

xjyutping: Jyutping above traditional Chinese characters

Version 1.2.0

28 September 2026

Abstract

xjyutping typesets Jyutping (粵拼), the Linguistic Society of Hong Kong romanisation of Cantonese, above Chinese characters. Readings are chosen from the surrounding words, so 行 is *hong4* in 銀行 and *haang4* in 行路. Every character sits in a cell of the same width, wide enough for its Jyutping, so the text stays evenly spaced and no Jyutping touches its neighbours or the line above. The package works with Xe_{La}T_EX (through xeCJK) and with Lua_{La}T_EX (through Lua_{TeX}-ja), and with the ctex package and classes under either engine.

1 Getting started

```
\documentclass{article}
\usepackage[fontset=none]{ctex} % for xelatex or lualatex
\setCJKmainfont{Songti TC}
\usepackage{xjyutping}
\begin{document}
\begin{jyutpingscope}
我哋去銀行，行路返屋企。校長嘅學校長大。
\end{jyutpingscope}
\end{document}
```

ngo5 dei6 heoi3 ngan4 hong4 haang4 lou6 faan1 uk1 kei2 haau6 zoeng2 hai2 hok6 haau6 zoeng2 daai6
我 哋 去 銀 行， 行 路 返 屋 企。 校 長 嘅 學 校 長 大。

Compile with xelatex or lualatex. ctex sets up Chinese typesetting for either engine; you can instead load xeCJK under Xe_{La}T_EX or luatexja-fontspec (with \setmainjfont) under Lua_{La}T_EX, and if none of them is loaded, xjyutping loads xeCJK or Lua_{TeX}-ja itself. Any traditional Chinese font will do; the Jyutping uses the document's main Latin font unless you choose another.

2 Commands

2.1 Annotating text

\begin{jyutpingscope}[\textit{options}] ... \end{jyutpingscope} annotates every Chinese character in the block. The block ends the paragraph, and its line spacing is raised where needed so that the Jyutping of one line never comes close to the line above.

\xjyutping*[\textit{options}]{\textit{text}} does the same for a piece of running text inside an ordinary paragraph.

香港人講\xjyutping*{廣東話}，寫\xjyutping*{繁體字}。

gwong2 dung1 waa2 faan4 tai2 zi6
香 港 人 講 廣 東 話， 寫 繁 體 字。

2.2 Choosing a reading yourself

Cantonese pronunciation shifts with meaning and with colloquial tone change, and no word list is perfect. There are three ways to say what you want.

`\xjyutping[options]{characters}{readings}` gives the characters an explicit reading, here and nowhere else. It works inside a scope and on its own, and it also marks a word boundary.

`\setjyutping{character}{reading}` changes the reading a character gets when it is not part of a known word.

`\setjyutping{word}{readings}` adds a word, or overrides the reading of a known one. Words always win over single-character defaults.

Readings are Jyutping syllables separated by spaces, one per Chinese character. Punctuation, Latin letters and commands among the characters take none, so `\xjyutping{A行}{hong4}` reads 行 as *hong4*. `\setjyutping` is global and takes effect from where it appears, so it can go in the preamble or in the middle of a scope.

In the next example 重話 means ‘also said’ (*zung6*, not the default *cung5* ‘heavy’), and 家行 is a name read *gaa1 hang4*:

```
\setjyutping{重話}{zung6 waa6}  
\begin{jyutpingscope}  
佢重話\xjyutping{家行}{gaa1 hang4} 聽日㗎。  
\end{jyutpingscope}
```

keoi5 zung6 waa6 gaa1 hang4 ting1 jat6 lai4
佢 重 話 家 行 聽 日 㗎。

2.3 Formatting inside words

Braces and formatting commands do not split a word, so a character can be highlighted without losing its reading in context:

```
\begin{jyutpingscope}  
佢喺銀\textbf{行} 返工，係校\textcolor{red}{長} 嘅朋友。  
\end{jyutpingscope}
```

keoi5 hai2 ngan4 hong4 faan1 gung1 hai6 haau6 zoeng2 ge3 pang4 jau5
佢 喺 銀 行 返 工，係 校 長 嘅 朋 友。

2.4 Switching off

`\disablejyutping` stops annotation until `\enablejyutping` or the end of the current group or environment. Use it for passages that should stay plain inside a scope.

