ICOT Technical Memorandum: TM-0092

TM-0092

日本語の漢字・用語の校正のための知識

石　井　　暁

January, 1985

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03) 456-3191～5
Telex ICOT J32964

Institute for New Generation Computer Technology

TM-0072

日本語の漢字・用語の校正のための知識

石井　暁

1. はじめに

文章の校正作業は計算機化が望まれ、またそれに向いた作業と考えられる。英語においては綴りの誤りの検出や正しい綴りの推定が実際に行なわれている。[2] しかし、日本語は単語がはっきり分離されていない、多くの文字を用いる等の英語との相違があるため、日本語の校正支援システムは、英語の場合とは異った機能が必要と考えられる。

実際に行なわれている校正作業を新聞社の例により調査した結果、ことば使いの校正（例えば漢字となかの使いわけや送りがなのつけ方）が計算機化が比較的容易で、有効と考えられる。[1] そこで新聞社で用いられている漢字の知識と用語の知識を計算機に扱える型式に変換し、ＤＥＣ２０上のＰｒｏｌｏｇにより校正システムの構築を試みている。

本文ではこれらの知識の型式や量について述べる。

2．　漢字の知識

漢字の知識としては朝日新聞社の用語集[4]の漢字表を用いた。これは常用漢字表[5]と内容はほとんど同じであり、約１，９４０字の漢字とその読み方の知識である。この中にない漢字の使用及びここにない読み方での使用は原則として使用しない事になっている。

述語の型式は次の様になっている。

ａｓａｈｉ＿ｚｙｏｙｏ（´亜´，［´ア´］，４２）．

引数は各々漢字、読み、用語集のこの情報が載っているページである。最後のページは、校正システムの説明機能として用いる。

行数は約４，１００行である。

３．　一般的な知識

校正作業とは直接関係しない、文章処理に一般的に必要となる知識として次の様な物を用意した。

３．１　活用語尾の知識

動詞、形容詞の活用語尾の知識で、資料［３］を基に次の様な形で述語を手作業で作成した。

```
katuyou __gobi (ka__gyou, godan, mizen, [´か´]).*

keiyousi__katuyou __gobi ( [´か´, ´ろ´], mizen ).
```

行数は約１４０で行である。

＊ ここで“か”は２バイトのＪＩＳ漢字コードを使用する。ただしＤＥＣ２０上 Pro logでは漢字というデータタイプがないため、実際は＄＋の２バイトの ascii文字を用いる。全て同様である。

３．２　助詞、助動詞の知識

これについても資料を基に手作業で次の様な述語を作成した。

（１）助詞について

```
zyosi ( [´か´], __, meisi , __).
zyosi ( [´か´], __, __, syuusi, __).
```

各引数の意味は順に助詞、それが接続する語、品詞、活用形である。
行数は約７０行である。

（２）助動詞の活用形について

```
zyodousi_katuyo ( [´う´], [´う´], syuusi).
```

“う”は助動詞の“う”の終止形であることを表わしている。
助動詞については活用が特殊な物が多いので、語幹と活用語尾の区別をせず、全ての助動詞の全ての活用した形を上記の述語で表した。
行数は約１００行である。

（３）助動詞の接続条件について

```
zyodousi_setuzoku ( [ ´さ´ , ´せ´ , ´る´ ] , _ , _ ,
          _ , godan , ng) .
zyodousi_setuzoku ( [ ´さ´ , ´せ´ , ´る´ ] , _ , _ ,
          _ , sahen , ng) .
zyodousi_setuzoku ( [ ´さ´ , ´せ´ , ´る´ ] , _ , _ ,
          mizen , _ , ok) .
```

引数の意味は順に、助動詞の終止形、その助動詞が接続する単語・品詞、活用形、活用種類、 ok が ng である。上記の例は、“させる”という助動詞は五段・サ変以外の未然形に接続することを表している。

