encoded at U+2F800 through U+2FA1D
 - Encoded in Unicode 3.1
 - Includes new compatibility ideographs for CNS 11643-1992

How Ideographs are Encoded

- All ideographs encoded in Unicode prior to 3.1 are encoded in the Basic Multilingual Plane (BMP)
 - One to three bytes per character in UTF-8, two bytes in UTF-16
- All ideographs encoded in Unicode from 3.1 onwards are encoded in the Supplementary Ideographic Plane (SIP)
 - Four bytes per character in UTF-8 and UTF-16
- An additional plane for ideographs may be allocated if necessary

Future Compatibility Ideographs

- New compatibility ideographs proposed for
 - JIS X 0213
 - GCCS HKSCS
 - Z29.64-1989 (EACC)
- Some on BMP (U+FA2E–U+FAFF), some in SIP

Ideographic Variation Mark

- Allows an encoded ideograph to stand for a similar, unencoded ideograph
- Must have a visual appearance
- Encoded at U+303E

⚭ 乾 for 𤄎

犧牲既成，粢盛既潔，祭祀以時；
然而旱⚭乾水溢，則變置社稷。

Ideographs & Composition: Splitting Ideographs Up

- The vast majority of ideographs are combinations of a radical plus a phonetic element
 - Dictionaries even assign radicals where characters don't have them naturally
- Ideographs are also traditionally seen as built up of a small number of frequently-used stroke shapes
- Decomposition of ideographs into pieces is a common, natural approach

Ideographs & Composition: Some Approaches

- Composition schemes for ideographs have a long history on computers
- Decomposition can be:
 - Decomposition into strokes
 - Frequently used in fonts
 - Also frequent in input methods
 - Glyph-oriented, rather than character oriented
 - Decomposition into semantic pieces
 - Frequent in dictionaries, infrequent on computers
 - What are the correct semantic pieces
 - How far do you decompose?

CJK Radical Sets

- Based on the 214 radicals of the KangXi dictionary
 - Universally understood, not universally used
- Radicals may have multiple shapes
 - Dictionaries usually list *all* a radical's shapes
 - Multiple shapes encoded

人
亻

- "Basic" shape in the KangXi radicals block (U+2F00–U+2FD5)
- Additional shapes in the CJK Radicals Extension block (U+2E00–2EF3)

CJK Radical Sets

- Radicals have been given English names
 - Most are common and readily recognized by English-speaking sinologists

人	U+2F08	KANGXI	RADICAL	MAN	亻	U+2E85	CJK	RADICAL	PERSON
心	U+2F3C	KANGXI	RADICAL	HEART	冫	U+2E95	CJK	RADICAL	SNOUT TWO
手	U+2F3F	KANGXI	RADICAL	HAND	忄	U+2E96	CJK	RADICAL	HEART ONE
水	U+2F54	KANGXI	RADICAL	WATER	扌	U+2E98	CJK	RADICAL	HAND
金	U+2FA6	KANGXI	RADICAL	GOLD	氵	U+2EA1	CJK	RADICAL	WATER ONE
阜	U+2FA9	KANGXI	RADICAL	MOUND	钅	U+2ED0	CJK	RADICAL	C-SIMPLIFIEDED GOLD

CJK Radical Sets

- Radicals should be treated differently from ideographs
 - U+4E00 ≠ U+2F00
 - Radicals are "symbols", ideographs are "letters"
 - May (or may not) distinguish them visually