3 How readings are chosen

1. The text is split into runs of Chinese characters. Punctuation, Latin text and most commands end a run. Spaces and line breaks between two characters do not, and neither do braces or formatting commands (`\textbf`, `\emph`, `\color`, `\textcolor`, size commands ...). The text of a `\footnote` is a run of its own, and the text around the footnote mark carries on.

2. Each run is divided into words from a list of about 100 000 Cantonese words: the division with the fewest words wins, then the one with the fewest single characters, and on a tie the longer final word.
3. A word takes its reading from the list. Your `\setjyutping` words are checked first.
4. A character that is not part of a word takes your `\setjyutping` reading if you gave one, otherwise its default reading. A few characters read differently at the end of a run: 呢 is *ni1* in 呢張床 but the particle *ne1* in 你呢?
5. `\xjyutping{characters}{readings}` always wins, and ends the run.

Traditional characters have several accepted shapes (為 and 爲, 裡 and 裏, 說 and 説, 衛 and 衛, 線 and 綫, 溫 and 温 ...). For word lookup they are folded together, so a word is found whichever shape you type.

3.1 Proofreading

About four thousand characters have more than one common reading. When such a character is not settled by a word or by you, its reading is only a guess. The `multiple` option formats those guesses, and the `debug` option writes every run to the log with the alternatives:

```
\xjyutpingsetup{multiple=\color{red}, debug}
\begin{jyutpingscope}
佢重未嚟。
\end{jyutpingscope}
```

keoi5 cung5 mei6 lai4
佢 重 未 嚟。

The log then contains a line such as

```
xjyutping> 佢 keoi5:s | 重 cung5:m(zung6 cung4) | 未 mei6:s | 嚟 lai4:m(lei4) |
```

Each character is followed by its reading and a type: **w** from the word list, **u** set by you, **m** a guessed polyphone (other readings in brackets), **s** a character with a single common reading. Here 重 means ‘still’, so `\setjyutping{重未}{zung6 mei6}` fixes it (and `\setjyutping{重}{zung6}` would make *zung6* the default for 重 on its own).

4 Options

Options can be given to `\usepackage`, to `\xjyutpingsetup`, or in the optional argument of `jyutpingscope` and `\xjyutping`.

Key	Default	Meaning
<code>ratio</code>	0.45	Jyutping size relative to the text
<code>vsep</code>	1.05em	height of the Jyutping baseline above the character’s
<code>hsep</code>	0.15em plus 0.4em	space between cells: the least gap between two Jyutping, plus stretch for justified lines
<code>width</code>	auto	cell width: auto , natural or a length
<code>font</code>	<code>\normalfont</code>	font of the Jyutping (font commands only)
<code>format</code>		extra formatting, e.g. <code>\color{gray}</code>
<code>multiple</code>		formatting for guessed polyphones
<code>fancy</code>	<code>false</code>	tones as pitch strokes (section 5)
<code>linebreak</code>	<code>false</code>	keep the lines of the source (section 6)
<code>debug</code>	<code>false</code>	log every run

4.1 Cell width

With `width=auto` every character in a block gets the same cell, wide enough for the longest Jyutping in that block. With a length you get a fixed grid across the whole document; a Jyutping longer than the cell widens its own cell rather than overlap. `width=natural` makes each cell only as wide as its own character or Jyutping: tighter, but uneven.

```
\xjyutping*[width=auto]{香港中文大學}
\xjyutping*[width=natural]{香港中文大學}
\xjyutping*[width=2em]{香港中文大學}
```

hoeng1 gong2 zung1 man4 daai6 hok6
香 港 中 文 大 學
hoeng1 gong2 zung1 man4 daai6 hok6
香 港 中 文 大 學
hoeng1 gong2 zung1 man4 daai6 hok6
香 港 中 文 大 學

