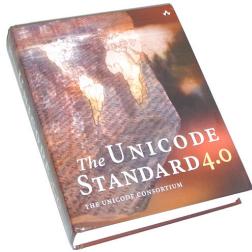


# The Multilingual Lion: T<sub>E</sub>X learns to speak Unicode



Jonathan Kew  
SIL International

April 7, 2005



## Background

- TEX: free typesetting system with a 25-year history
  - stable, reliable, flexible, widely implemented
  - experienced user community
  - rich collection of supporting tools
- Originally designed for English typesetting
  - support for accents and other European characters
  - language support extended via custom fonts, macros, and preprocessors

## Traditional T<sub>E</sub>X input conventions

- Input text is ASCII (or 8-bit codepage)

<i>Source text</i>	<i>Typeset output</i>	<i>Notes</i>
<code>\'{a}</code>	á	<i>typical accent command</i>
<code>\c{c}</code>	ç	
<code>\aa</code>	å	
<code>---</code>	—	<i>ligature in typical T<sub>E</sub>X fonts</i>
<code>\$\alpha\$</code>	α	<i>math mode symbol</i>
<code>\dn acchaa}</code>	অঞ্চাট	<i>using custom preprocessor</i>

## Multilingual typesetting with TEX

- Text input
  - Escape sequences for non-ASCII characters
  - Multiple 8-bit codepages
  - Preprocessors for complex scripts
- Font support
  - Fonts limited to 256 glyphs
  - Custom-encoded fonts with specific glyph sets
- All tied together via complex TEX macros
  - Difficult to understand and extend
  - Difficult to integrate with other packages

## Towards a cleaner solution

- Unicode: all required characters directly represented
  - no need for “escape sequences” to access characters not included in the current codepage
  - no need to switch between codepages according to the language/script being typeset
  - characters rendered via standard access codes
- Character/glyph model and modern font rendering technologies
  - complex script handling moved out of the domain of the text data stream

## Typesetting Unicode text with X<sub>E</sub>T<sub>E</sub>X

- Accented characters

```
\halign{#\hfil\quad&
      #\hfil\cr
dan&   dan\cr
dubok& dubok\cr
džabe& đak\cr
džin&  džabe\cr
Džin&  džin\cr
đak&   Džin\cr
Evropa& Evropa\cr}
```

dan	dubok	džabe	džin	Džin	đak	Evropa
dan	dubok	đak	džabe	džin	džin	Evropa

## Typesetting Unicode text with XeTeX

- CJK ideographs

```
\font\han="STSong" at 16pt
\font\rom="Gentium" at 8pt
\def\hc#1#2{\vtop{\hbox{\han #1}
\hbox{\kern10pt\rom #2}}}
\vtop{\hc{書<}{ka-ku}}
\hc{最も}{motto-mo}
\hc{最後}{sai-go}
\hc{働く}{hatara-ku}
\hc{海}{umi}}
```

書 <  
ka-ku  
最 も  
motto-mo  
最 後  
sai-go  
働 <  
hatara-ku  
海  
umi

## Typesetting Unicode text with XETEX

- Complex scripts

\c ١

شئادیپ یج ایند \s

\p

ء نیمز ادخ ۾ تاعورش ۱

. ويک اديپ يك نامس آ

بیترتیب نیمز تقو نا ۲

دنمس یهنا . یئه ناري و ئ

وه ليکيد ناس هدنوا ورچاتم جو

ادخ ناثم یج گئٹاپ ئ

يک یئپ اريق حور جي

ینشور ”ـت ونـڈ مـکـح اـدـخ نـهـذـت ۳

يئيپ يـت ینـشـور وـس ”ـ. ـيـئـثـ

دنيا جي پيدائش

١ شروعات ۾ خدا زمين ۽ آسمان کي  
پيدا ڪيو. ٢ ان وقت زمين بي ترتيب ۽  
ويران هئي. اونهي سمند جو مٿاچرو اوندهه  
سان ڏکيل هو ۽ پاڌئي جي مٿان خدا جي  
روح ڦيرا پئي ڪي ٣ تـهـنـ خـداـ حـڪـمـ ڏـنوـ  
ته ”روشنـيـ ٿـئـيـ.“ سـوـ روـشـنـيـ ٿـئـيـ.