行数は約４０行である。

３．３　単語辞書

文章の単語分けに当っては、特に表記に問題のない普通の単語辞書が必要であり、適当な辞書が得られるか否かが、単語分けの優劣に影響する。

しかし、本研究では汎用の単語辞書の入手は不可能であったため、常用漢字表[５]の“例”の欄にある語を集めて単語辞書とした。単語数は約８，５００語である。これでは語数が不足するため、４で述べる用語全てを単語辞書としても兼用している。その結果、単語数は約２７，０００語となった。

これらの資料から集めた単語は品詞、活用の種類が付いていない。そこで、これらの単語を一度紙に印刷し、その２つを手作業で記入し、入力する事により、単語辞書とした。ここで品詞は、名詞、動詞、形容詞の３種に分類した。手数を減らす目的で、活用のない語は全て名詞とした。

述語の例は次の様な物である。

```
kanzi _dict ( [ ´相´ , ´容´ , ´れ´ , ´な´ ] ,
      [ ´あ´ , ´い´ , ´い´ , ´れ´ , ´な´ ] ,
        keiyousi, _, _, _, iikae , 000090) .

kanzi _dict ( [ ´御´ ] , [ ´あ´ , お´ ] ,
        dousi , ga_gyou, godan , _ , _, 000285) .
```

```
kanzi _dict( [´鮮´, ´か´],
      [´あ´, ´ざ´, ´や´, ´か´],
       meisi , _, _, ayamari , _, 000669) .
```

ここで引数の意味は順に、語（語幹）の表記、その読みがな、品詞、活用行、活用種類、ayamari （その語が誤りの場合）、iikae （その語は他に言い換えるべきである場合）、見出し番号は後に述べる出典の述語との結合に用いるための語を一意に識別するための番号である。

各単語の出典を別の述語として扱っている。
述語の形式は次の様になっている。

```
syutten (000090, asahi_youzi , 242) .

syutten (000285, zyoyo , 028) .

syutten (000285, asahi_kana, 158) .

syutten (000669, asahi_ayamariyasue_kana, 238) .
```

各引数の意味は見出し番号（前の述べた単語辞書との結合に用いる。）、出典コード、その中でページ数である。最後のページ数はシステムの説明機能のためのものである。この述語は１つの単語が多くの出典に表われている際は、それだけ作られる。
行数は約３５，０００行である。

既に述べた様に、今回は単語辞書の作成に必ずしも充分な資料を用いる事ができなかったので、単語辞書は上記の様に比較的単純な物とした。もし多くの資料も基づくなら、資料ごとに含まれている情報が異り（例えばある資料には品詞が付いていない、またある資料には使って良いか悪いかが付いていない）、また更には内容が矛盾する事も当然あろう。優れたシステムはこれらの情報を見較べながら処理をすると考えられるので、単語辞書は大きく、複雑になろう。

4. 用語の知識

用語の知識源として、朝日新聞社の用語集[4]を選び、その中で用語についての知識の計算機化を試みた。

ただし、次の3点については扱っていない。

- 漢字の知識
  既に2で述べた通り、別のやり方で扱う。
- 誤り易い慣用句
  “愛きょうを崩す”は“相好を崩す”が“愛想をふりまく”とせよ等の知識である。用語の知識というより用語のくみ合せ方の知識であり、別の扱いが必要と考えたため、ここでは扱わない。
- 説明した文章
  記事の表記の原則等は文章（例えば“記事は、分かりやすい口語体を使う”）により表わされる事が多く、扱っていない。また皇室用語、裁判用語には詳細な説明文があるが扱っていない。送り仮名や外来語については原則の説明があるが、これも扱っていない。

ここで扱った知識は次の2種に分けられる。

（1） 正しい使い方のみの知識

次の様な項ではあることば使いが正しいということのみが記述されている。

- 熟字訓　（例：“明日”（あす）は使ってよい）　149件
- 送り仮名　（例：“受付係”，“受け付け中”などの様に送り仮名を付けよ）　8010件
- 裁判用語　（例：“上訴”，“上訴”などの用語）　42件
- 運動用語　（例：アイスホッケーで“アイシング”，“アシスト”などの用語）　1373件（登山用語については言い換え方もあるが、扱っていない）
- 難読集　（例：“和物”は“あえもの”と読む）　447件
- 外来語　（例：“アーケード”）　1277件
- 外国地名　（例：“アーカンソー川”（米国））　915件

