The Localization (CJK) Challenges and Possibilities in Taiwan

曾政嘉 Cheng-Chia Tseng (zerng07)

zerng07@fedoraproject.org

ROME | 11th October 2017
The Overview of L10N

5 Major aspects

1. Message translation (UI, Help, Website)
2. Text output (display on the screen, print on papers)
3. Text input (type with a keyboard or select with programs)
4. Information processing (supported by the program)
5. Adapt to the local culture (such as the calendar, the cultural difference on color psychology, conventions on icon design, etc.)
The Taiwan community is getting more and more active these years

- I maintain the translation of the UI, Website and LibreOffice Online
- Jeff Huang and I work on the native language website
- Mark Hung works on the CJK support
- Franklin Weng lead the work on LibreOffice training, migration and marketing in Taiwan
Challenges & Possibilities
The translation of LibreOffice Help has been discontinued for some years

<table>
<thead>
<tr>
<th>Name</th>
<th>Progress</th>
<th>Last Activity</th>
<th>Critical</th>
<th>Suggestions</th>
<th>Incomplete</th>
<th>Total</th>
<th>Last updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>LibreOffice 5.3 – Help</td>
<td></td>
<td>Pootle 6 months</td>
<td>7,361</td>
<td>0</td>
<td>92,039</td>
<td>459,736</td>
<td>6 months</td>
</tr>
<tr>
<td>LibreOffice 5.4 – Help</td>
<td></td>
<td></td>
<td>6,772</td>
<td>0</td>
<td>85,561</td>
<td>459,867</td>
<td>2 weeks</td>
</tr>
<tr>
<td>LibreOffice 5.2 – Help</td>
<td></td>
<td></td>
<td>7,774</td>
<td>0</td>
<td>76,227</td>
<td>456,049</td>
<td>1 year</td>
</tr>
<tr>
<td>LibreOffice 5.1 – Help</td>
<td></td>
<td>wck317 1 year</td>
<td>7,516</td>
<td>0</td>
<td>61,403</td>
<td>451,642</td>
<td>1 year</td>
</tr>
<tr>
<td>LibreOffice 5.1 – UI</td>
<td></td>
<td>pesder 6 months</td>
<td>135</td>
<td>4</td>
<td>1,370</td>
<td>101,140</td>
<td>1 year</td>
</tr>
<tr>
<td>LibreOffice 5.2 – UI</td>
<td></td>
<td>pesder 6 months</td>
<td>142</td>
<td>16</td>
<td>1,087</td>
<td>101,144</td>
<td>6 months</td>
</tr>
<tr>
<td>LibreOffice 5.3 – UI</td>
<td></td>
<td>Aaron 1 hour</td>
<td>134</td>
<td>0</td>
<td>951</td>
<td>100,198</td>
<td>6 months</td>
</tr>
<tr>
<td>LibreOffice 5.4 – UI</td>
<td></td>
<td>Cheng-Chia Tseng 40 seconds</td>
<td>126</td>
<td>0</td>
<td>0</td>
<td>96,100</td>
<td>2 weeks</td>
</tr>
<tr>
<td>LibreOffice Online</td>
<td></td>
<td>Cheng-Chia Tseng 36 minutes</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2,768</td>
<td>3 hours</td>
</tr>
<tr>
<td>Terminology</td>
<td></td>
<td>wck317 1 year</td>
<td>0</td>
<td>0</td>
<td>1,119</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Website</td>
<td></td>
<td>Cheng-Chia Tseng 1 year</td>
<td>0</td>
<td>118</td>
<td>0</td>
<td>1,912</td>
<td>1 year</td>
</tr>
<tr>
<td>Impress Remote (iOS)</td>
<td></td>
<td>Cheng-Chia Tseng 3 years</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>346</td>
<td></td>
</tr>
<tr>
<td>Impress Remote (Android)</td>
<td></td>
<td>Cheng-Chia Tseng 4 years</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>259</td>
<td></td>
</tr>
</tbody>
</table>
Minguo Calendar Identification

Minguo calendar support is not intuitive enough

- The ruling government in Taiwan is the Republic of China (ROC) which was founded in 1912 in mainland China.