4.2 Size and font

```
\xjyutping*[ratio=0.6, font=\sffamily, format=\color{gray}]{早晨，食咗飯未？}
```

zou2 san4 sik6 zo2 faan6 mei6
早 晨， 食 咗 飯 未？

5 Fancy tones

With the `fancy` option, `\usepackage[fancy]{xjyutping}`, each tone number is shown the way Visual Jyutping shows it: a small stroke that traces the pitch of the tone, then a smaller tone number, raised for the two high tones and lowered for the others. The strokes follow Visual Jyutping's symbols: tone 1 high, tone 2 rising to high, tone 3 mid, tone 4 falling to low, tone 5 rising from low, tone 6 low.

```
{\Large\xjyutping*[fancy]{詩史試時市事}}
\begin{jyutpingscope}[fancy]
我哋去銀行，行路返屋企。校長喺學校長大。
\end{jyutpingscope}
```

si⁻¹ si^{/2} si⁻³ si^{\4} si^{/5} si⁻⁶
詩 史 試 時 市 事
ngo^{/5} dei⁻⁶ heoi⁻³ ngan^{\4} hong^{\4} haang^{\4} lou⁻⁶ faan⁻¹ uk⁻¹ kei^{/2} haa^{u-6} zoeng^{/2} hai^{/2} hok⁻⁶ haa^{u-6} zoeng^{/2} daai⁻⁶
我 哋 去 銀 行， 行 路 返 屋 企。 校 長 喺 學 校 長 大。

The strokes are drawn, not taken from a font, so they work with any `font` and take the colour of `format` and `multiple`. The spacing follows the new shape: cells are measured with the strokes and small numbers in place (so `width=auto` cells come out a little wider), line spacing makes room for the raised numbers, and if the lowered numbers reach further down than the letters in the chosen font, the Jyutping is lifted by the difference so that it keeps its distance from the character. Like every option, `fancy` can be switched on or off for one scope: `\begin{jyutpingscope}[fancy=false]`. A syllable given without a tone number (`\xjyutping{唔}{m}`) is printed as it is.

6 Verse and lyrics

With the `linebreak` option the lines of the source are kept: each line end inside `jyutpingscope` becomes a line break, as if it were written `\\` or `\newline`, and each blank line a paragraph break, so that the next stanza starts a new paragraph by the document’s usual rules (indented by default; with no indent and some space before it under `\usepackage[parfill]{parskip}`).

```
\begin{jyutpingscope}[linebreak]
床前明月光，
疑是地上霜。

舉頭望明月，
低頭思故鄉。
\end{jyutpingscope}
```

cong4 cin4 ming4 jyut6 gwong1
床 前 明 月 光，
ji4 si6 dei6 soeng6 soeng1
疑 是 地 上 霜。
geoi2 tau4 mong6 ming4 jyut6
舉 頭 望 明 月，
dai1 tau4 si1 gu3 hoeng1
低 頭 思 故 鄉。

Line ends at the start and the end of the scope, after a `\\` and before `\begin`, `\end` or a new paragraph add no break, and a line ending in `%` joins the next one as usual. A line break also ends a run of characters, so no word is looked up across two lines. The option works for the environment only, and only where the line ends are still in the source when the scope begins: not in `\xjyutping*`, nor in a scope inside the argument of a command (`\parbox{...}`) or inside a scope without the option, whose text has been read already. Inside an environment such as `minipage` it works.

7 Xe_{La}T_EX and Lua_{La}T_EX

The readings, the options and the layout are the same under both engines; only the machinery differs. Under Xe_{La}T_EX, `xeCJK` hands each Chinese character to a hook in which `xjyutping` builds the character’s cell. Under Lua_{La}T_EX the rubies are built as the text is read, and a Lua function (in `xjyutping.lua`, which must be installed next to `xjyutping.sty`) puts each cell together after Lua_{TeX}-ja has laid out the line. So under Lua_{La}T_EX:

- Lua_{TeX}-ja’s own rules for punctuation widths and line breaks apply; for example, a line never breaks just before a dash (—) or an ellipsis (⋯).
- A paragraph can be as long as you like (a paragraph of 40 000 characters compiles).
- Compiling takes about 1.6 times as long as with Xe_{La}T_EX.
- Text that comes from a macro is annotated with the options and the size of the place where it stands, but with the `\setjyutping` readings in force when its paragraph ends.

8 Limits

- The body of `jyutpingscope` and the argument of `\xjyutping*` are read in full before anything is typeset, like any argument, so `\verb` and `verbatim` environments cannot go inside (put them outside the scope), and in a `\url` or `\href` inside a scope a `%` must be written `\%`.

