# 伝統的かな遣い変換システムの試作

3N－9

松尾美一　佐藤王正

（島根大学大学院理学研究科）

## 1. 序論

日本は古来中国から漢字を取り入れて独自の文字文化を作り、漢字かなまじり文を国語としてきた。この伝統的かな遣い*は戦後、簡略的な表音式のかな遣いに変わった。このかな遣いの変化は、文化の継承から問題がある[1]とされている。戦前の文学作品はもちろんであるが、今日においても文献2を始めとして、伝統的かな遣いで書かれた書物が出版されている。しかし、伝統的かな遣いの文章に馴染みがない人達は少なくない。伝統的かな遣いで書かれている文章を容易に読むには、伝統的かな遣いの文章に慣れるまで習熟する方法と、これを馴染みのある当用かな遣いの文章に変換する方法とが考えられる。前者の方法では、習熟するまで時間がかかる上に習熟する労力が負担である。しかも習熟するインセンティヴは得られにくい。後者の方法だと読者本人が習熟する必要がなく取り付きやすい。そこで、当用文に変換する方法の実現を試みたので報告する。

## 2. 実現方法

### (1)かな変換

伝統的かな遣いの文章を当用かな遣いの文章に変換するには、伝統的かな遣いをしている個所を見つけ出し、それに対応している当用かな遣いに変えることで実現できる。最も単純な変換は単に伝統的かなの全てを当用かなに変換する方法がある。しかし、これは格助詞の「は」や「を」が「わ」や「お」に変換されしまう。「私は」が「私わ」になってしまっては困る。この不都合を解消するには、文の組み立てを解体して字句を変換すべきものと、そうでないものを区別しなければならない。このために、一般的には単語辞書を用いた形態素解析が用いられる。ところが、この方式では辞書が鍵となるが、伝統的かな遣いに適した辞書を新たに作成する必要がある。しかし、これは費用の点からも労力の点からも実現性は薄い。よって簡単な方法が望まれる。そこで、文中の格助詞などの特定の字句に着目する[3]。これらを鍵語として文を句切ってからかな文字を変換すれば、正当な結果が得られる。

### (2)バックトラッキングの半自動化

上の方法で、誤って字句分けが生ずることがある。字句に鍵語が含まれていると、これが優先されるからである。例えば、「君はをかしい」という文で、格助詞の「は」と「を」を鍵語とすれば、「君/は/を/かしい」のようになるからである。こうした不都合を修正するためのバックトラッキングが必要となる。鍵語にある文字を含む単語を予め字句解析の禁止処理してから字句解析を行う。ここでは「をかし」という単語を予め字句解析禁止とする。これにより「君/は/をかしい」となる。このような単語をまとめて簡単な単語辞書を作成しておく。

### (3)システムの構成

システム構成は、資料の読み取り、字句解析、かな変換、バックトラッキングから成る。資料の読み取りは、スキャナからの資料の文字像を文字認識するもので、既存の OCR 機能を利用する。この結果を字句解析の入力とする。これは第（1）項の機能をもつ。ここでの鍵語には句読点と格助詞が含まれている。字句解析の結果を字句列として出力する。この字句列には誤りが含まれている可能性があるので、バックトラッキングを半自動化する。これは、自動的に誤りの補正の行えないものへの配慮である。かな変換は、字句列に対して伝統的かな遣い部分を当用かな遣いに変換する。

### (4)変換例

字句解析で鍵語による字句分けを行い字句列 f を作成する。上表で下線は鍵語を示す。かな変換では鍵語以外のすべての伝統的かな遣いの文字を変換し

---

* 「歴史的かな遣い」ともいう。ここでは伝統的に使用しているという意味でこうよぶ。

出力fに出力する。「彼は」の「は」は鍵語なので変換しない。「言はれて」の「は」は「わ」に変換される。

表1 変換例

| | |
|---|---|
| 入力ファイル | 彼は機敏であったと言はれてゐるが、さういふ性格はそそっかしいといふことになりかねない。 |
| 字句列ファイル | 彼/は/敏感であった/と/言はれてゐる/が/、/さういふ性格/は/そそっかしい/と/いふことになりかねない/。/ |
| 出力ファイル | 彼は機敏であったと言われているが、そういう性格はそそっかしいということになりかねない。 |

## 3. 評価実験

### (1)方法

伝統的かな遣いの書物全体を当用かな遣いに変換する。この変換結果を分析して、鍵語による字句解析方法の有効性、バックトラッキングの状況、方式上の限界等について考察する。

### (2)資料

ここでかな変換の対象とする資料は
文献2の論説文は215ページで113,540文字が含まれている。

### (3)処理結果

・切り出された字句数 18982

・伝統かなの混じり具合

切り出された字句の中に、変換の対象となる伝統かながどの程度含まれているかを表2に示す。

表2 伝統かなの混じり具合

| 伝統的かな遣い文字の数 | 字句数 | 比率(%) |
|---|---|---|
| 0 | 16193 | 85.3 |
| 1 | 2505 | 13.2 |
| 2～4 | 284 | 1.5 |
| 合計 | 18982 | |

表2より伝統かなが含まれない字句が大部分で85%である。残りの15%の字句に変換の対象となる伝統かなが含まれている。

### (4)変換の正当性

正当性を評価するために、字句解析なしにかな変換を行った結果と、字句解析付きで変換した結果を比較する。更に、字句解析に関してバックトラッキングの有無についての比較も行う。ここではバックトラッキング処理無しの場合を「字句解析付き変換1」とし、その反対にバックトラッキング処理を行った場合を「字句解析付き変換2」とした。表3において、「かな変換の正当性」とは（正当な変換個所数）／（変換すべき個所数）であり、「字句解析の正当性」は（字句解析された個所数）／（正当な字句解析個所数）を意味する。

表3 かな変換と字句解析の正当性

| (%) 変換方式 | かな変換の正当性 | 字句解析の正当性 |
|---|---|---|
| 単純変換 | 40 | ――― |
| 字句解析付き変換1 | 83 | 81.5 |
| 字句解析付き変換2 | 95 | 99.9 |

かな変換の正当率を見ると単純変換では40%で、字句解析を行うことで95%の正当率を得られることが分かる。字句解析の正当率はバックトラッキングを行うことで99.9%になった。

### (5)方式上の限界

字句解析を行ってからのかな変換の正当率は95%であり、残りの5%はうまくいかなかった。それは正しい字句解析を行っても変換間違いが起こる場合があるからである。例えば、「なりふり」という単語の「ふ」を「う」と変換してしまう間違いがある。かな変換を行うとすべての「ふ」を「う」に変換してしまうので「なりうり」となってしまう。「なりふり」は鍵語が含まれていないため、字句解析による変換可能かどうかの判別ができない。

## 4. 結論

本システムにより、伝統的かな遣いを当用かな遣いに変換することができた。これにより、伝統的かな遣いに馴染みがなくても、伝統的かな遣いの文章を簡便に読むことができるようになった。本システムでは、字句解析を行ってから、かな変換を行う方法を採用した。結果、かな変換の正当率をあげることができた。

## 参考文献

1）林武：国語の建設，講談社，(1971)．
2）土屋道雄：言論の責任，高木書房，(1998)．
3）佐藤圭正：流れ図文の性質－文記述の違いに着目した分析，情報処理学会論文誌，(1995)．