- 2017 is 民國 106 年 (Minguo year 106). (2017-1912+1)

- In general, people in Taiwan (zh_TW locale) use both systems in living.
  - 2 or 3 digits (such as 99 or 106) for Minguo year
  - 4 digits (such as 2017) for common era year

- The public servants in the government only use Minguo calendar. “106/10/11” for 2017/10/11 => LibreOffice identifies as “0106-10-11”
  https://bugs.documentfoundation.org/show_bug.cgi?id=113184
Workaround: user have to type common era year first then convert
The support of vertical layout of CJK text is a mess (5.3, 5.4)

- vertical glyphs cannot be displayed in the slide show (including Latin text)
- Han glyphs are rendered at a far distance higher from the cursor position

To know more about the details of the CJK vertical layout, please read the Chapter 7: Typography, *CJKV Information Processing, 2nd Edition* by Ken Lunde

https://bugs.documentfoundation.org/show_bug.cgi?id=103729
Vertical Layout of the CJK text

- Here is a screencast video.
Asian Phonetic Guide (Ruby)

- Bopomofo ruby is the way Taiwanese teach children to learn the Mandarin reading of the Han (Chinese) characters

- One Han character may read in more than one pronunciation. Some can read in 6 different pronunciations.

Author: Bobby Tung, CC-by-SA 4.0 International
Asian Phonetic Guide (Ruby)

- Mono Ruby

One or more ruby glyphs serve to annotate only a single base glyph. Used in Chinese, Japanese and Korean text to annotate Han characters.

一見如故
Asian Phonetic Guide (Ruby)

- **Group Ruby** (LibreOffice design takes group ruby in mind)

Ruby glyphs serve to annotate two or more base glyphs. Used in Japanese text to read the kanjis (Han characters).

[Image of Japanese text and Ruby guide interface]

Asian Phonetic Guide (Ruby)

- **The Problem for Chinese**

  LibreOffice separates the phrase automatically for you to annotate, so 一見如故 is treated as one phrase after selection (group ruby).

  **As Bopomofo we have to annotate one by one (mono ruby).**

  [Link](https://bugs.documentfoundation.org/show_bug.cgi?id=113189)
Asian Phonetic Guide (Ruby)

- Horizontal Ruby

Demo by Bobby Tung to demonstrate the HTML5 ruby module implementation, tweaked with OpenType features

Directly typed with LibreOffice Phonetic Guide

Reference: https://bobbytung.github.io/BopomofoLayoutTest/case01/index.html
Author: Bobby Tung, CC-by-SA 4.0 International
Asian Phonetic Guide (Ruby)

- Vertical Ruby

Demo by Bobby Tung to demonstrate the HTML5 ruby module implementation, tweaked with OpenType features

Directly typed with LibreOffice Phonetic Guide

Author: Bobby Tung, CC-by-SA 4.0 International
Based on grid system and multiples between scales

- ZiTien Movable Type Box
Typographic Scale Convention (Japan)

This system was invented in Japan and introduced to Taiwan.

<table>
<thead>
<tr>
<th>Set A</th>
<th>Set B (mostly used scale)</th>
<th>Set C (frequently used)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>初號 Primary ： 42 pt (4) Title</td>
<td></td>
</tr>
<tr>
<td>一號 One ： 27.5 pt (2)</td>
<td>二號 Two ： 21 pt (2) Heading</td>
<td>三號 Three ： 16 pt (2)</td>
</tr>
<tr>
<td>四號 Four ： 13.75 pt (1)</td>
<td>五號 Five ： 10.5 pt (1) Body text</td>
<td>六號 Six ： 8 pt (1)</td>
</tr>
<tr>
<td></td>
<td>七號 Seven ： 5.25 pt (0.5) Ruby</td>
<td>八號 Eight ： 4 pt (0.5)</td>
</tr>
</tbody>
</table>

There are 3 sets of scale. Color in yellow: base factor
Typographic Scale Convention (Japan)

Color in yellow: regularly used in combination

<table>
<thead>
<tr>
<th>Set A</th>
<th>Set B (mostly used scale)</th>
<th>Set C (frequently used)</th>
</tr>
</thead>
<tbody>
<tr>
<td>一號  One  :  27.5 pt</td>
<td>初號 Primary  :  42 pt (4) Title</td>
<td>三號  Three  :  16 pt if treated as 15.75 pt (1.5)</td>
</tr>
<tr>
<td>四號  Four  :  13.75 pt</td>
<td>二號 Two  :  21 pt (2) Heading</td>
<td>五號  Five  :  10.5 pt (1) Body text</td>
</tr>
<tr>
<td></td>
<td>四號  Four  :  13.75 pt</td>
<td>六號  Six  :  8 pt if treated as 7.875 pt (0.75)</td>
</tr>
<tr>
<td></td>
<td>七號 Seven  :  5.25 pt (0.5) Ruby</td>
<td>八號  Eight  :  4 pt Ruby</td>
</tr>
</tbody>
</table>
Typographic Scale Convention (China)

