LibreOffice
The Document Foundation

ROME
CONFERENCE

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 psycology, convensions on icon design, etc.)

L10N Work by the Community in Taiwan

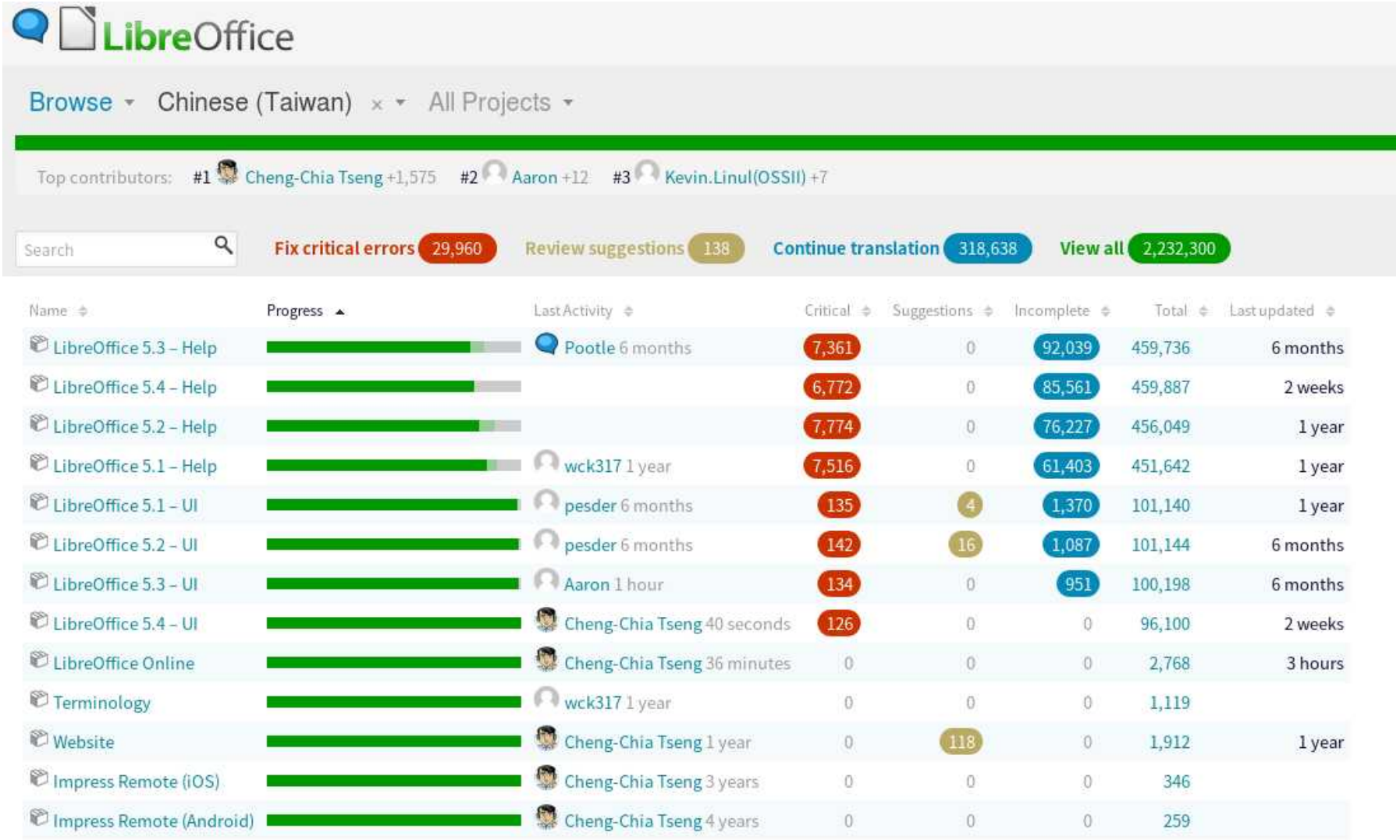
The Taiwan community is getting more and more active these years

- 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

Translation

The translation of LibreOffice Help has been discontinued for some years



LibreOffice

Browse · Chinese (Taiwan) × · All Projects

Top contributors: #1 Cheng-Chia Tseng +1,575 #2 Aaron +12 #3 Kevin.Linul(OSSII) +7

Search | Fix critical errors 29,960 | Review suggestions 138 | Continue translation 318,638 | View all 2,232,300

Name	Progress	Last Activity	Critical	Suggestions	Incomplete	Total	Last updated
LibreOffice 5.3 – Help		Pootle 6 months	7,361	0	92,039	459,736	6 months
LibreOffice 5.4 – Help			6,772	0	85,561	459,887	2 weeks
LibreOffice 5.2 – Help			7,774	0	76,227	456,049	1 year
LibreOffice 5.1 – Help		wck317 1 year	7,516	0	61,403	451,642	1 year
LibreOffice 5.1 – UI		pesder 6 months	135	4	1,370	101,140	1 year
LibreOffice 5.2 – UI		pesder 6 months	142	16	1,087	101,144	6 months
LibreOffice 5.3 – UI		Aaron 1 hour	134	0	951	100,198	6 months
LibreOffice 5.4 – UI		Cheng-Chia Tseng 40 seconds	126	0	0	96,100	2 weeks
LibreOffice Online		Cheng-Chia Tseng 36 minutes	0	0	0	2,768	3 hours
Terminology		wck317 1 year	0	0	0	1,119	
Website		Cheng-Chia Tseng 1 year	0	118	0	1,912	1 year
Impress Remote (iOS)		Cheng-Chia Tseng 3 years	0	0	0	346	
Impress Remote (Android)		Cheng-Chia Tseng 4 years	0	0	0	259	

