コーパス内の文字出現頻度による言語間漢字変換ツールの作成

Agro Rachmatullah　小川 泰弘　外山 勝彦　(名古屋大学)

背景

- 地域による漢字の字体の違い
  - 繁体字：台湾・香港・韓国 ------ 微妙な違い
  - 簡体字：中国
  - 新字体：日本
- 共通している単語が多い ------ 日本語の「証人」 ↕ 台湾の「證人」

専門用語の翻訳に有用

言語間漢字変換ツールが必要

中国	台湾	香港	韓国	日本
骨 U+9AA8	骨 U+9AA8	骨 U+9AA8	骨 U+9AA8	骨 U+9AA8
运 U+8FD0	運 U+904B	運 U+904B	運 U+904B	運 U+904B
号 U+53F7	號 U+865F	號 U+865F	號 U+865F	号 U+53F7
证 U+8BC1	證 U+8B49	證 U+8B49	證 U+8B49	証 U+8A3C
恶 U+6076	惡 U+60E1	惡 U+60E1	惡(악) U+60E1	惡(오) U+F9B9	悪 U+60AA
卫 U+536B	衛 U+885B	衞 U+885E	衛 U+885B	衛 U+885B
叙 U+53D9	敘 U+6558	敍 U+654D	敍 U+654D	叙 U+53D9
青 U+9752	青 U+9752	青 U+9752	青 U+9751	青 U+9752
查 U+67E5	查 U+67E5	查 U+67E5	查 U+67FB	査 U+67FB
收 U+6536	收 U+6536	收 U+6536	收 U+6536	収 U+53CE
乡 U+4E61	鄉 U+9109	鄉 U+9109	鄉 U+9115	郷 U+90F7

図1：地域による字体の違い

アプローチ

- Unihanの漢字関係データから漢字の同値類を作成
- 各地域のコーパスから出現回数が高い文字を選出

Unihan

- Unicode Consortiumが提供
- 11,136字の間の関係

関係の種類	例
Simplified	來 (U+4F86) → 来 (U+6765)
Traditional	来 (U+6765) → 來 (U+4F86)
Semantic	鄉 (U+9109) ↔ 郷 (U+90F7)
SpecializedSemantic	井 (U+4E3C) ↔ 井 (U+4E95)
Compatibility	惡 (U+F9B9) → 惡 (U+60E1)
Z	説 (U+8AAC) ↔ 說 (U+8AAA)

「井」には井戸の意味

韓国では読みごとに個別の文字コード

字体のわずかな違い

同値類の生成

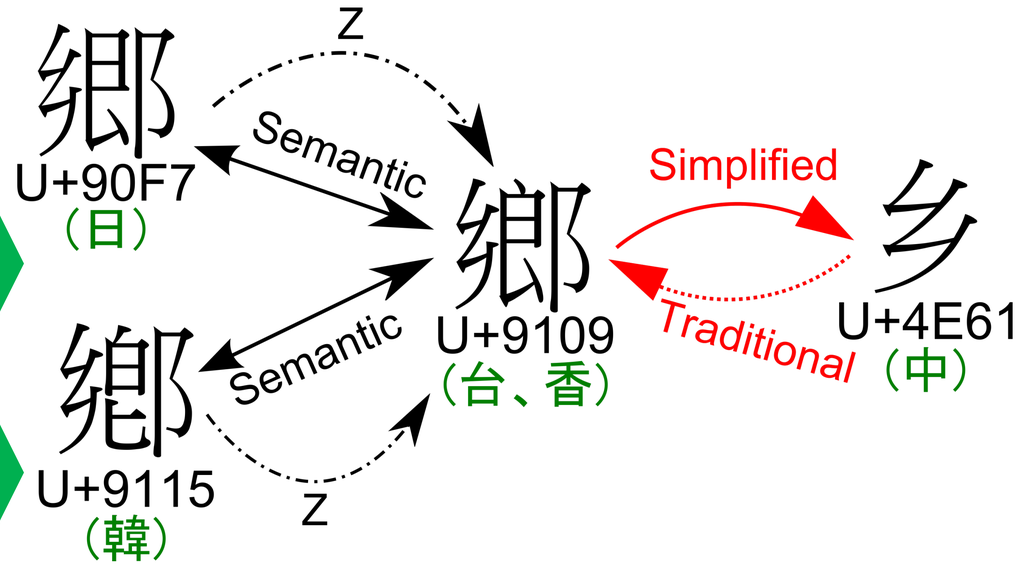

図2：漢字の同値類の例

コーパス

- CEDICT, 2012 (中国、台湾)
- 漢英法律詞彙, 1999（香港）
- 법령용어한영사전 (The Korean-English Glossary of Legal Terms), 2nd ed., 2009 (韓国)
- EDICT, 2012 (日本)

出現回数の計算

実験結果

- 4,570個の同値類を生成
- コーパスに3,170個出現 (網羅率 69.4%)
- コーパスにない例：
  - {疋,⺪, 𤴔} (部首＝漢字ではない)
  - {萹, 萹} (稀な漢字)

同値類の要素数	数	割合	例
2字	3,355個	73.4%	{負, 负}
3字	763個	16.7%	{軽, 輕, 轻}
4字	270個	5.9%	{諸, 诸, 諸, 諸}
5字	109個	2.4%	{駆, 驱, 駈, 驅, 敺}
6字～11字	73個	1.6%	{歓, 歡, 懽, 驩, 讙, 欢}

問題

- 望ましくない同値類
- 同値類から欠落する漢字
- 出現回数に差がない場合
- コーパスに出現しない場合

今後の課題

- 変換データの修正・補完

試験運用中

http://www.kl.i.is.nagoya-u.ac.jp/~agro/jitai

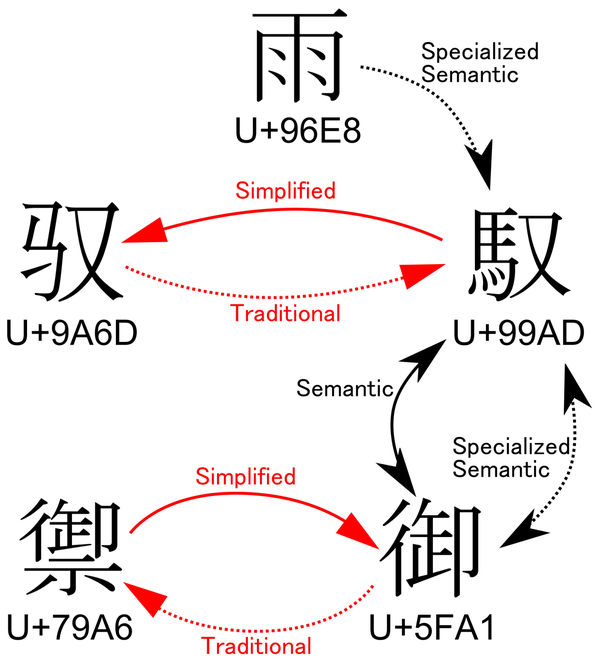

図3：望ましくない同値類