この種の知識は独立した述語とはせず、既に3．3で述べた汎用の単語辞書に含める事によって表現している。

（２）　言いかえ方の知識

あることば使いを使ってはいけない、さらにあることば使いに言いかえよとの記述がされている。

- 表 外 訓　（例：“予め”は“あらかじめ”と言い換えよ）　８３件
- 誤り易い送り仮名　（例：“失なう”は“失う”とせよ）　６１件
- 用字用語　（例：“挨拶”は“あいさつ”とせよ，“愛玩”は“愛用”または“愛がん（用）”とせよ）　１５４３０件（ただし、同音意義語の使いわけについては扱っていない）
- 皇室用語　（例：“巡幸”，“迎啓”，“行幸”，“行啓”は使わず、“ご旅行”とせよ）　３８１件
- 選挙用語　（例：“個別訪問”は“戸別訪問”とせよ）　２７件
- 市場用語　（例：“青伝票”は“売り伝票”とせよ）　１６１件
- 誤りやすい用字（例：“悪体をつく”は“悪態をつく”とせよ）３７３件

この種の知識はつのぎの様な述語で表した。

```
asahi_youzi （［〝相〟，〝容〟，〝れ〟，〝な〟，〝い〟］，
              ［〝あ〟，〝い〟，〝い〟，〝れ〟，〝な〟，〝い〟］，
              ［〝相〟，〝入〟，〝れ〟，〝な〟，〝い〟］，
              ［〝あ〟，〝い〟，〝い〟，〝れ〟，〝な〟，〝い〟］）．

asahi_ayamariyasui_kana（
              ［〝鮮〟，〝か〟］，
              ［〝あ〟，〝ざ〟，〝や〟，〝か〟］，
              ［〝鮮〟，〝や〟，〝か〟］，
              ［〝あ〟，〝ざ〟，〝や〟，〝か〟），
```

述語名は出典を表し、各引数は使わない表記、その読み、代りに用いる表記、その読みである。

行数は約６，９００行である。

5. おわりに

日本語の校正支援の初めの段階として、Prologにより漢字や用語の校正システムの構築を試みている。漢字や用語の知識としては主に新聞社の用語集の物を用い、その型式や量について述べた。

知識は次の３種に大別される。行数はPrologで記述した際の述語の行数である。

・ 漢字について	４，１００行
・ 活用、助動詞、助詞について	３５０行
・ 単語について	６２，０００行
・ 言い換えについて	６，９００行

単語の知識はごく普通のことばの知識を増加させる事が優れたシステムのためには必要となろう。

大筆ながら本研究に当り、朝日新聞社より用語集の使用の承認を頂いた。また知識の計算機への入力、編集はシャープ（株）の援助を頂いた。関係各位に深く感謝するものである。

文　献

[1] 石井　暁：“新聞における校正・校閲の実データによる調査”、ＴＲ－０３９、新世代コンピュータ技術開発機構（1983）
これの概要が次の物になっている。
Ishii,S.：“Study of Proofreading Techniques Used at a Japanese Newspaper ”、情報処理学会第２８回（昭和５９年前期）全国大会講演論文集（II）2M-7、pp.1205 ～1206、情報処理学会（1984）

[2] 川合　慧：“英文綴り検査法”、情報処理Vol.２４、No. ４、pp.507～ 513、情報処理学会(1983)

[3] 久松他：“角川国語辞典”６３版、角川書店(1982)

[4] 朝日新聞社用語幹事編：“朝日新聞の用語の手びき”第１８刷、朝日新聞社(1984)

[5] 大蔵省印刷局編：“常用漢字表”、大蔵省印刷局（1982）