Minguo Calendar Identification

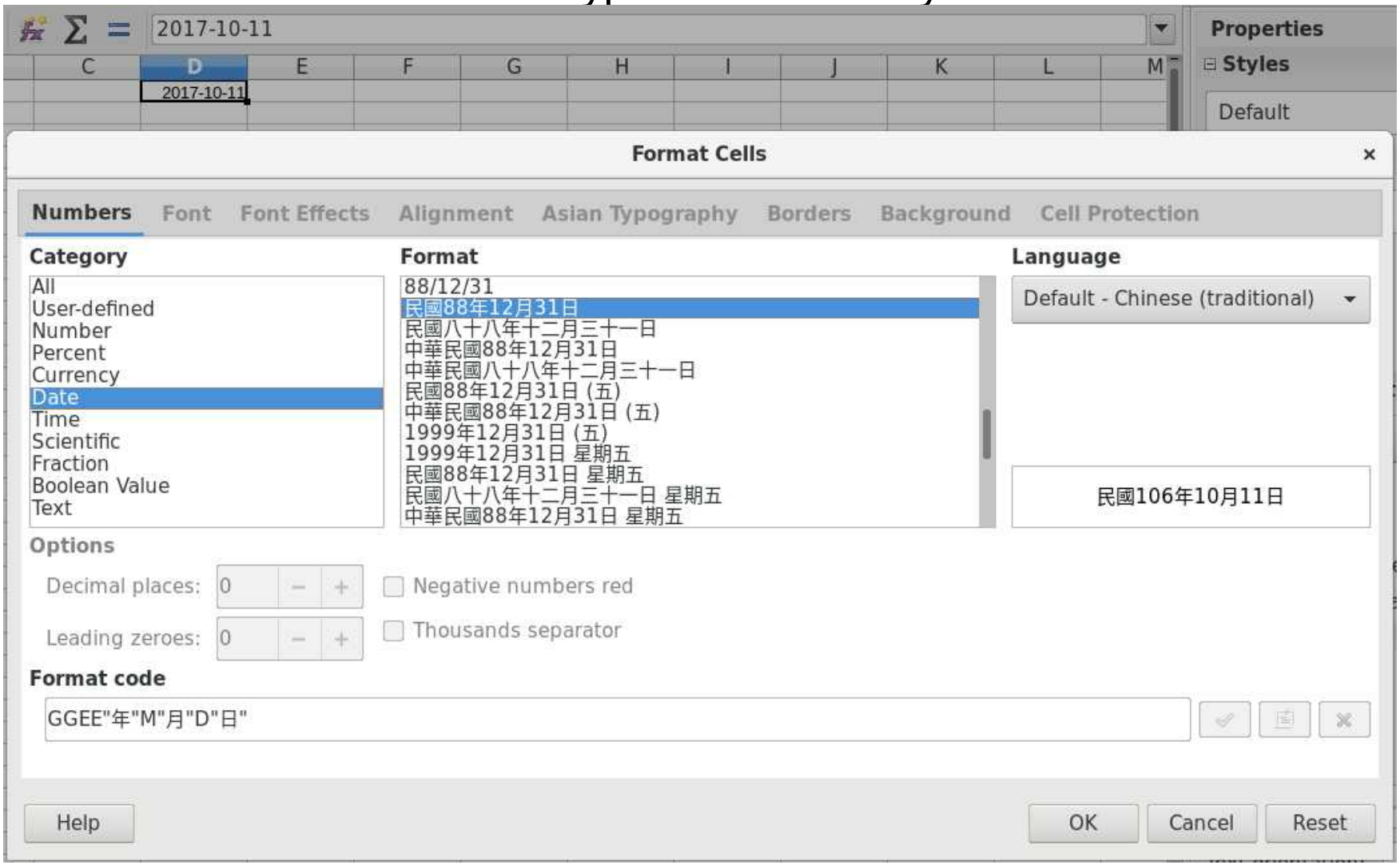
Minguo calendar support is not intuitive enough

- 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

Minguo Calendar Identification

Workaround: uesr have to type common era year first then convert



Vertical Layout of the CJK text

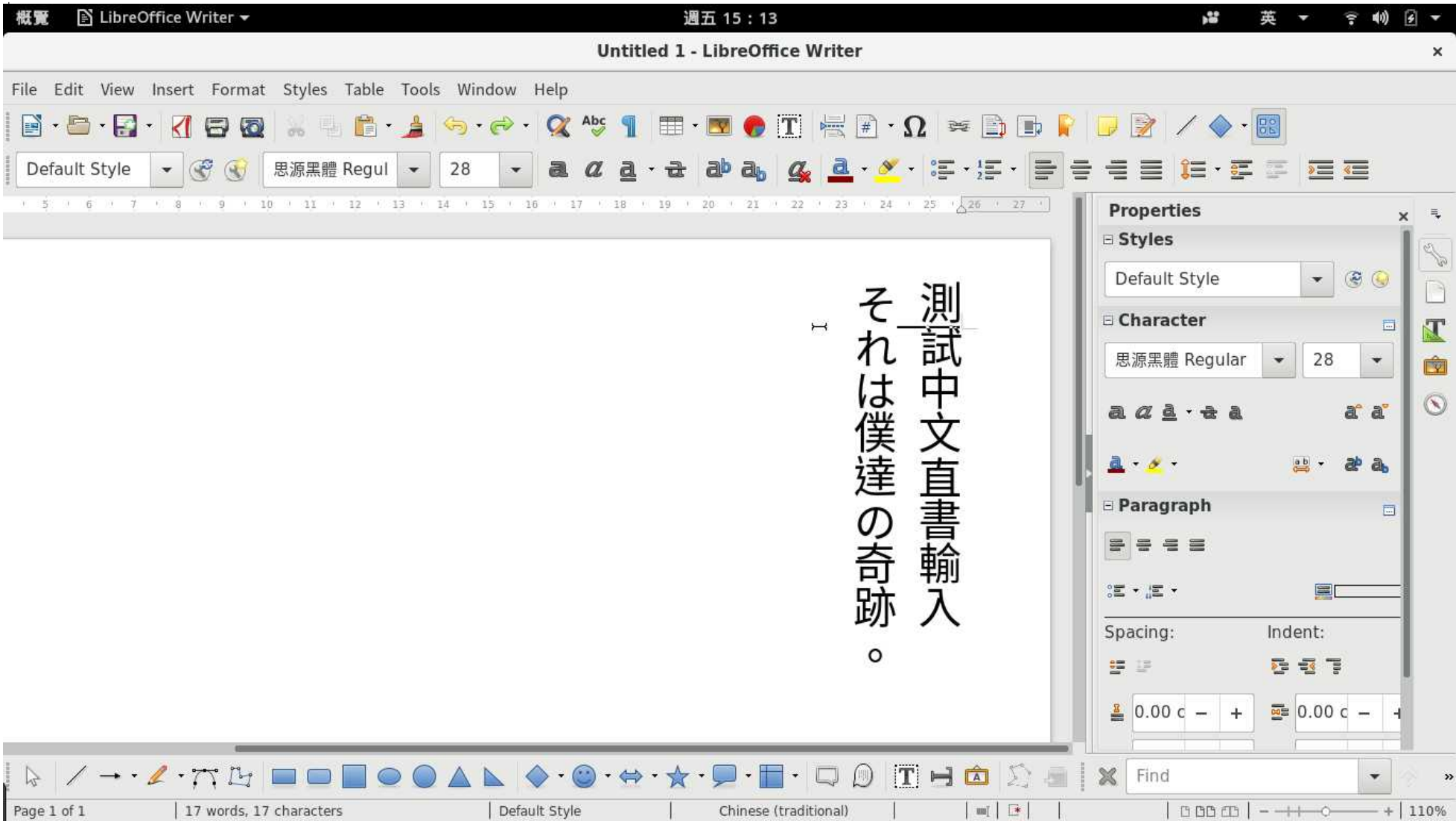
- The support of vertical layout of CJK text is a mess (5.3, 5.4)
  - vertical glyphs cannot be displayed in the slide show (including Latin text)
  - Han glyphs are rendered at a far distance higher from the cursor position

