What is Unicode? Translation

Behdad Esfahbod
unicode@behdad.org

دانشکدهی مهندسی کامپیوتر
دانشگاه صنعتی شریف
تهران، ایران
اولین چهارشنبهی آبان ماه هر سال
There were some slides from old presentations
There were some slides from old presentations

It was boring to redesign new ones
So, I turned to Technical Details here...
What is Unicode?

It's all about characters and their character codes.

Unicode uses a 20-bit character set!

U+06CC is the ARABIC LETTER FARSIYEH (٢)

Characters are arranged in blocks so one can find them easily (all Arabic letters are in range U+0600 and U+06FF).
What is Unicode?

- It's all about characters

U+06CC is the ARABIC LETTER Farsi Yeh (Ù)
What is Unicode?

- It's all about characters

- $0 \leq \text{character code} \leq 10FFFF_{16} = 17 \times 2^{16} - 1 = 1,114,111$
What is Unicode?

- It's all about characters

- $0 \leq \text{character code} \leq 10FFFF_{16} = 17 \times 2^{16} - 1 = 1,114,111$

- The minimum number of bits enough for encoding every character is 21, but that's almost nowhere used. (It's actually a 20.08746-bit character set!)
What is Unicode?

- It's all about characters

- $0 \leq \text{character code} \leq 10FFFF_{16} = 17 \times 2^{16} - 1 = 1,114,111$

- The minimum number of bits enough for encoding every character is 21, but that's almost nowhere used. (It's actually a 20.08746-bit character set!)

- U+06CC is the ARABIC LETTER FARSI YEH (ي)
What is Unicode?

- It's all about characters

- \[ 0 \leq \text{character code} \leq 10FFFF_{16} = 17 \times 2^{16} - 1 = 1,114,111 \]

- The minimum number of bits enough for encoding every character is 21, but that's almost nowhere used. (It's actually a 20.08746-bit character set!)

- U+06CC is the ARABIC LETTER Farsi YEH (ی)

- Characters are arranged in blocks so one can find them easily (all Arabic letters are in range U+0600 and U+06FF)
But what is a character?

We don't know!
But what is a character?

We don't know!
... but we know some things that are not characters:
But what is a character?

We don't know!
... but we know some things that are not characters:

- Glyphs: there is only one ARABIC LETTER BEH (ب)
But what is a character?

We don't know!

... but we know some things that are not characters:

- Glyphs: there is only one ARABIC LETTER BEH (ب)
- Ligatures: there is no ARABIC LIGATURE LAM ALEF (لا)
But what is a character?

We don't know!

... but we know some things that are not characters:

- Glyphs: there is only one ARABIC LETTER BEH (ب)
- Ligatures: there is no ARABIC LIGATURE LAM ALEF (ﻻ)
- Markup: there is no START BOLDFACE (＜b＞)
But what is a character?

We don't know!

... but we know some things that are not characters:

- **Glyphs**: there is only one ARABIC LETTER BEH (ب)
- **Ligatures**: there is no ARABIC LIGATURE LAM ALEF (لا)
- **Markup**: there is no START BOLDFACE (\<b\>)
- **Logos and emblems**: there is no APPLE SIGN
But what is a character?

We don't know!

... but we know some things that are not characters:

- Glyphs: there is only one ARABIC LETTER BEH (ب)
- Ligatures: there is no ARABIC LIGATURE LAM ALEF (لا)
- Markup: there is no START BOLDFACE (<b>)
- Logos and emblems: there is no APPLE SIGN

So, did you found out what's it?
Glyphs: there are four different presentation forms of ARABIC LETTER BEH (پ، ب، پ، ب)، in addition to one general one, but...
Thats a big lie!

- Glyphs: there are four different presentation forms of ARABIC LETTER BEH (ب، ب، ب، ب)، in addition to one general one, but...

- Ligatures: there is an ARABIC LIGATURE LAM ALEF (لا)، among many others
Glyphs: there are four different presentation forms of ARABIC LETTER BEH (ب، ب، ب)， in addition to one general one, but...