- `\input{<file>}` inside a scope reads the file as part of the scope, with the same restrictions as the body; a file that uses `\endinput`, `\verb`, verbatim environments, `\makeatletter` or catcode changes is input normally instead. Such files, text that comes from a macro, an `\included` file or `\maketitle` are annotated character by character, without word context; the debug log marks them `[no context]`. Put `\xjyutping*` inside such a macro, or a scope inside the included file.
- The arguments of `\label`, `\ref`, `\cite`, `\index`, `\url`, the first argument of `\href`, the optional argument of `\hyperref`, `\hyperlink`, `\pdfbookmark`, `\includegraphics`, the `cleveref` commands and environment names are passed through untouched.
- Give other commands their Chinese argument in braces (`\textbf{行}`, not `\textbf 行`).
- Section titles, captions and footnotes inside a scope are annotated. The table of contents is annotated only if `\tableofcontents` is itself inside a scope (an explicit `\xjyutping` in a title is annotated there too); running heads and PDF bookmarks are always plain.
- In `beamer` a `\frametitle` inside a scope stays plain (it is typeset after the scope has ended): write `\frametitle{\xjyutping*{...}}`.
- The underline and emphasis-mark commands of `xeCJKfntef` (Xe $\LaTeX$ only) do not keep the cell spacing inside a scope; use `\underline` or `ulem`'s `\uline` there.
- A scope can hold well over 100 000 characters. Under Xe $\LaTeX$ one paragraph is limited to about 15 000 by $\TeX$'s memory (about 13 000 with `fancy`); Lua $\LaTeX$ has no such limit.
- A `\setjyutping` inside a scope is noticed while the text is read, so one inside `\iffalse...\fi` or an unused macro definition still applies to the text after it.
- Characters missing from the data are typeset in their cell without Jyutping.
- Xe $\LaTeX$ and Lua $\LaTeX$ are supported; pdf $\LaTeX$ is not.

9 Data and licences

The readings come from these sources (the README lists them with their folders and commits):

- the LSHK *Cantonese Pronunciation List of Characters for Computers* (粵拼表) and the rime-cantonese dictionaries (both CC BY 4.0), which give the character readings, the defaults and most of the words; rime-cantonese is authoritative;
- CC-Canto and the Cantonese readings of CC-CEDICT (both CC BY-SA 3.0, by Pleco), as distributed with Jyut Dictionary, which add about 2 300 words, each only where it changes a reading and passes checks against rime-cantonese;
- the book data of 粵音資料集叢, which gives the readings of 640 rare characters the others lack;
- OpenCC's variant tables (Apache-2.0).

`tools/build-data.py` regenerates `xjyutping-chars.def` and `xjyutping-words.def` from these sources, and `tools/fetch-sources.sh` fetches them; the tables at the top of the build script hold the hand-checked corrections.

The package code (`xjyutping.sty`, `xjyutping.lua`) is under the $\LaTeX$ Project Public License 1.3c. The data files are distributed under CC BY-SA 4.0, because they adapt the CC BY-SA word lists. The readings taken from 粵音資料集叢 come from data published without a licence statement and are used with attribution.

10 Acknowledgements

`xjyutping` was inspired by the authors of the $\text{\LaTeX}$ package `xpinyin` by Qing Lee (李清), which puts Hanyu Pinyin above simplified Chinese characters, and follows its way of annotating characters through `xeCJK`'s `\CJKsymbol` hook. The `fancy` option was inspired by Visual Jyutping by Vincent Tam, whose tone symbols it draws, and by the Visual Cantonese Fonts (粵語字體) by Jon Chui / A3I Ltd. (canto.hk, docs.visual-fonts.com).

Many thanks to the authors of every source of the readings: the Jyutping Workgroup of the Linguistic Society of Hong Kong, and those it thanks, Prof Lu Qin, Dr Cheung Kwan Hin and Nathan Hammond; the Cantonese Computational Linguistics Infrastructure Development Workgroup (CanCLID) and the contributors of rime-cantonese; 石見田 and 粵音資料集叢, with the authors and editors of the dictionaries digitised there; Aaron Tan's Jyut Dictionary; Pleco for CC-Canto and the Cantonese readings of CC-CEDICT, and MDBG and the contributors of CC-CEDICT; and Carbo Kuo (BYVoid) and the contributors of OpenCC.