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

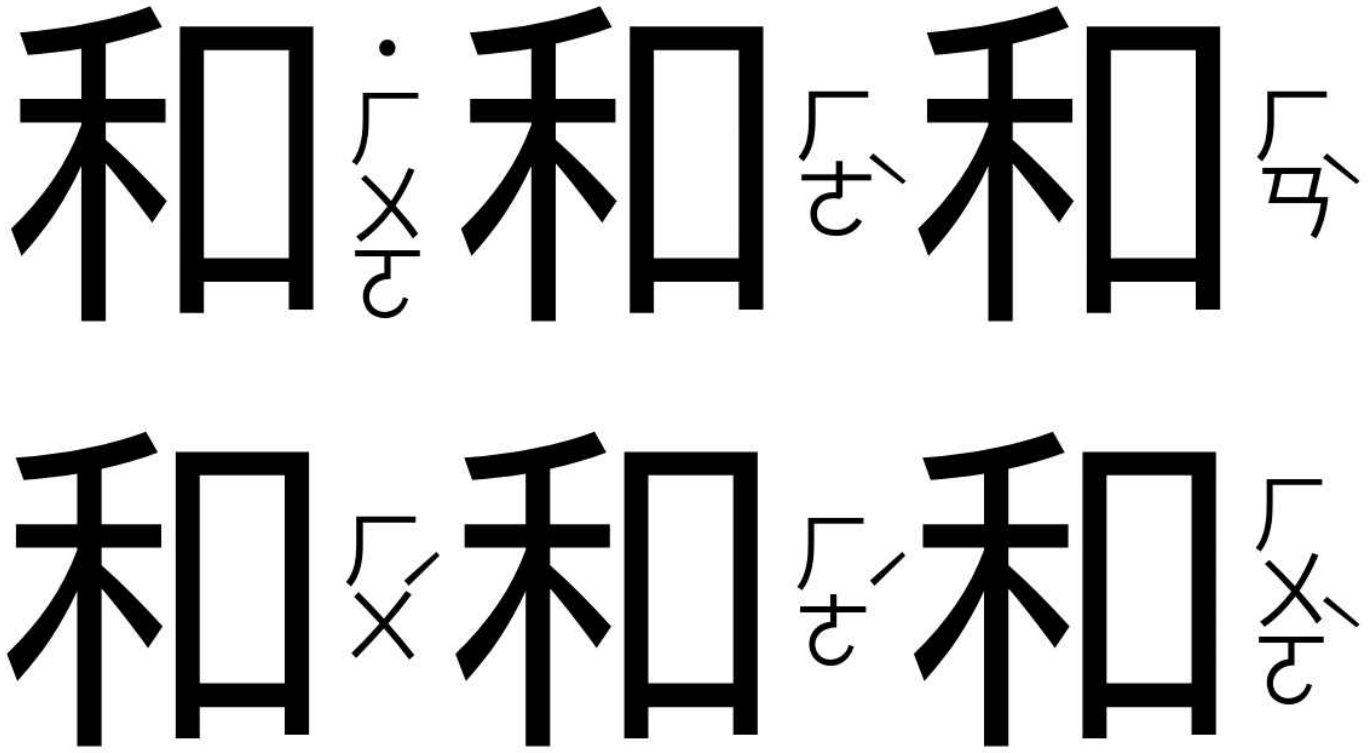
Vertical Layout of the CJK text

- 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.