Ligatures: there is an ARABIC LIGATURE LAM ALEF (ل، ل)، among many others

Markup: there are control character everywhere, from a PARAGRAPH SEPARATOR to something named POP DIRECTIONAL FORMATTING.
Thats a big lie!

- Glyphs: there are four different presentation forms of ARABIC LETTER BEH (ب، ب، ب، ب)، in addition to one general one, but...

- Ligatures: there is an ARABIC LIGATURE LAM ALEF (ل), among many others

- Markup: there are control character everywhere, from a PARAGRAPH SEPARATOR to something named POP DIRECTIONAL FORMATTING

- Logos and Emblems: Farsi Symbol (U+262B) is there, as well as playing cards suits.
Not just codes, names or shapes
Not just codes, names or shapes

- Several informative or normative properties and descriptions are available to disambiguify the characters:

  *general category, combining class, bidirectional category, decomposition mapping, numeric value, mirroring property, case mappings, joining class and group, line breaking property, ...*
Some character properties

Decomposition, recomposition, reordering, equivalence, and normalization: to make sure that you and I encode the same string the same way.

Bidirectional properties and behavior: to make sure logically encoded bidirectional scripts are not displayed differently on my computer than on yours.
Some character properties

- Decomposition, recomposition, reordering, equivalence, and normalization: to make sure that me and you encode the same string the same way.
Some character properties

- **Decomposition, recomposition, reordering, equivalence, and normalization**: to make sure that me and you encode the same string the same way.

- **Bidirectional properties and behavior**: to make sure logically encoded bidirectional scripts are not displayed differently on my computer than yours.
Bidirectional Algorithm

Providing an *exact* and *implicit* mechanism for converting a logically stored stream of characters including some characters of a right-to-left script, to a visually ordered one suitable for display.
Bidirectional Algorithm

Providing an exact and implicit mechanism for converting a logically stored stream of characters including some characters of a right-to-left script, to a visually ordered one suitable for display.

This is needed for Arabic (incl. Persian, Urdu, Sindhi, ...), Hebrew (incl. Yiddish), Syriac, and Thanaa.
Bidirectional Algorithm

Providing an exact and implicit mechanism for converting a logically stored stream of characters including some characters of a right-to-left script, to a visually ordered one suitable for display.

This is needed for Arabic (incl. Persian, Urdu, Sindhi, ...) Hebrew (incl. Yiddish), Syriac, and Thanaa.

A car is called THE CAR in Hebrew

↓

A car is called THE CAR in Hebrew
Bidirectional Algorithm (continued)

Many implicit and explicit *bidirectional categories*:

left-to-right, right-to-left, right-to-left Arabic, European number, Arabic number, European number separator, European number terminator, common number separator, non-spacing mark, boundary neutral, paragraph separator, segment separator, whitespace, other neutrals, *left-to-right embedding*, *right-to-left embedding*, *left-to-right override*, *right-to-left override*, *pop directional format*
A few interesting features (continued)
A few interesting features (continued)

- Line breaking properties
A few interesting features (continued)

- Line breaking properties
- Mirroring characters
A few interesting features (continued)

- Line breaking properties

- Mirroring characters

- All characters and symbols needed for mathematical typesetting (thanks to AMS)
With enough care, the above algorithms can be applied in some different order.
Joining and Shaping Algorithms

Two adjacent letters may join to each other, or may not, forming 1, 2, or 4 glyphs for each character (for example ‘•’, ‘”’). The Joining Algorithm is for deciding if two adjacent letters do join or not. The Shaping Algorithm is for selecting the proper glyph, based on the results of the Joining Algorithm.
Joining and Shaping Algorithms

- Two adjacent letters may *join* to each other, or may not...
Joining and Shaping Algorithms

- Two adjacent letters may *join* to each other, or may not
- ... forming 1, 2, or 4 glyphs for each character (for example ﺖ، ﺪ، ﺪ، ﺳ)
Joining and Shaping Algorithms

- Two adjacent letters may *join* to each other, or may not

- ... forming 1, 2, or 4 glyphs for each character (for example ﺤ، ﺤ، ﺤ، ﺤ)

- The Joining Algorithm is for deciding if two adjacent letters do join or not
Joining and Shaping Algorithms

- Two adjacent letters may join to each other, or may not
- ... forming 1, 2, or 4 glyphs for each character (for example س س س س)
- The Joining Algorithm is for deciding if two adjacent letters do join or not
- The Shaping Algorithm is for selecting the proper glyph, based on the results of the Joining Algorithm
Unicode Transformation Formats

These are just *encodings*...
Unicode Transformation Formats

These are just *encodings*...