馬 0x2FBA KANGXI RADICAL HORSE

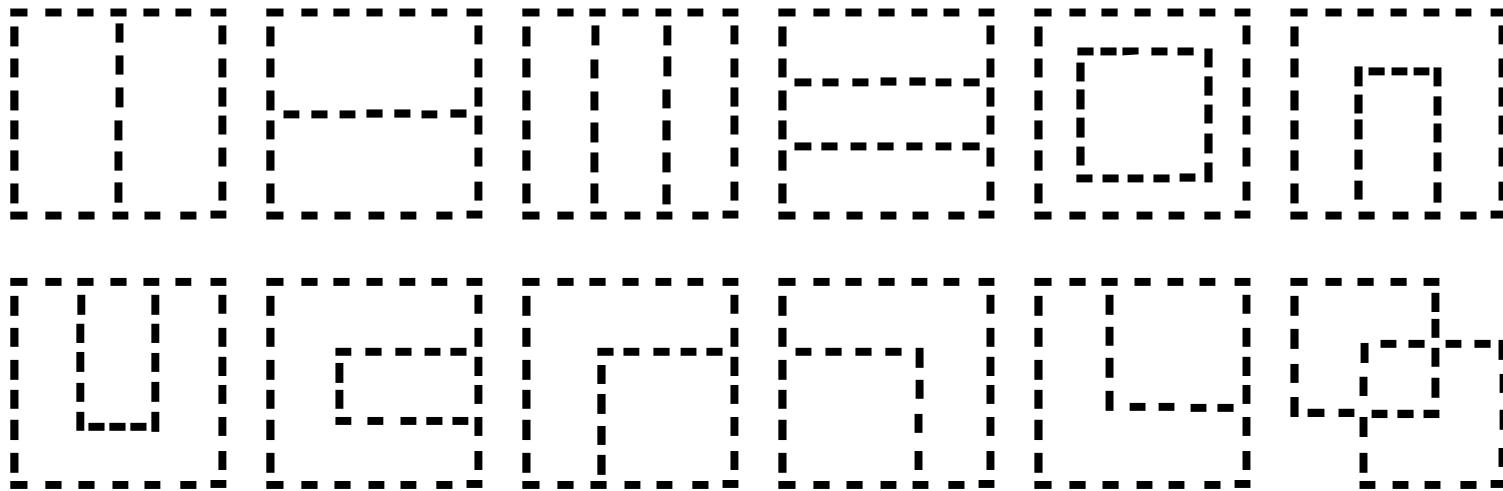
馬 U+99AC CJK UNIFIED IDEOGRAPH-99AC

Describing Ideographs

- Intended to be a stop-gap to provide a way of representing as-yet unencoded ideographs
- Two main features:
 - A set of description operators
 - Rules for using them to describe ideographs
- Note: This is a description, not an encoding
 - Equivalent to “little ‘e’ with an acute accent on it”
 - Not equivalent to U+0065 U+0301

Ideographic Description Characters

- Set of twelve characters in Unicode 3.0
- Encoded starting at U+2FF0
- Used to describe the geometric relationship between two pieces of an ideograph