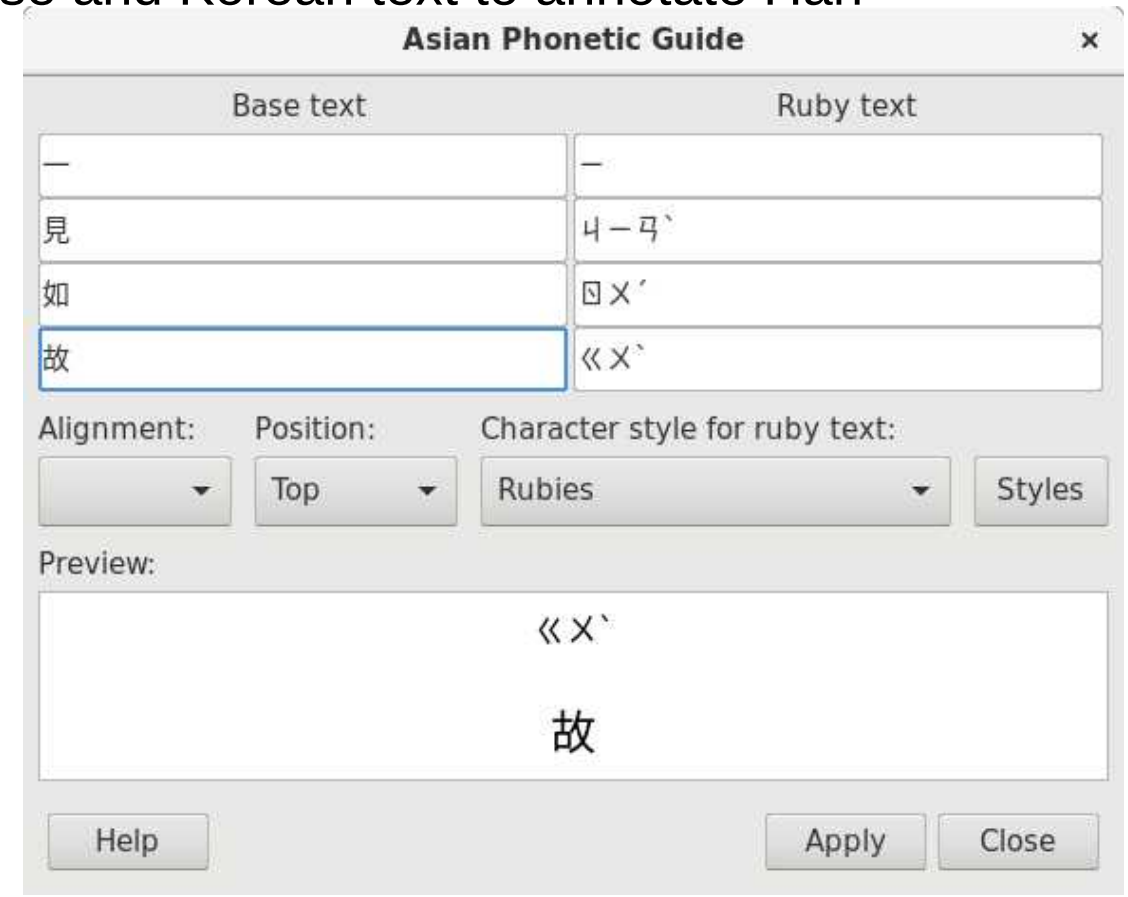


Picture source: https://speakerdeck.com/bobbytung/du-2017-liao-zhu-yin-huan-mei-gao-ding-ma
Author: Bobby Tung, CC-by-SA 4.0 International

Asian Phonetic Guide (Ruby)

- **Mono Ruby**

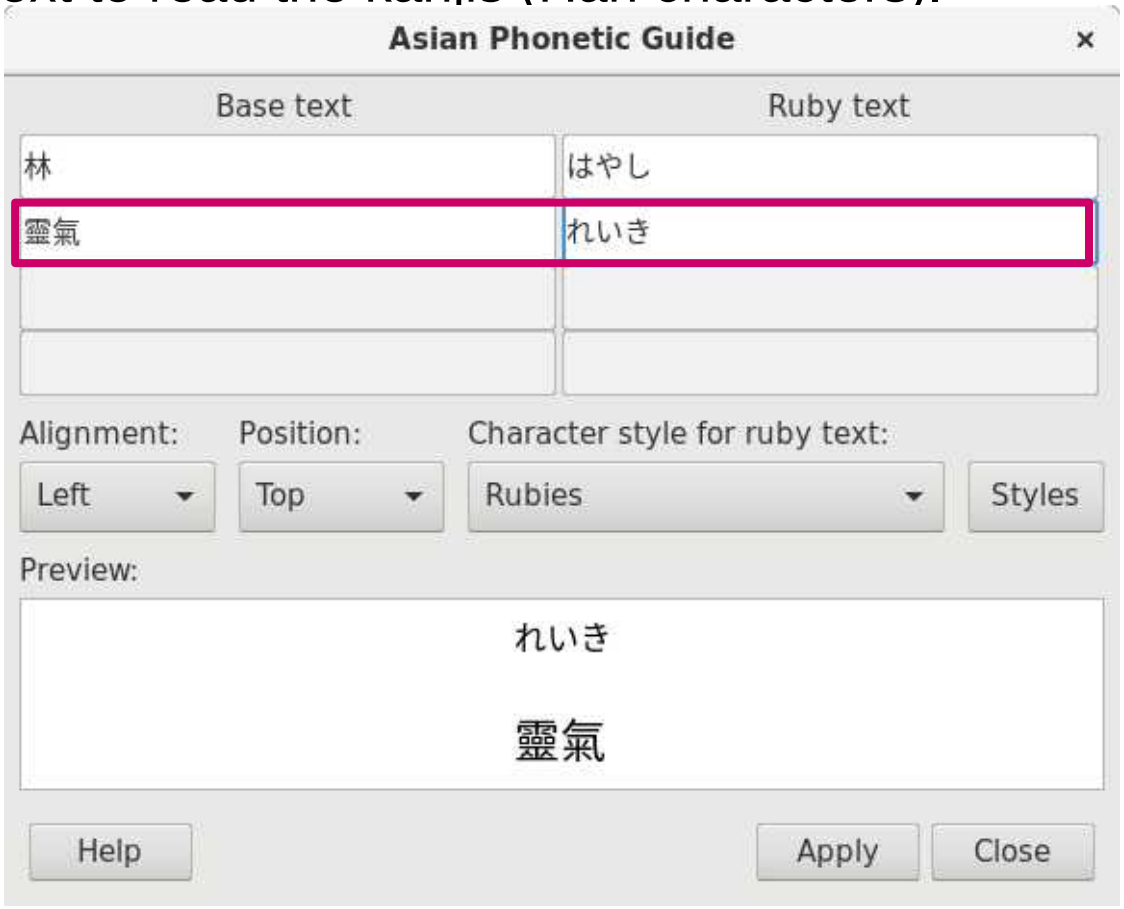
  One or more ruby glyphs serve to annotate only a single base glyph. Used in Chinese, Japanese and Korean text to annotate Han characters.



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).

