## И S G C P C T C S И S П Ц И Я, S A III SIDE II SIDE II Sh

ى لليها بألغة العربية (العالم العربي)

- Technical background of i18n for software and Debian packages -

Enialte esonvava ouve:

בוור כדי לחווטיד בעבריו

Kenshi Muto <a href="mailto:kmuto@debian.org">kmuto@debian.org</a>

Fingerprint: 360B B879 6957 5D6F EED3 D986 40A5 BEED 72D0 3CB1

110n/i18n talk session with Christian Perrier, debconf4 at POA.

## 

ى لليها بألغة العربية (العالم العربي)

- Following Christian's talk,
- How to make your packages better about internationalization and localization
- Abbreviations
- Fi18n: Internationalization

  18characters
- > I10n : Localization
  10characters

ى لليعالم بألفة العربية (العالم العربي) И S G C P C T C T A S И O П Ц И Я, S A Thoose this to proceed in English thoisissez ceci pour continuer en français Eniale te solo v<u>i</u>o vo <u>o</u>ouve: letuvikkai tesite diegimą, iksirinkę kią selecione esta para prosseguir em l'ortugué YTOOLINPOA,OAKATЬ PAGOT! lürkçe olarak devam etmek için seçiniz виберіть цей рядок, щобі then cai này để tiếp tạc bảng tiếng Việt

### There are TWO Word-Processors here...



Menu is localized, Content is NOT localized

TOO DEIN NORMONDE ON THE SECOND

Menu is NOT localized, Content is localized

# Needs to understand somewhat...

> To understand a basic knowledge about

En il 18n/110ns sooy y lo vo ouvs;

国本語で鼓煙を行うにはこれを近んでください Listuviëkai tesite diegimą, iësirinke ëiq Belegione osta para prosseguir em Portugut Чтобы продолжать работ:

lurkoe olarak devem etmek ioin seoiniz Виберіть цейрядок, щоб

thọn cái này để tiếp tạo bằng tiếng Việt

## 1800 POTCLING MONUMENT SA

ى لليها بألفة العربية (العالم العربي)

- > Many many languages exist in the world
- Latin (English, French, German, ...)
- > BIDI (Arabic, Hebrew)
  - > CJK (Chinese, Japanese, Korean) 歡迎光臨

lurkoe olarak devam etmek ioin seoiniz

виберіть цей рядок, щоб

Mon cái này để tiếp tực bảng tiếng Việt

> More (Thai, Hindi, ...) 대비만이란

ما هو ديبيان

## Character set and Encoding

A = 0x41 (ASCII, IS08859-1) € = 0xa4 (IS08859-15)

あ = 0xa4 0xa2 (EUC-JP) 0x82 0xa0 (Shift\_JIS)

- Character set
  - > set of acceptable characters for each languages (ASCII, JISX0208, ...)
- Encoding (Encoded character set)
  - > map character set and ID number for computer
- Many encodings...
  - ISO-8859-1 (Latin-1), ISO-8859-15 (Euro),
     ISO-2022-JP, EUC-JP, ShiftJIS, Big5,
     GB2312, KOI8-R, ....

## Mojibake

ى لليها بألغة العربية (المالم العربي)

- > What's "mojibake"? "文字化け"
  - > Broken screen (we can't read characters)
    - It's far from English developpers, but we
       (Japanese) meet very often
- lett Why? i tesite dieging, issirinke sig
- Mismatch encoding or mismatch font

thon cái này để tiếp tực bảng tiếng Việt

- Screen problem Screen problem
  - > Bad toolkit or bad design

ى للبعدا بألفة العربية (العالم العربي) NSOCPOTE TASMONUMA, SA 計能器は関熱以続に中支性流道術 shoose this to proceed in English hoisise How toumake your ansals En 1 Act to 8 60 y 10 y 0 0 0 0 v 8 17日12 Typackage more 日本語で設定を行うにはこれを比んでください Lighty iskai Id On and id 18n? jkg 8ig selecione esta para prosseguir em l'ortugué YTOGEINPOA,OAKATE PŁGOT! lurkoe olarak devam etmek için seçiniz виберіть цей рядок, щоб then cai này để tiếp tực bảng tiếng Việt

msgid "Debian installer main menu" (message ID)

#### gettext

msgstr "Debianインストーラメインメニュー" (Japanese)

Message catalog database

msgstr "Menu principal de l'installateur Debian" (French)

Switch messages using LANG<sub>m</sub> environment variable in

msgstr "Menu principal do instalador Debian" (Brazil Portuguese)

- > Relates msgid and localized message
- many packages use this for 110n (such as debconf messages)
- > core architecutre of debian-installer 110n
- Many bindings
  - > C, shell, Perl, Python, Ruby, Java, ...

#### NSGODOTE TABUOTUMA, SA 注述提供以採**Settext**维行

ى لليها بألئة العربية (العالم العربي)

- Misunderstood implementation...
- Don't use non ASCII characters in msgid
- Use s(n)printf and %s in msgid for dynamic variable

Виберіть цей рядок, щобі Яфп odi này để tiếp tạo báng tiếng Việt

- > Does the msgid really need to make 110n?
  - Gettext isn't the "Silver bullet"...
- Remember Word Processor question

### Toolkit library

ى لليها بألغة العربية (العالم العربي)

- Use i18n ready libraries
- > GTK+ (2.0 is better)
  - > Qt
- > Don't set specific font as default
- XLFD: fixed (?)
  - FreeType: serif, sans-serif, or monospace

thon cái này để tiếp tực bảng tiếng Việt

Configurable is better



## NSOCETE TOOIKIT LIBRARY

ى لليها بألغة العربية (العالم العربي)

- house this to proceed in English
  Input method problem of the français
- How to input your local characters?
- > Application on terminal
- > Depends on terminal software
- Application on X Window System
- Only a few of modern Toolkit can handle
- lurkoc, Immodule, XIM cimek join scoiniz виберть цем радок, щоб hon cai nay da tiap tuo bang tiang Viat

### Use internal encoding

ى لليبه أ بألغة العربية (العالم العربي)

- > Before treating a string, unify its encoding
- Use character unit instead of byte unit
- $\rightarrow UCS-4$  あ = 0xa4 0xa2 (EUC-JP)
- > 32bit Unicode b = 0x00003042 (UCS-4)
- GNU libc (iconv(3))
- Wide character (wchar\_t)
- lurkoe, clarak devem etmek ioin seoiniz вибер тъцей радок, щоб hon ambstocws(3) ( iép tao béng tiéng Việt

#### Use internal encoding

لبغا بألغة العربية (العالم العربي)

- Respect user's locale setting. User's locale (encoding) can be gotten by:
  - locale charmap
  - returns such as ANSI\_X3.4-1968, EUC-JP)
    - Use 'LANG=C' or 'LC\_ALL=C' if you want to ignore locale setting
  - > Applications should provide choice for user:

Mon cái này để tiếp tực bảng tiếng Việt

### Conclusion 4 M 7 8 A

لبعلاأ بألغة العربية (المحالم العربي

- There are many languages in the world! Wrong implementation causes "Mojibake".
- Modern Toolkit for X application is recommended.
- Use internal encoding, such as UCS-4 or Wide character.
- Let's try to make your application 110n/i18n ready! More 110n/i18n makes users more happy. :-)
   Happy Hacking!