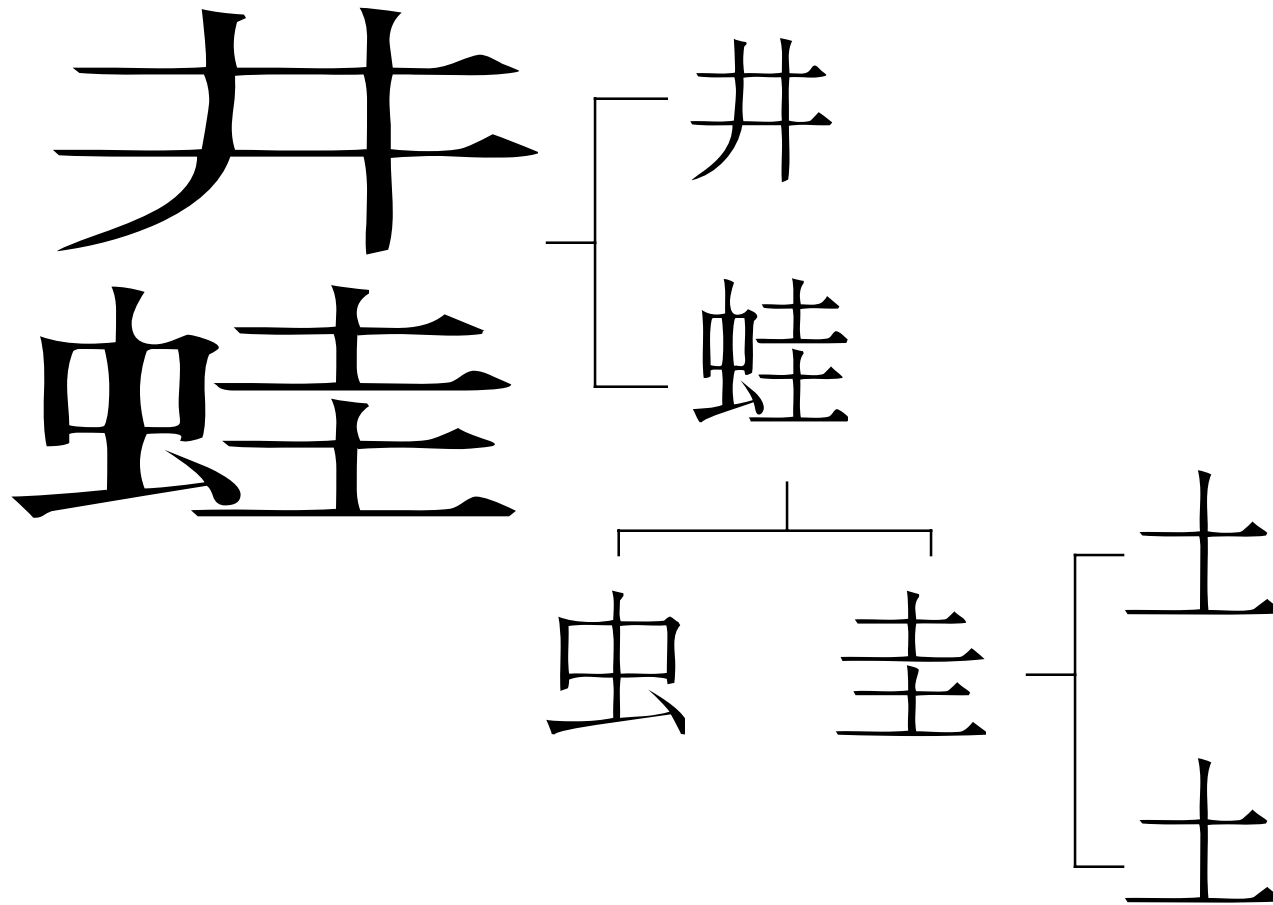


Ideographic Description Sequences

- Grammar defined in Unicode 3.1

```
IDS ::= Ideograph |  
       Radical |  
       BinaryDescriptionOperator IDS IDS |  
       TrinaryDescriptionOperator IDS IDS IDS  
  
BinaryDescriptionOperator ::= U+2FF0 | U+2FF1 | U+2FF4 |  
                               U+2FF5 | U+2FF6 | U+2FF7 | U+2FF8 | U+2FF9 | U+2FFA |  
                               U+2FFB  
  
TrinaryDescriptionOperator ::= U+2FF2 | U+2FF3  
  
Radical ::= U+2E80 | U+2E81 | ... | U+2EF2 | U+2EF3 |  
            U+2F00 | U+2F01 | ... | U+2FD4 | U+2FD5  
  
Ideograph ::= U+3400 | U+3401 | ... | U+4DB4 | U+4DB5 |  
             U+4E00 | U+4E01 | ... | U+9FA4 | U+9FA5 | U+FA0E |  
             U+FA0F | U+FA11 | U+FA13 | U+FA14 | U+FA1F | U+FA21 |  
             U+FA23 | U+FA24 | U+FA27 | U+FA28 | U+FA29 |  
             U+20000 | U+20001 | ... | U+2A6D5 | U+2A6D6
```

Describing Ideographs



OCLC CJK Users Group Annual Meeting
Chicago, Illinois, March 24, 2001

Using Ideographic Descriptions

井蛙

井 蛙

井 虫 圭

井 虫 土 土

Describing Ideographs

Describe 乾

犧牲既成，粢盛既潔，祭祀以時；然而旱日工日，水溢，則變置社稷。

IDS: Rendering

犧牲既成，粢盛既潔，
祭祀以時；然而旱□□□
□□□日日工日乞水
溢，則變置社稷。

IDS: Rendering

犧牲既成，粢盛既潔，
祭祀以時；然而旱 []
[] [] [] 日日工日万乞]
水溢，則變置社稷。

IDS: Rendering

- Use a special font
- "Undescribe" on the fly??!?

犧牲既成，粢盛既潔，
祭祀以時；然而旱^暵水
溢，則變置社稷。

IDS: Back-scan Length

- How many ideographs do you have to look back through before you *know* you're not part of an IDS?
- Maximum allowed is six

日 日 工 日 𠂇 乞

Back-scan length is 6

井 虫 土 土

Back-scan length is 2

IDS: Recursion Depth

- What is the maximum number of pending operations in the IDS?
- No limit set

日 日 工 日 万 乞

Recursion depth is 4

井 土 土 土 土

Recursion depth is 3

IDS: Total Length

- Total number of characters in IDS
- *Should* be short as possible
- No more than sixteen

日 日 工 日 丐 乞

Total length is 11

井 虫 土 土

Total length is 7—**Illegal!**

IDS: Total Length

- Total number of characters in IDS
- Should be as short as possible
- No more than sixteen

