东方歌词翻译迁移至 sak.uy

最近几个月在这个博客发了不少歌词翻译 ~~似乎有要转型成音乐博主的趋势~~，前段时间买了个新域名 sak.uy，准备专门用来放这些东方歌曲的歌词翻译，于是分设了单独的博客「Sakuya的音乐盒」。主博客这边右侧边栏会有到音乐盒的链接。

曾经在这边的那些歌尽量保持 URL 跳转过去，新的歌词翻译会发到那边去，还想继续听歌的话请继续订阅那边的 RSS 呀。

主博客这边还是像往常一样保持记录生活点滴和技术经验好了。说道介绍技术， 有人问过我那些日语歌词上给汉字标注的假名都是我一个个手输的么？ 一开始是手输的，后来发现了不错的自动化方案，于是这里介绍一下。

首先是 python-furigana

这是个 python 写的小程序（严格说是库），可以把一段日文转换成标准的 HTML 形式的 `<ruby>` 标签的振假名(振(ふ) り 仮名(かな))。 它本身只是个方便的格式化库，实际工作是用 python-mecab 这个 binding 去查询 mecab 这个著名的日语语料分析库。要用它还得配合一些开源的 mecab 词典，这些在 `[archlinuxcn]` 都有打好的包了，直接安装：

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

装好之后用法也很直接，甚至没有 binary 直接调用 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>
```

就是提供日语作为输入，然后输出 HTML 形式的 `<ruby>` 标签而已。 像上面的例子中出现的错误（「振り仮名」完整的一个词中「仮名」意思是「平仮名」应该发音「がな」而非意为「假的人名」的「かめい」）可以看出其实标注的准确率还是有些问题的。嘛日语作为一个非常依赖上下文判断的语言， 经常日本人都会搞错某些汉字的发音，这些也不能强求机械化的算法能 100% 正确实现。 好在单纯的词典匹配也能满足大部分标注的需要了，用这个标注总体来说 95% 以上的情况都是正确的（歌词的话正确率低一些，毕竟歌词中古语啦当て字啦训読み这些情况很常见）。

把输出插入我的博客

然后我的博客用 reStructuredText 语法写，不能直接用 HTML 标签（虽然我加了 `:html:` 这个 行内角色 (inline role) 但是大量用也不方便）。这个博客一开始用 Pelican 重写

主题的时候 我就实现了个自己的 :ruby: 行内角色 用来标发音，于是一段 sed 就能把 python-furigana 的输出转换成我用的 rst 语法：

```
 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  }
```

上面这些 alias 在我的 .bashrc 中。有了这些之后，我只要把需要标注的日语文本放入剪切版，执行 Ct ，再粘帖结果就好了。

```
 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`
```

然后所有那些歌词上标注的假名都是这样一句一句标注好之后，再手动校对修改的。