- UTF-8 for 8-bit environments
Unicode Transformation Formats

These are just *encodings*...

- UTF-8 for 8-bit environments
- UTF-16 for 16-bit environments
Unicode Transformation Formats

These are just *encodings*…

- UTF-8 for 8-bit environments
- UTF-16 for 16-bit environments
- UTF-32 for 32-bit environments

… best one depends on the environment
UTF-32
UTF-32

- The identity mapping to Unicode values
The identity mapping to Unicode values

Easiest to process
UTF-32

- The identity mapping to Unicode values
- Easiest to process
- Most storage
UTF-32

- The identity mapping to Unicode values
- Easiest to process
- Most storage
- Four times in size for ASCII text
The identity mapping for BMP, so misleading for novice developers.

Good compromise on ease and reasonable storage.

Widely used on Microsoft platforms.
The identity mapping for BMP, so misleading for novice developers
UTF-16

- The identity mapping for BMP, so misleading for novice developers
- Good compromise on ease
UTF-16

- The identity mapping for BMP, so misleading for novice developers
- Good compromise on ease
- Reasonable storage
UTF-16

- The identity mapping for BMP, so misleading for novice developers
- Good compromise on ease
- Reasonable storage
- Widely used on MicroSoft platforms
Upward compatible with ASCII
Designed for replacing ASCII transparently
Least storage, still simple
Most recommended and actually in use
Great fun to learn
• Upward compatible with ASCII
UTF-8

- Upward compatible with ASCII
- Designed for replacing ASCII transparently
• Upward compatible with ASCII

• Designed for replacing ASCII transparently

• Least storage, still simple
UTF-8

- Upward compatible with ASCII
- Designed for replacing ASCII transparently
- Least storage, still simple
- Most recommended and actually in use
• Upward compatible with ASCII

• Designed for replacing ASCII transparently

• Least storage, still simple

• Most recommended and actually in use

• Great fun to learn
Frequently Asked/Answered Questions

What is crap with these fonts?

What is the problem with FARSIYEH (۷۸۸۷)?

What is the problem with HEHWITHYEH ABOVE (ی)?

I cannot type my name (ب), where are the letters?

Is there any good fonts around?
Frequently Asked/Answered Questions

- What is all crap with these fonts?

- What is the problem with FarsiYeh («نژ» ᵁ)?

- What is the problem with Hebrew with Yeh above (י)?
Frequently Asked/Answered Questions

- What is all crap with these fonts?
- What is the problem with FARSI YEH (کارت گرافیکی)?
Frequently Asked/Answered Questions

- What is all crap with these fonts?
- What is the problem with Farsi Yeh (کارت گرافی کی)؟
- What is the problem with HEH WITH YEH ABOVE (ة)؟
Frequently Asked/Answered Questions

- What is all crap with these fonts?
- What is the problem with FARSI YEH (کارت گرافی کی)?
- What is the problem with HEH WITH YEH ABOVE (ه)?
- I cannot type my name پژن, where are the letters?
Frequently Asked/Answered Questions

- What is all crap with these fonts?
- What is the problem with Farsi Yeh (کارت گرافیکی)؟
- What is the problem with Heh with Yeh above (ة)؟
- I cannot type my name پوژن, where are the letters?
- Is there any good fonts around?
ISIRI 2901
Standard keyboard layout

First released in 1989, and revised in 1994. All characters are accessible at most by shift key. Should be revised to reflect the new standard. Drivers available for Windows 2000/XP, also all Linux environments.