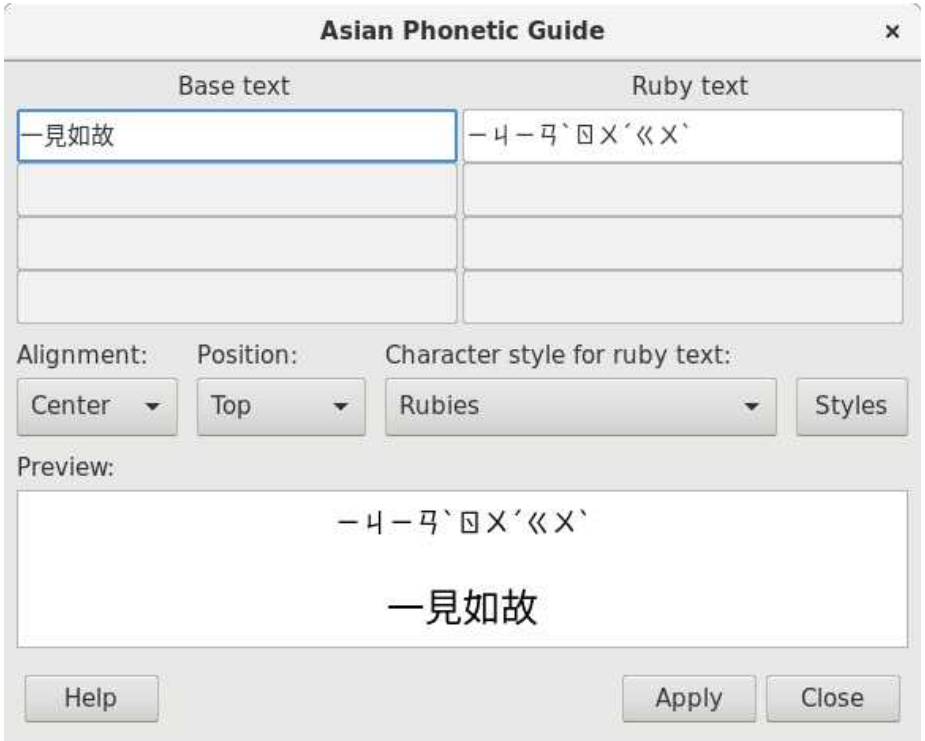


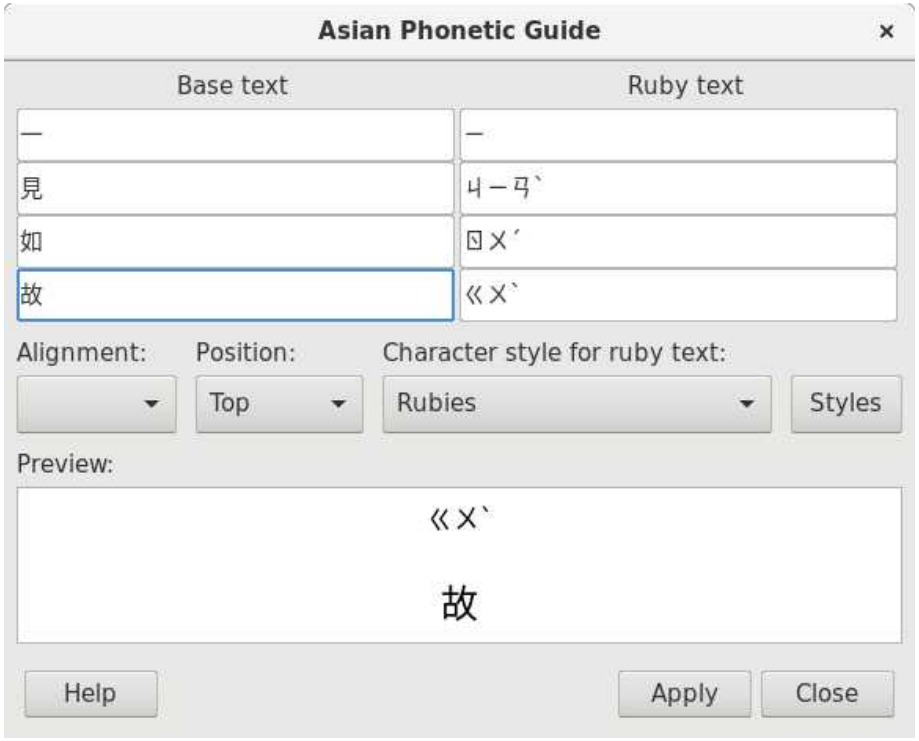
Asian Phonetic Guide (Ruby)

- **The Problem for Chinese**

  LibreOffice seperates 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).
  https://bugs.documentfoundation.org/show_bug.cgi?id=113189





Asian Phonetic Guide (Ruby)

- Horizontal Ruby

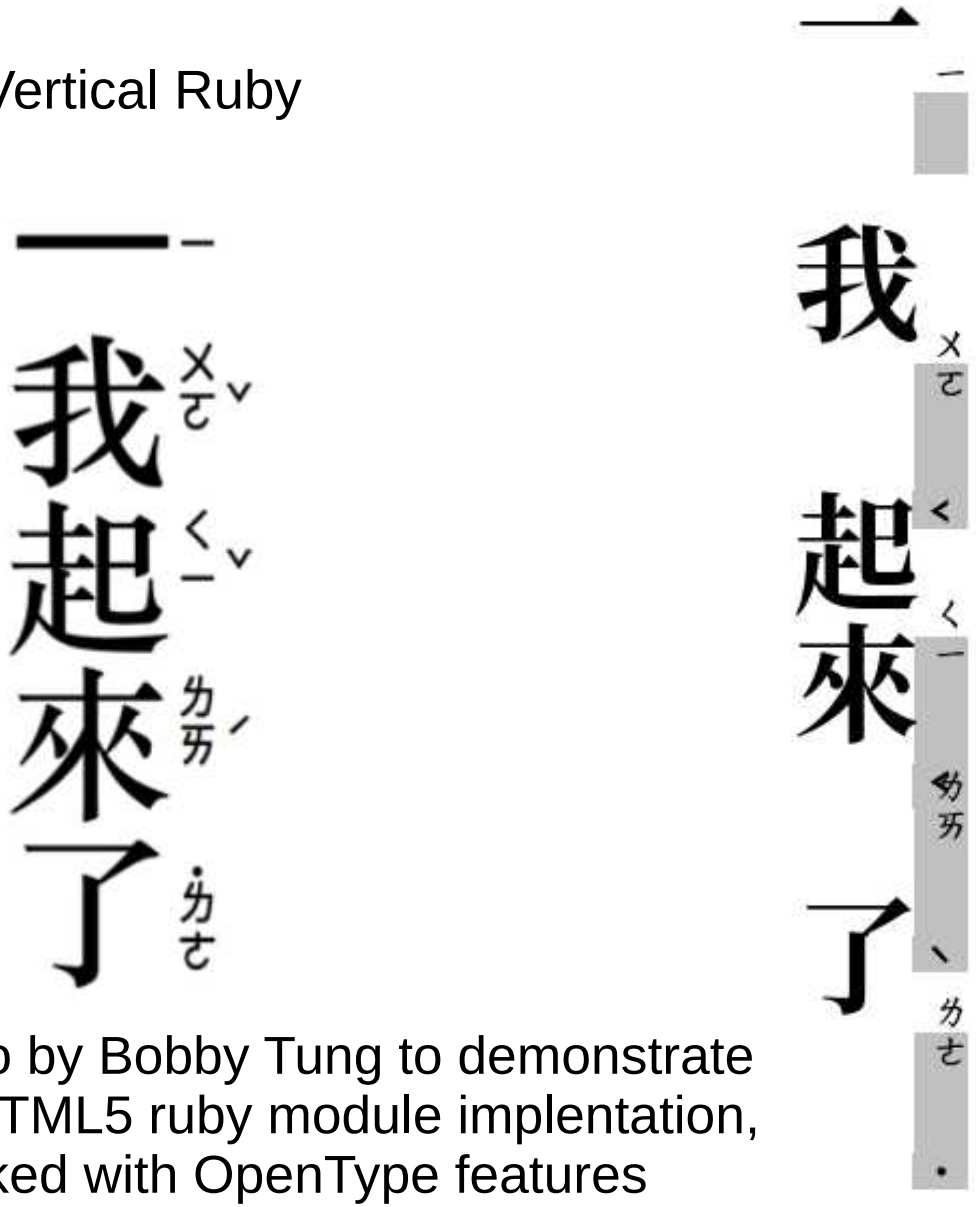
Demo by Bobby Tung to demonstrate the HTML5 ruby module implentation, tweaked with OpenType features

