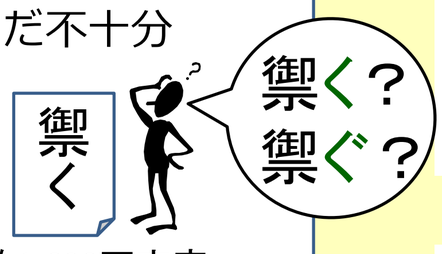
近代文語論説文を対象とした濁点の自動付与アプリケーション

岡 照晃（奈良先端科学技術大学院大学）

背景

- BCCWJのような現代語コーパスの普及に伴い，コーパスを利用した日本語研究が増加
- しかし，歴史コーパスの整備はまだ不十分
- 原因：校訂作業のコストが高い！
  1) 専門家にしか行えない
     ☞ 作業人員の確保が難しい
  2) 作業対象が膨大
     - e.g., 太陽コーパスの総文字数は約1,600万文字
     - 1人の作業者が1万6,000文字の資料に濁点付与するのに1日は必要

     ☞ 少人数では作業完了までに時間がかかる
  3) 人手の作業にミスはつきもの



研究目的

- 校訂作業を機械学習を使って完全自動化し，校訂にかかる負担を減らす
- 第１目標： 濁点付与の自動化[1, 2, 3]

☞ 文献[1,2,3]の手法を実装

校訂作業の例：濁点付与

濁点無表記文字

~Before

今や廣島は其名大に内外國に顯はれ苟も時事を談するものは同地の形勢如何を知らんと欲せさるはあらす是れ征清の大師一たひ海に航せしより　大元帥陛下大纛を此に駐め大本營となし軍務を親裁し玉ふに因てなり先つ其大勢より叙述して次第に細事に及はんとす

~After

今や廣島は其名大に内外國に顯はれ苟も時事を談ずるものは同地の形勢如何を知らんと欲せざるはあらず是れ征清の大師一たび海に航せしより　大元帥陛下大纛を此に駐め大本營となし軍務を親裁し玉ふに因てなり先づ其大勢より叙述して次第に細事に及ばんとす

「明六雑誌」第1号中の濁点無表記文字の数

音の清濁	濁音	清音	計
濁点文字（e.g., が,ざ…）	78	0	78
濁点を付けることが可能な文字（e.g., か, さ…）	368	1,273	1,641
計	446	1,273	

濁音の仮名文字のうち，83%が濁点無表記（総文字数：8,423 の4%）

インストール不要・幅広い環境で動作可能！

- Microsoft Silverlight アプリケーションとして開発
- ブラウザ上で動作するアプリケーションなので，オンライン環境さえあれば，自前のPCにインストールすることなく使用可能！
- Microsoft IE 以外のブラウザ上でもモチロン動く！
  - 対応ブラウザ： IE, Firefox, Chrome, Safari
- Mac OSにも対応

直観的な操作で難しい設定は一切ナシ！

- 濁点付与のための２種類のモードを搭載
- キーボード入力モード
  - キーボードから入力された文章に対して，リアルタイムで濁点付与を実行(左下の図)
- ファイル入出力モード
  - 入出力テキストファイルを指定するだけで，あとはアプリケーションが自動で濁点付与を実行
  - 太陽コーパスと同形式のXMLファイルにも対応
  - 濁点付与の確信度までわかる！


```
<記事 題名="太陽の発刊" 著者="大橋新太郎" 欄名="＊＊" 文体="文語" ジャンル="NDC051">
<s>　太陽の發刊　<l 位置="P001D01" /><br /></s>
<s>一陽復歸して萬象維れ新に、</s>
<s>凜々たる明治二十八年の新旭光は至渥甚深なる<敬意欠字>__</敬意欠字>皇恩の下に生等をして同胞四千餘萬の愛讀者諸君と共に紙上に相見るを得せしむ。</s>
```

自前のPCへインストールすればオフラインでも使用可能！

- インストールもカンタン！

本アプリケーションの濁点付与性能

評価用コーパス	適合率[%]	再現率[%]	F値
太陽コーパス	70.6	96.0	81.4
国民之友	95.8	98.0	96.9
明六雑誌	94.5	98.2	96.3

参考文献

[1] Teruaki Oka, Mamoru Komachi, Toshinobu Ogiso and Yuji Matsumoto (2011)「Automatic Labeling of Voiced Consonants for Morphological Analysis of Modern Japanese Literature」In *Proc.of IJCNLP 2011*, pp. 292-300.
[2] 岡照晃, 小町守, 小木曽智信, 松本裕治（2011）「機械学習による近代文語文への濁点の自動付与」情報処理学会研究報告自然言語処理研究会報告, 2011-NL-201:6, pp. 1-8.
[3] 岡照晃（2012）「統計的機械学習による歴史的資料への濁点の自動付与」第１回コーパス日本語学ワークショップ予稿集, pp. 13-22.