## Key changes from T<sub>E</sub>X to X<sub>E</sub>T<sub>E</sub>X

- Unicode as the text encoding
  - directly use Unicode input text, Unicode-encoded fonts
- Fonts and rendering technologies
  - use any fonts available in the host computer
  - use existing smart-font rendering systems
- Additional features for multilingual typesetting
  - optional font features
  - line breaking for Asian scripts
- Backward compatibility issues
  - support for legacy T<sub>E</sub>X fonts and documents

## From 8 to 16 bits...

- Character type in T<sub>E</sub>X code was 8-bit value
  - one option: process text as UTF-8
- Character codes used to index a number of tables
  - character category, case pairs, etc.
- Decision to use 16-bit character codes
  - all 256-element tables enlarged to 65,536 elements to match the extended character set
  - extended T<sub>E</sub>X commands that refer to character codes

## From 8 to 16 bits... and beyond?

- Unicode does not fit in 16 bits either!
- X<sub>E</sub>T<sub>E</sub>X handles non-BMP characters as UTF-16 surrogate pairs
  - properties of individual characters cannot be set
  - unlikely to matter for typesetting usage: all surrogate codes can be treated as simple printable characters
  - keeps size of internal tables moderate, without extensive restructuring
- Using UTF-16 happens to match the font rendering APIs that X<sub>E</sub>T<sub>E</sub>X uses

## Implementing the character/glyph model

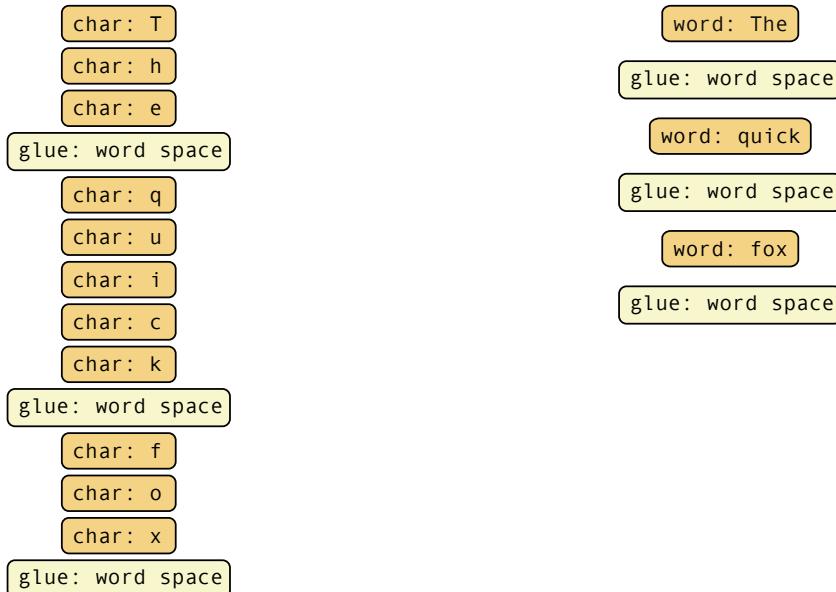
- Required for support of complex scripts in Unicode
- Significant change from traditional T<sub>E</sub>X model
  - T<sub>E</sub>X regards “a specific character code in a specific font” as the fundamental unit of text to be typeset
  - assumes such a character has known, fixed dimensions
  - provision for ligatures by character substitutions
  - a paragraph consists of sequence of “character” nodes, to be precisely placed, and intervening “glue” nodes
- A Unicode character may not map to a single, known glyph
  - many scripts require contextual selection of glyphs
  - must measure characters in context, not in isolation

## Implementing the character/glyph model

- Initial implementation using ATSUI on Mac OS X
  - typesetting process collects runs of characters (words)
  - calls ATSUI text layout APIs to measure width
  - a Xe<sup>T</sup>E<sub>X</sub> paragraph consists of sequence of “word” nodes separated by “glue”
- Typesetting engine positions words, not glyphs
  - this is the job of the font rendering engine

## Implementing the character/glyph model

Nodes in a T<sub>E</sub>X paragraph      Corresponding nodes in X<sub>E</sub>T<sub>E</sub>X