In China, a system comparable to the Japan system developed

<table>
<thead>
<tr>
<th>Set A</th>
<th>Set B (mostly used scale)</th>
<th>Set C (frequently used)</th>
</tr>
</thead>
<tbody>
<tr>
<td>初號 Primary： 42 pt (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>小初號 (small)： 36 pt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>一號 One： 26 pt</td>
<td>二號 Two： 21 pt (2)</td>
<td>三號 Three： 16 pt (2)</td>
</tr>
<tr>
<td>小一號 (small)： 24 pt (2)</td>
<td>小二號 (small)： 18 pt</td>
<td>小三號 (small)： 15 pt</td>
</tr>
<tr>
<td>四號 Four： 14 pt</td>
<td>五號 Five： 10.5 pt (1)</td>
<td>六號 Six： 8 pt (1)</td>
</tr>
<tr>
<td>小四號 (small)： 12 pt (1)</td>
<td>小五號 (small)： 9 pt</td>
<td>小六號 (small)： 6.5 pt</td>
</tr>
<tr>
<td></td>
<td>Body text</td>
<td></td>
</tr>
<tr>
<td>七號 Seven： 5.5 pt</td>
<td></td>
<td>八號 Eight： 5 pt</td>
</tr>
<tr>
<td></td>
<td>Ruby</td>
<td></td>
</tr>
</tbody>
</table>
Typographic Scale Convention (China)

There are more scales relate to one another in multiples

<table>
<thead>
<tr>
<th>Set A</th>
<th>Set B (mostly used scale)</th>
<th>Set C (frequently used)</th>
</tr>
</thead>
<tbody>
<tr>
<td>一號 One：26 pt&lt;br&gt;小一號 (small): 24 pt (8/3)</td>
<td>初號 Primary: 42 pt (14/3)&lt;br&gt;小初號 (small): 36 pt (4)&lt;br&gt;Title</td>
<td>三號 Three：16 pt&lt;br&gt;小三號 (small): 15 pt (5/3)</td>
</tr>
<tr>
<td>四號 Four：14 pt&lt;br&gt;小四號 (small): 12 pt (4/3)</td>
<td>二號 Two：21 pt (7/3)&lt;br&gt;小二號 (small): 18 pt (2)&lt;br&gt;Heading</td>
<td>六號 Six：8 pt&lt;br&gt;小六號 (small)：6.5 pt</td>
</tr>
<tr>
<td>七號 Seven：5.5 pt&lt;br&gt;Ruby</td>
<td>五號 Five：10.5 pt&lt;br&gt;小五號 (small)：9 pt (1)&lt;br&gt;Body Text</td>
<td>八號 Eight：5 pt</td>
</tr>
</tbody>
</table>
Typographic Scale Convention (Taiwan)

In Taiwan, new fonts were imported from China and then adapted to the original Japan system.

<table>
<thead>
<tr>
<th>Set A</th>
<th>Set B (mostly used scale)</th>
<th>Set C (frequently used)</th>
</tr>
</thead>
<tbody>
<tr>
<td>初號 Primary: 42 pt (14/3) Larger Title</td>
<td></td>
<td></td>
</tr>
<tr>
<td>新五號四行 New Five*4：36 pt (4) Title</td>
<td></td>
<td></td>
</tr>
<tr>
<td>一號 One：27.5 pt</td>
<td>二號 Two：21 pt (7/3) Heading</td>
<td>三號 Three：16 pt</td>
</tr>
<tr>
<td>四號 Four：13.75 pt</td>
<td>五號 Five：10.5 pt (7/6) Larger Body Text</td>
<td>六號 Six：8 pt</td>
</tr>
<tr>
<td>新四號 New Four:12pt (4/3) Section</td>
<td>新五號 New Five：9 pt (1) Body Text</td>
<td></td>
</tr>
<tr>
<td></td>
<td>七號 Seven：5.25 pt Ruby</td>
<td>八號 Eight：4 pt</td>
</tr>
</tbody>
</table>

Green ones: fonts imported from China
Cells with yellow color: regularly used in combination
Typographic Scale Convention (Taiwan)

The listed sizes of typographic scale in LibreOffice are:

6, 7, 8, 9, 10, 10.5, 11, 12, 13, 14, 15, 16, 18, 20, 22, 24, 26, 28, 32, 36, 40, 44, 48, 54, 60, 66, 72, 80, 88, 96 pt

- Regardless of the small sizes, the mostly used 21 pt & 42 pt of size in Taiwan, Japan and China are missing in the size list.

