東方歌詞翻譯遷移至 sak.uy

最近幾個月在這個博客發了不少歌詞翻譯 ~~似乎有要轉型成音樂博主的趨勢~~，前段時間買了個新域名 sak.uy，準備專門用來放這些東方歌曲的歌詞翻譯，於是分設了單獨的博客「Sakuya的音樂盒」。主博客這邊右側邊欄會有到音樂盒的鏈接。

曾經在這邊的那些歌儘量保持 URL 跳轉過去，新的歌詞翻譯會發到那邊去，還想繼續聽歌的話請繼續訂閱那邊的 RSS 呀。

主博客這邊還是像往常一樣保持記錄生活點滴和技術經驗好了。說道介紹技術， 有人問過我那些日語歌詞上給漢字標註的假名都是我一個個手輸的麼？ 一開始是手輸的，後來發現了不錯的自動化方案，於是這裏介紹一下。

首先是　python-furigana

這是個 python 寫的小程序（嚴格說是庫），可以把一段日文轉換成標準的 HTML 形式的 `<ruby>` 標籤的振假名(振 (ふ) り 仮名 (かな))。 它本身只是個方便的格式化庫，實際工作是用 python-mecab 這個 binding 去查詢 mecab 這個著名的日語語料分析庫。要用它還得配合一些開源的 mecab 詞典，這些在 `[archlinuxcn]` 都有打好的包了，直接安裝：

```
 1  $ sudo pacman -Syu python-furigana m
ecab-git python-mecab mecab-ipadic
```

裝好之後用法也很直接，甚至沒有 binary 直接調用 python 的 module 就可以：

```
$ python -m furigana.furigana "振り仮名の例"
<ruby><rb>振</rb><rt>ふ</rt></ruby>り<ruby><rb>仮名</rb><rt>かめい</rt></ruby>の<ruby><rb>例</rb><rt>れい</rt></ruby>
```

就是提供日語作爲輸入，然後輸出 HTML 形式的 `<ruby>` 標籤而已。 像上面的例子中出現的錯誤（「振り仮名」完整的一個詞中「仮名」意思是「平仮名」應該發音「がな」而非意爲「假的人名」的「かめい」）可以看出其實標註的準確率還是有些問題的。嘛日語作爲一個非常依賴上下文判斷的語言， 經常日本人都會搞錯某些漢字的發音，這些也不能強求機械化的算法能 100% 正確實現。 好在單純的詞典匹配也能滿足大部分標註的需要了，用這個標註總體來說 95% 以上的情況都是正確的（歌詞的話正確率低一些，畢竟歌詞中古語啦当て字啦訓読み這些情況很常見）。

把輸出插入我的博客

然後我的博客用 reStructuredText 語法寫，不能直接用 HTML 標籤（雖然我加了 `:html:` (inline role) 這個 行內角色 但是大量用也不方便）。這個博客一開始用 Pelican 重寫

主題的時候 我就實現了個自己的 :ruby: 行內角色 用來標發音，於是一段 sed 就能把 python-furigana 的輸出轉換成我用的 rst 語法：

```
 1  $ which clipboard Co Ci Ct
 2  clipboard: aliased to xclip -selecti
on clipboard
 3  Co: aliased to clipboard -o
 4  Ci: aliased to clipboard -i
 5  Ct () {
 6      t=$(mktemp /tmp/furigana-XXXX)
 7      python -m furigana.furigana $(Co
) | sed 's@<ruby><rb>@ :ruby:`@g;s@</rb
><rt>@|@g;s@</rt></ruby>@` @g' | sponge
 $t
 8      cat $t | tee /dev/tty | perl -pe
 'chomp if eof' | Ci
 9  }
```

上面這些 alias 在我的 .bashrc 中。有了這些之後，我只要把需要標註的日語文本放入剪切版，執行 Ct ，再粘帖結果就好了。

```
 1  $ echo "振り仮名の例" | Ci
 2  $ Ct
 3  :ruby:`振|ふ` り :ruby:`仮名|かめい` の
 :ruby:`例|れい`
```

然後所有那些歌詞上標註的假名都是這樣一句一句標註好之後，再手動校對修改的。