## Implementing the character/glyph model

- OpenType Layout support using ICU library
  - alternative font layout engine
  - provides support for OpenType features in Latin fonts
  - supports a number of complex (Indic/Asian) scripts
- X<sub>E</sub>T<sub>E</sub>X uses either ATSUI or ICU according to layout tables found in fonts
  - overall typesetting process is independent of font technology in use
  - distinction required only at lowest level of measuring a run of text in a given font
  - documents may freely mix AAT and OT fonts

## Implementing the character/glyph model

- ATSUI APIs used in typesetting
  - `ATSUCreateStyle`, `ATSUSetAttributes`
  - `ATSUCreateTextLayout`, `ATSUSetTextPointerLocation`,  
`ATSUSetRunStyle`
  - `ATSUGetUnjustifiedBounds`, `ATSUDrawText`
- ICU APIs used in typesetting
  - `ubidi_open`, `ubidi_close`, `ubidi_setPara`,  
`ubidi_getDirection`, `ubidi_countRuns`,  
`ubidi_getVisualRun`
  - `LayoutEngine::layoutChars`, `getGlyphs`,  
`getGlyphPositions`

## Hyphenation support

- Paragraphs formed of lists of “word boxes”
  - treated as indivisible units in the token list
  - allows T<sub>E</sub>X to remain unaware of low-level details
- If acceptable line breaks not found, hyphenation required
  - extract text characters from word nodes
  - find hyphen positions using T<sub>E</sub>X’s algorithm
  - repackage words as word fragments and discretionary break nodes

## Hyphenation support

- Modifying the node list to allow hyphenation

Two glue different glue foxes

Two glue dif hyphen? fer hyphen? ent glue foxes

- Problem: unused hyphen points break rendering

Two glue dif - fer -

ent glue foxes

*Two differ-  
ent foxes*

- Need to re-merge word nodes after choosing breaks

Two glue differ-

ent glue foxes

*Two differ-  
ent foxes*

## Advanced font features

- OpenType language systems

```
\font\Doulos="Doulos SIL/ICU"
```

```
\font\DoulosViet="Doulos SIL/ICU:language=VIT"
```

Unicode cung cấp

một con số duy

nhất cho mỗi ký tự

Unicode cung cấp

một con số duy

nhất cho mỗi ký tự

```
\font\Brioso="Brioso Pro"
```

```
\font\BriosoTrk="Brioso Pro:language=TRK"
```

... gelen firmaları

... tarafından ...

... gelen firmaları

... tarafından ...

## Advanced font features

- Custom AAT features

```
\font\Doulos="Doulos SIL/AAT"
```

```
\font\DoulosAlt="Doulos SIL/AAT:
```

Alternate forms=Literacy alternates,

Small v-hook straight style;

Uppercase Eng alternates=Capital N with tail"

Xəsee na Mose dō  
Njutitotoŋkeke la anyi,  
eye wònà wohlē vu qe  
vɔtrutiwo ɲu bene dɔla  
si atsrɔ̄ ɳgɔgbəviwo la  
nagawɔ̄ nuvevi Israel  
viwo ya o.

Xəsee na Mose dō  
Njutitotoŋkeke la anyi,  
eye wònà wohlē vu qe  
vɔtrutiwo ɲu bene dɔla  
si atsrɔ̄ ɳgɔgbəviwo la  
nagawɔ̄ nuvevi Israel  
viwo ya o.

## East Asian languages

- Line breaking without word spaces
  - T<sub>E</sub>X normally breaks lines at “glue” arising from spaces
  - Chinese, Japanese, Thai, etc. do not use word spaces
  - โดยพื้นฐานแล้ว, คอมพิวเตอร์จะเกี่ยวข้องกับเรื่องของตัวเลข. คอมพิวเตอร์จัดเก็บโดยการกำหนดหมายเลขให้สำหรับแต่ละตัว. ก่อนหน้าที่ Unicode จะถูกสร้างขึ้น, ได้มีระบบ encoding อญ্ত์หลายร้อยระบบสำหรับการกำหนดหมายเลขเหล่านี้.
- Use ICU line-break: \XeTeXlinebreaklocale "th"
  - โดยพื้นฐานแล้ว, คอมพิวเตอร์จะเกี่ยวข้องกับเรื่องของตัวเลข. คอมพิวเตอร์จัดเก็บตัวอักษรและอักษรระเอื่นๆ โดยการกำหนดหมายเลขให้สำหรับแต่ละตัว. ก่อนหน้าที่ Unicode จะถูกสร้างขึ้น, ได้มีระบบ encoding อญ্ত์หลายร้อยระบบสำหรับการกำหนดหมายเลขเหล่านี้.