Directly typed with LibreOffice Phonetic Guide

Reference: https://bobbytung.github.io/BopomofoLayoutTest/case01/index.html
Upper picture source: https://speakerdeck.com/bobbytung/du-2017-liao-zhu-yin-huan-mei-gao-ding-ma

Asian Phonetic Guide (Ruby)

- Vertical Ruby



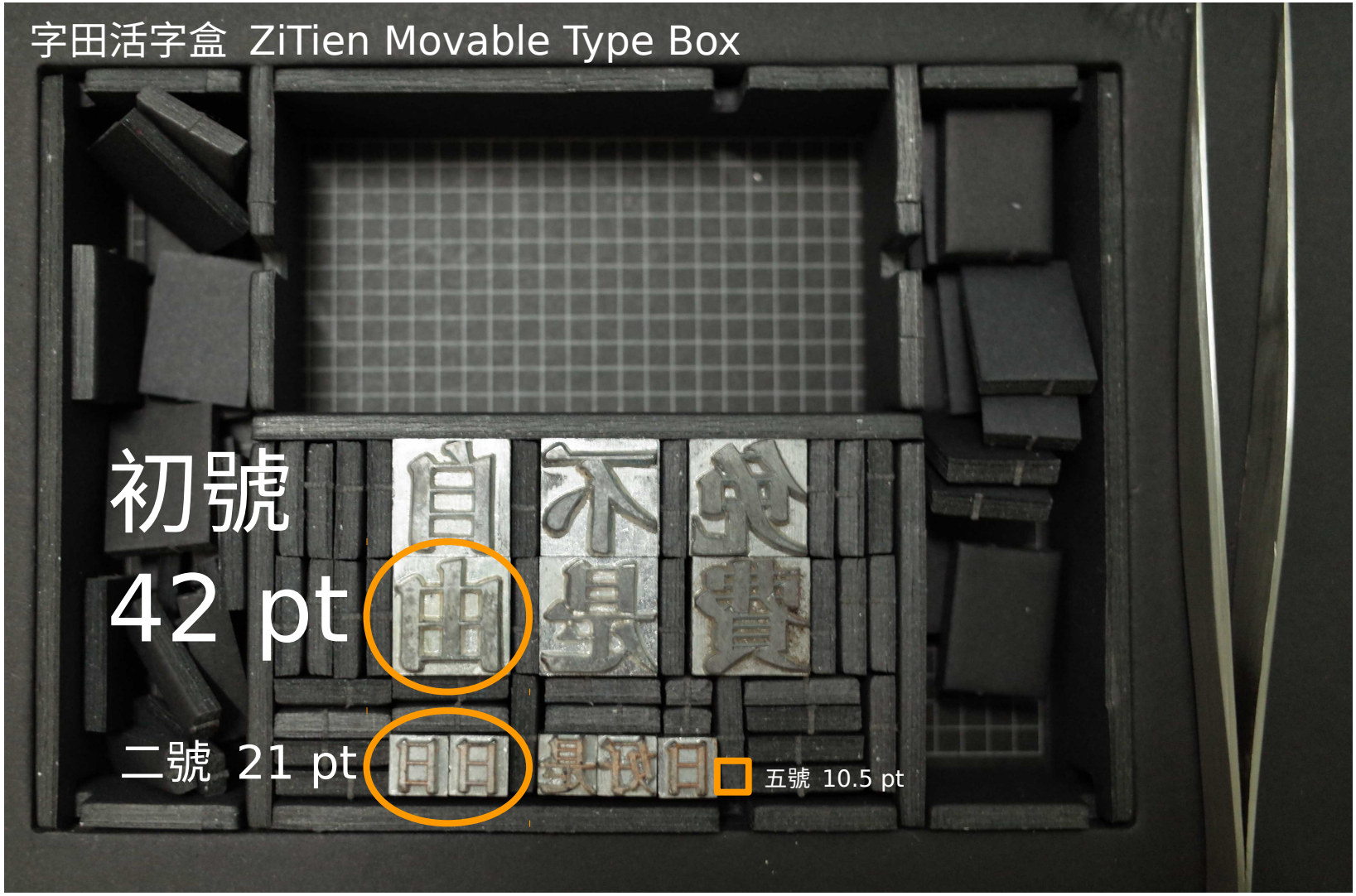
Demo by Bobby Tung to demonstrate the HTML5 ruby module implentation, tweaked with OpenType features

Directly typed with LibreOffice Phonetic Guide

Left picutre source: https://speakerdeck.com/bobbytung/du-2017-liao-zhu-yin-huan-mei-gao-ding-ma

Typographic Scale Convention (Taiwan)

Based on grid system and mutiples between scales

- 

Typographic Scale Convention (Japan)

This system was invented in Japan and introduced to Taiwan.