- The typographic scale convention is getting widely known theses year in Taiwan due to the popularity of the movable type preserved by RiXing Type Foundry and other projects such as 字田活印盒.

- It is better to implement a toggle to switch to the typographic scale convention for ease of use by professional typographic designers.

https://bugs.documentfoundation.org/show_bug.cgi?id=113191
Typically, Chinese paragraphs are indented by 2 characters.

- If you indent the first line by 2 characters, then it is fixed to 21 pt due to the default size as 10.5 pt.

- However, when you adjust the size of the paragraph into 12 pt, the indentation is still 21 pt.

https://bugs.documentfoundation.org/show_bug.cgi?id=36709
Line breaking and word wrapping problems (not yet reported):

1. In Asian Layout setting, “Not to be broken on either side” or 分離禁止文字 (inseparable characters) rule is not supported in LibreOffice, eg. —— and ……

2. There are 3 fundamental methods used to line-break or word-wrap CJK text.
   - Push-in-first
   - Push-out-first
   - Push-out-only, or hanging punctuation (LibreOffice behavior)
Advanced CJK typography

Line breaking and word wrapping

- The red circles in the beginning and the end of the line are forbidden punctuations which shouldn’t be there.

*Table 7-18. Before line breaking*

<table>
<thead>
<tr>
<th>Glyph strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>こうほうこうほう　「ありがとう」 こうほうこうほう</td>
</tr>
<tr>
<td>、 こうほうこうほう　「ありがとう」</td>
</tr>
</tbody>
</table>

Figure from Chapter 7: Typography, *CJKV Information Processing*, 2nd Edition by Ken Lunde

Fare Use
Advanced CJK typography

Line breaking and word wrapping

- **Push-in-first**
  
  Move characters that are prohibited from the beginning back to the end of the previous line.

  Or shift up a character from the following line that are prohibited from terminating.

- It would be great if LibreOffice can support this strategy.

Figure from Chapter 7: Typography, CJKV Information Processing, 2nd Edition by Ken Lunde Fare Use
Advanced CJK typography

Line breaking and word wrapping

- Push-out-first (causes premature end)

Figure from Chapter 7: Typography, *CJKV Information Processing*, 1st Edition by Ken Lunde
Advanced CJK typography

Line breaking and word wrapping

- Push-out-only, or hanging punctuation (adopted in LibreOffice)

A punctuation is left hanging on the right margin (or bottom margin in vertical mode).

Figure from Chapter 7: Typography, CJKV Information Processing, 2nd Edition by Ken Lunde

Fare Use
There are always new Han characters added to each Unicode version

- Unicode 10 Standard: (2017) 136,690 CJK Han characters
- Max glyphs in a OpenType font: 65,535 glyphs
- Two problems of missing Han glyphs:
  - not encoded in the Unicode standard
  - not included in the font although encoded
- Use Unicode IDS (Ideographic Description Sequence) to describe the missing Han characters and compose the glyphs dynamically in 2D

- At http://組字.意傳.台灣/, it will return a rendered picture back. Written in Java, source code licensed under Affero General Public License, GitHub project han3_ji7_tsoo1_kian3.
Possibility of Unicode IDS Support

IDS combination syntaxs:

- ⿱ left to right  e.g. 話 vs ⿱言舌
- ⿰ above to below  e.g. 果 vs ⿰田木 / 翦 vs ⿰明空
- ⿲ left to middle and right  e.g. 湖 vs ⿲氵古月
- ⿳ above to middle and below  e.g. 舅 vs ⿳臼田力
- ⿴ full surround  e.g. 囚 vs ⿴□人
- ⿵ surround from lower left  e.g. 翅 vs ⿵支羽 / 過 vs ⿵辶呚
- ⿶ etc.
Possibility of Unicode IDS Support

- ⼤木木 ➞ 林
- 三木三木木 ➞ 森

This glyph is implemented as a “ligature” feature of Source Han Serif, can be shown before being encoded into the Unicode standard.

Reference site: [http://組字.意傳.台灣/](http://組字.意傳.台灣/)
Special thanks to:

- Dr. Ken Lunde, for his great work on CJKV information processing
- Bobby Tung, for his talk on Bopomofo ruby
- Shoichi Chou, for his talk on Unicode IDS Support
- And the whole LibreOffice community!