日 日 工 日 万 乞

Total length is 11

井 蛙

Total length is 3—**OK!**

IDS: Equivalence

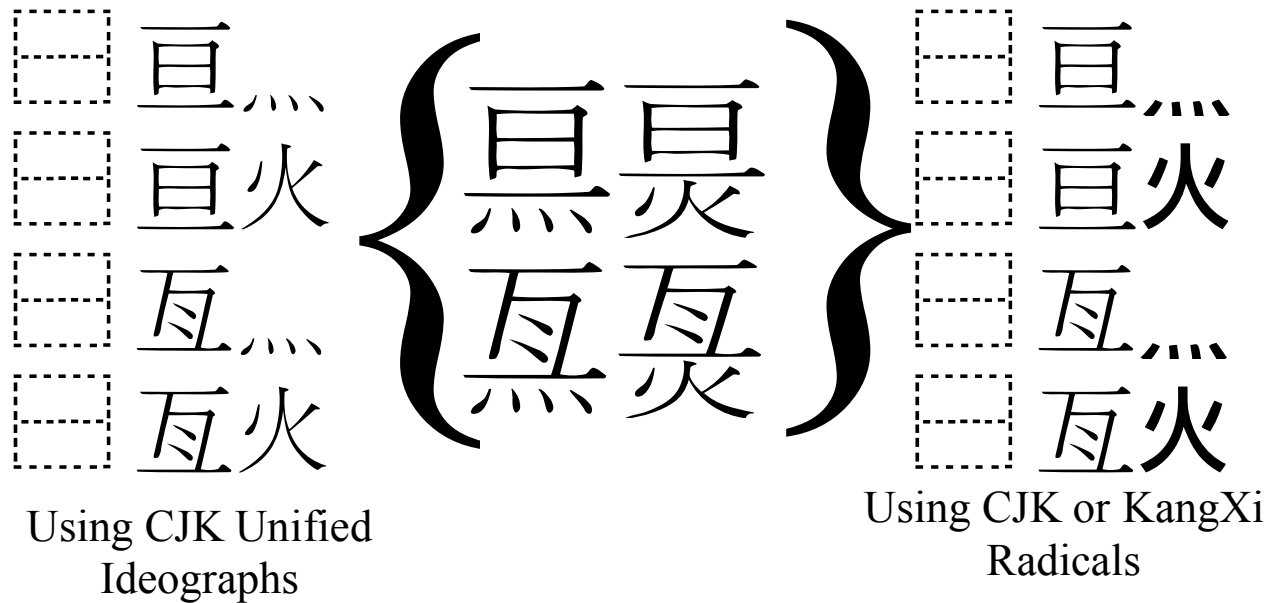
- Two Ideographic Description Sequences are equivalent if and only if they are identical
- "Should" of length rule is unenforceable!

井蛙

≠ 井 虫圭

≠ 井 虫 土土

IDS: Equivalence vs. Variant Glyphs

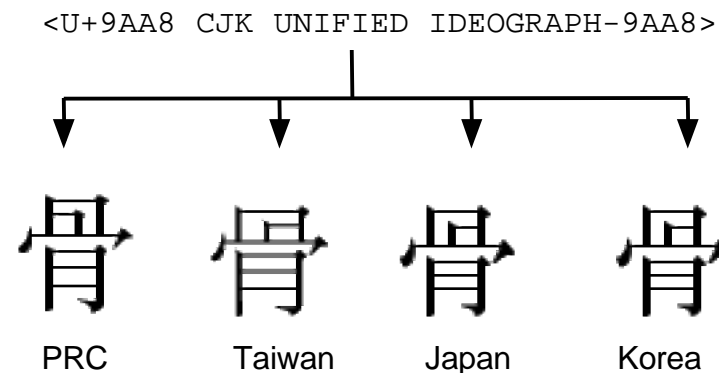


Solving the Variants Problem

- Fundamental problem: Indicating which of multiple glyphic variations of a character should be used
 - Analogous to insisting an “*a*” be used instead of an “*a*”
 - E.g., in pinyin
- Multiple problems
 - Unihan glyph selection
 - Well-known variations (primarily personal names)
 - Newly coined variations (primarily personal names)
- Severity of the problem is unclear

Solutions to the Glyph Selection Problem

- Restrict use to an appropriate locale-specific font
- Use glyph shape selection features in OpenType or Apple Advanced Typography (AAT)



Solutions to the Variants Problem

- Out-of-band information to select Unihan glyphs
 - Plane-14 language tags
 - Other out-of-band information
- Variation tags
 - Research is ongoing at Apple and elsewhere
 - Needs a registry of variants

Conclusion

- Unicode provides one of the most extensive standard ideograph sets available
- Planned enhancements to the standard will further extend its ability to solve the problems of user communities
 - Governments and other agencies that keep track of people's names
 - Scholars and others interested in rare, obscure, and nonce forms