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

There are 3 sets of scale. Color in yellow: base factor

Typographic Scale Convention (Japan)

Color in yellow: regularly used in combination

Set A	**Set B (mostly used scale)**	Set C (frequently used)
	初號 Primary ： 42 pt (**4**) Title	
一號 One ： 27.5 pt	二號 Two ： 21 pt (**2**) Heading	三號 Three ： 16 pt if treated as 15.75 pt (**1.5**)
四號 Four ： 13.75 pt	**五號 Five ： 10.5 pt (1)** **Body text**	六號 Six ： 8 pt if treated as 7.875 pt (**0.75**)
	七號 Seven ： 5.25 pt (**0.5**) Ruby	八號 Eight ： 4 pt Ruby

Typographic Scale Convention (China)

In China, a system comparable to the Japan system developed

Set A	**Set B (mostly used scale)**	Set C (frequently used)
	初號 Primary： 42 pt (4) 小初號 (small)： 36 pt Title	
一號 One： 26 pt 小一號 (small)： 24 pt (2)	二號 Two： 21 pt (2) 小二號 (small)： 18 pt Heading	三號 Three： 16 pt (2) 小三號 (small)： 15 pt
四號 Four： 14 pt **小四號 (small)： 12 pt (1)**	**五號 Five： 10.5 pt (1)** **小五號 (small)： 9 pt** **Body text**	**六號 Six： 8 pt (1)** **小六號 (small)： 6.5 pt**
	七號 Seven： 5.5 pt Ruby	八號 Eight： 5 pt

Typographic Scale Convention (China)

There are more scales relate to one another in multiples

Set A	**Set B (mostly used scale)**	Set C (frequently used)
	初號 Primary: 42 pt (**14/3**) 小初號 (small)： 36 pt (**4**) Title	
一號 One： 26 pt 小一號 (small): 24 pt (**8/3**)	二號 Two： 21 pt (**7/3**) 小二號 (small)： 18 pt (**2**) Heading	三號 Three： 16 pt 小三號 (small):15 pt (**5/3**)
四號 Four： 14 pt **小四號 (small):12 pt (4/3)**	**五號 Five： 10.5 pt** **小五號 (small)： 9 pt (1)** **Body Text**	**六號 Six： 8 pt** **小六號 (small)： 6.5 pt**
	七號 Seven： 5.5 pt Ruby	八號 Eight： 5 pt

Typographic Scale Convention (Taiwan)

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

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

Green ones: fonts imported from China
Cells with yellow color: regularly used in combination

Typographic Scale Convention (Taiwan)

The listed sizes of typpographic scale in LibreOffice are:

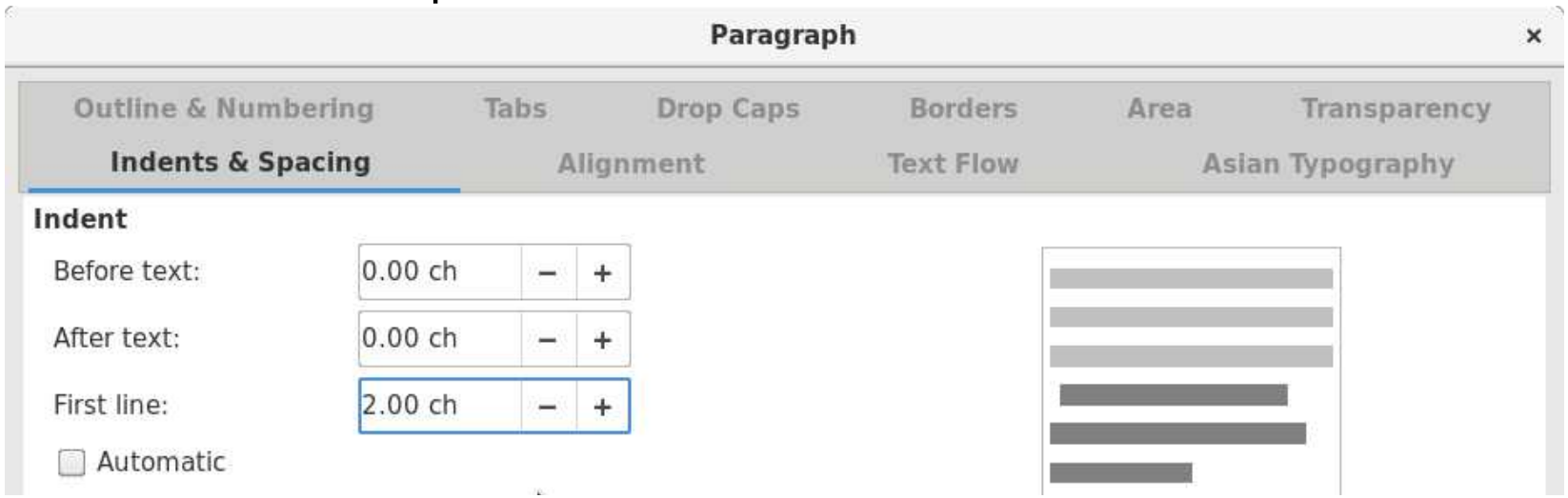
6, 7, 8, 9, 10, 10.5, 11, 12, 13, 14, 15, 16, 18, 20, 22, 24, 26, 28, 32, 36, 40, 44, 48, 54, 60, 66, 72, 80, 88, 96 pt

- 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 日星鑄字行 RiXing Type Foundry and other projects such as 字田活印盒 .
- It is better to **implent 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