After you learn, you will never switch.
ISIRI 2901
Standard keyboard layout

- First released in 1989, and revised in 1994
ISIRI 2901
Standard keyboard layout

- First released in 1989, and revised in 1994
- All characters are accessible at most by shift key
ISIRI 2901
Standard keyboard layout

- First released in 1989, and revised in 1994
- All characters are accessible at most by shift key
- Should be revised to reflect the new standard
ISIRI 2901
Standard keyboard layout

- First released in 1989, and revised in 1994
- All characters are accessible at most by shift key
- Should be revised to reflect the new standard
- Drivers available for Windows 2000/XP, also all Linux environments
ISIRI 2901
Standard keyboard layout

- First released in 1989, and revised in 1994
- All characters are accessible at most by shift key
- Should be revised to reflect the new standard
- Drivers available for Windows 2000/XP, also all Linux environments
- After you learn, you will never switch
Thats enough, lets talk about Online Games
John Carmack

The Prophet of Online Gaming
John Carmack, the commander keen
John Carmack, the commander keen

- 1990, just another 19 year old geek
John Carmack, the commander keen

- 1990, just another 19 year old geek
- Working at Softdisk Publishing
John Carmack, the commander keen

- 1990, just another 19 year old geek
- Working at Softdisk Publishing
- Starts to write *Commander Keen*, the EGA side-scroller game, with his team
John Carmack, the commander keen

- 1990, just another 19 year old geek
- Working at Softdisk Publishing
- Starts to write *Commander Keen*, the EGA side-scroller game, with his team
- The game is a huge success, so his team leave...
The more keens
The more keens

- ... to found id Software on February 1, 1991
The more keens

- ... to found id Software on February 1, 1991
- Writes several more Commander Keen games
• ... to found id Software on February 1, 1991

• Writes several more Commander Keen games

• After months of hard work, in May 1992, the first 3d game is born
Remember Nazy SS symbol?
Remember Nazi SS symbol?

- ... the good old *Wolfenstein 3d*
Remember Nazy SS symbol?

- ... the good old *Wolfenstein 3d*

- The very next year the great *Doom* is born
Remember Nazi SS symbol?

- ... the good old *Wolfenstein 3d*
- The very next year the great *Doom* is born
- It was a technical and creative milestone
Remember Nazi SS symbol?

- ... the good old *Wolfenstein 3d*
- The very next year the great *Doom* is born
- It was a technical and creative milestone
- Significantly raised the standards for game creators
The revolutionary step forward
The revolutionary step forward

- With Doom you could play through your modem
The revolutionary step forward

- With Doom you could play through your modem
- Better still, on a LAN with up to 8 people
The revolutionary step forward

- With Doom you could play through your modem
- Better still, on a LAN with up to 8 people
- No force to play against mindless computer opponents
The sequel to the Doom series
The sequel to the Doom series

- In 1996, Carmack creates Quake
The sequel to the Doom series

- In 1996, Carmack creates Quake
- Once again, the technology is completely new and totally astonishing
The sequel to the Doom series

- In 1996, Carmack creates *Quake*

- Once again, the technology is completely new and totally astonishing

- The first truly three-dimensional environment game
The sequel to the Doom series

- In 1996, Carmack creates Quake
- Once again, the technology is completely new and totally astonishing
- The first truly three-dimensional environment game
- Completely supporting internet play online
The sequel to the Doom series

- In 1996, Carmack creates *Quake*

- Once again, the technology is completely new and totally astonishing

- The first truly three-dimensional environment game

- Completely supporting internet play online

- This was the defining moment in first-person online gaming
Quake Clans

From the first week, people began to form teams. Several leagues were formed for expert players. The idea of a LAN party was formed. Within a year, the internet was a changed place.
Quake Clans

- From the first week, people began to form teams
Quake Clans

- From the first week, people began to form teams
- Several leagues were formed for expert players
Quake Clans

- From the first week, people began to form teams
- Several leagues were formed for expert players
- The idea of a LAN party was formed
Quake Clans

- From the first week, people began to form teams
- Several leagues were formed for expert players
- The idea of a LAN party was formed
- Within a year, the internet was a changed place
Quake III Arena

was the last big step
Doom III

is the next
Have a Look Yourself

THE END
Oh! Wait

And this f*cking bastard is just 31*

*see page number