## Backward compatibility

- Legacy T<sub>E</sub>X fonts, especially for math mode
  - supported via T<sub>E</sub>X font metrics and Type 1 font files
  - allow many existing T<sub>E</sub>X documents to work
  - not Unicode-compliant!

$$\begin{aligned} \left( \int_{-\infty}^{\infty} e^{-x^2} dx \right)^2 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} dx dy \\ &= \int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr d\theta \\ &= \int_0^{2\pi} \left( -\frac{e^{-r^2}}{2} \Big|_{r=0}^{r=\infty} \right) d\theta \\ &= \pi. \end{aligned}$$

## Backward compatibility

- Non-Unicode input text
  - by default, input read as Unicode (UTF-8 or UTF-16)
  - legacy codepages supported via ICU converters
  - set codepage of current input file:  
`\XeTeXinputencoding "charset-name"`
  - set initial codepage for newly-opened input files:  
`\XeTeXdefaultencoding "charset-name"`

## Backward compatibility

- Support for legacy keying practices
  - typical input:  
` ``\TeX' '---a typesetting system
  - generates: ``**\TeX**''---a typesetting system
- Font mapping for compatibility

; TECKit mapping for TeX input conventions

U+002D U+002D <> U+2013 ; -- -> en dash

U+002D U+002D U+002D <> U+2014 ; --- -> em dash

U+0027 <> U+2019 ; ' -> right single quote

U+0027 U+0027 <> U+201D ; '' -> right double quote

U+0022 > U+201D ; " -> right double quote

- generates: “**\TeX**”—a typesetting system

## More fun with font mappings

```
\def\SampleText{Unicode -  
    это уникальный  
    код для любого символа, \\  
    независимо от платформы, \\  
    независимо от программы, \\  
    независимо от языка.}  
\font\gen="Gentium"  
\gen\SampleText  
\bigskip  
\font\gentrans="Gentium:  
    mapping=cyr-lat-iso9"  
\gentrans\SampleText
```

Unicode - это уникальный  
код для любого символа,  
независимо от платформы,  
независимо от программы,  
независимо от языка.

Unicode - èto unikal'nyj  
kod dlâ lûbogo simvola,  
nezavisimo ot platformy,  
nezavisimo ot programmy,  
nezavisimo ot âzyka.

## X<sub>E</sub>T<sub>E</sub>X and other T<sub>E</sub>X extensions

- T<sub>E</sub>X<sub>G</sub>X
  - a direct ancestor of X<sub>E</sub>T<sub>E</sub>X, but now obsolete
- e-T<sub>E</sub>X
  - basis of current X<sub>E</sub>T<sub>E</sub>X implementation
  - provides a number of features, especially bidi support
- Omega, Aleph
  - ambitious project to extend T<sub>E</sub>X to all scripts
  - complex configuration, no direct smart-font support
- pdfT<sub>E</sub>X
  - widely-used extension providing rich PDF support
  - no native Unicode or smart-font support

# The Multilingual Lion: T<sub>E</sub>X learns to speak Unicode

## For more information

- X<sub>E</sub>T<sub>E</sub>X web site and mailing list
    - <http://scripts.sil.org/xetex>
    - <http://tug.org/mailman/listinfo/xetex>
  - Contact information
    - [mailto:jonathan\\_kew@sil.org](mailto:jonathan_kew@sil.org)
  - Questions... and answers?

፩፻፲፭ ምንም እውነት የዚህ ቀን አንድ ነው? የዚህ ቀን አንድ ነው? ما هي الشفرة الموحدة "يونيكود"؟ 什麼是Unicode (統一碼/標準萬國碼)? Što je Unicode? რა არის უნიკოდი? Tí eǐnai tò Unicode; ? מה זה יוניקוד? چیست یونیکوڈ کیا ہے؟ Hvað er Unicode? ユニコードとは何か？ 유니코드에 대해？ چیست？ Что такое Unicode？ Unicode គឺមីនៅទេ？ ፩፻፲፭ አንድ ነው?