Typography: first line indentation

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

Advanced CJK typography

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

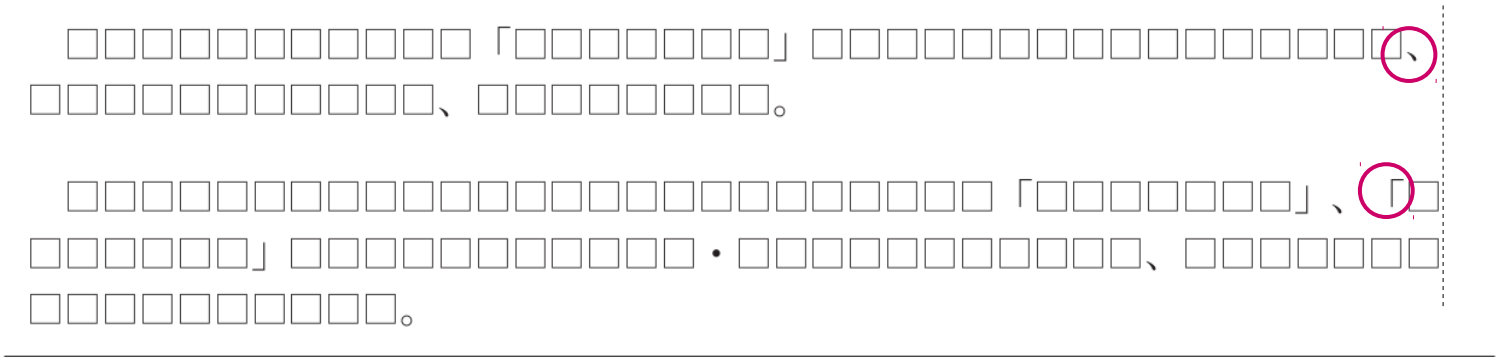
Glyph strings
□□□□□□□□□□□「□□□□□□□」□□□□□□□□□□□□□□□□□□□□、□□□□□□□□□□□□、□□□□□□□□。
□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□「□□□□□□□」、「□□□□□□□」□□□□□□□□□□□□・□□□□□□□□□□□□、□□□□□□□□□□□□□□□□。

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)

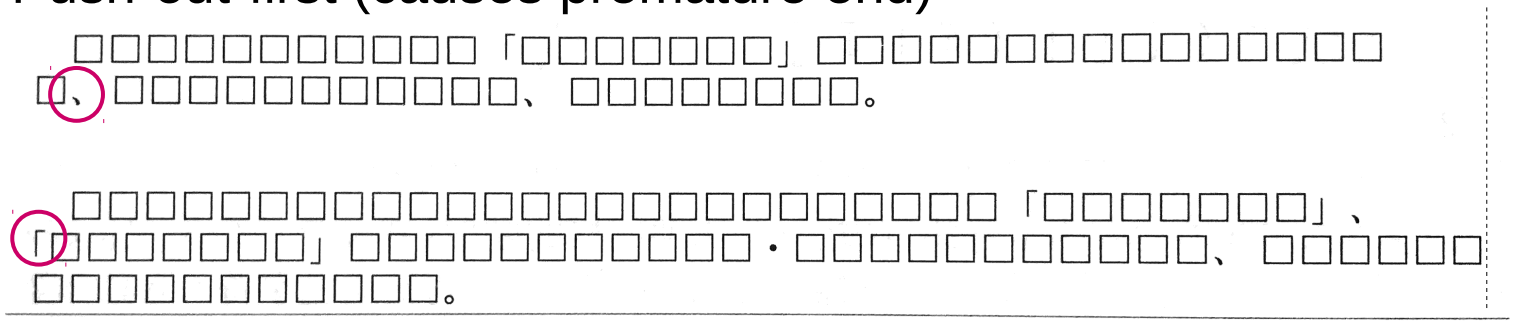


Figure from Chapter 7: Typography, *CJKV Information Processing*, 1st Edition by Ken Lunde
Fare Use

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

Possibility of Unicode IDS Support

There are always new Han chracters 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 dinamically 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

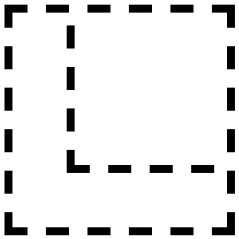
IDS combination syntaxs:

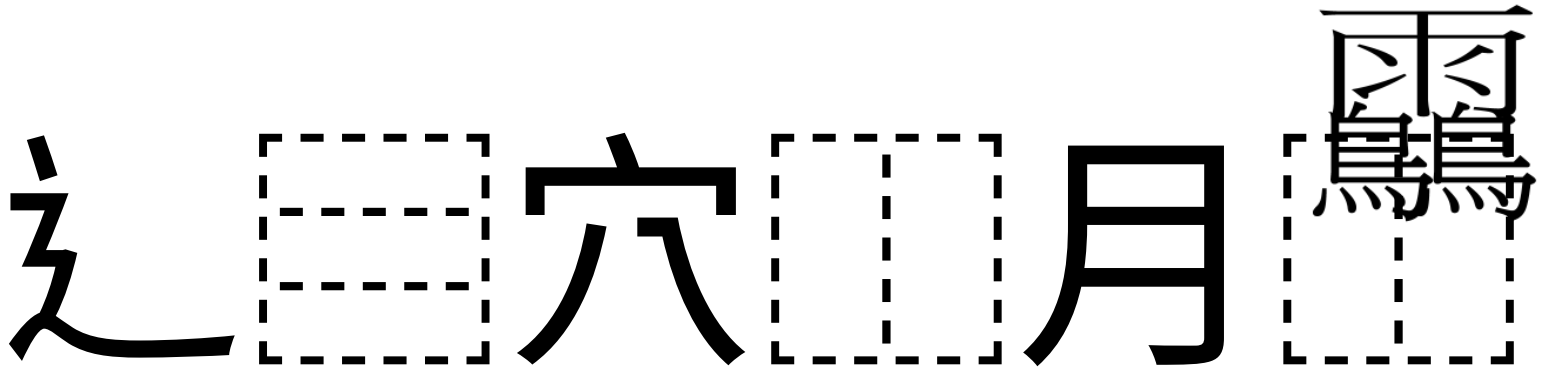
- ⿰ left to right e.g. 話 vs ⿰言舌
- ⿱ above to below e.g. 果 vs ⿱田木 / 嬰 vs ⿱明空
- ⿲ left to middle and right e.g. 湖 vs ⿲氵古月
- ⿳ above to middle and below e.g. 舅 vs ⿳臼田力
- ⿴ full surround e.g. 囚 vs ⿴口人
- ⿵ surround from lower left e.g. 翅 vs ⿵支羽 / 過 vs ⿵辶咼
- ⿶⿷⿸⿹⿺⿻ etc.

Possibility of Unicode IDS Support

- ⿰木木 => 林
- ⿱木⿰木木 => 森
- This glyph is implented as a “ligature” feature of Source Han Serif, can be shown before being encoded into the Unicode standard.

⿺辶⿱穴⿰月⿰⿰⿱幺長⿱言馬⿱幺長 刂 =>





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!

Questions?