誤パターンマッチ検出による日本語文章における表記誤りの定量的一評価

2V－1

天野純一，丸山芳男，並木美太郎，高橋延匡

（東京農工大学　工学部　電子情報工学科）

1．はじめに

日本語文章の作成に仮名漢字変換を用いると、「天候」を「天侯」と書くような誤りはなくなる。しかし、「一つ」と書くべき語を「１つ」という表記で書いたり、同じ文章の中に「たとえば」と「例えば」を混在させたりすることは、仮名漢字変換を用いるだけでは防ぐことは難しい。それらを効果的に防止あるいは訂正するシステムを開発するためには、まず、それらが持つ性質を調査する必要があると考えた。そこで今回われわれは、日本語文章の中に起りうるさまざまな誤りのうち、特にこのような「表記の誤り」や「表記のゆれ」に着目し、多くの文章からそれらを検出して性質を定量的に評価する実験を行った。本稿ではその結果を報告する。

2．「表記のゆれ」に対する実験

文章中に発生する表記のゆれの傾向を調べる実験を行った。この実験の主な目的は、語によるゆれやすさの差や、文章を書く人間のゆれの起しやすさの個人差を求めることにある。実験の概要は次のとおりである。

(1) 用いた文章……書籍の原稿（校正前）で、内容は CASE の解説。英語からの翻訳。16 個の章（平均約 3.0 万文字、合計約 47.5 万文字）からなっており、これらを 6 人の翻訳者がそれぞれ二つないし三つの章を担当して執筆している。原著者は一人。

(2) 用いた語・表記……それぞれが二つの表記を持つ 12 個の語を用いた。語の選定では、われわれの学科の過去の卒業論文の校正結果などを参考にして「ゆれやすそうな語のうち、特に学術論文に多く用いられると考える語」という基準をつくり、実験者の判断で選んだ。

(3) 実験の手順……一つ一つの章に対して、各表記を文字列パターンマッチで数えあげた。KWIC を見て誤検出（最も多い表記で 2 割程度）を人手で取り除いた。次にその結果から、各執筆者について、文章の中で語が「ゆれる割合」を計算した。それは次の式で求めた。

$$\sum_{i=1}^{k}\left(\frac{\min\left(W_i^1,\ W_i^2\right)}{W_i}\times\frac{W_i}{W_1+\cdots+W_k}\right)=\frac{\sum_{i=1}^{k}\min\left(W_i^1,\ W_i^2\right)}{W_1+\cdots+W_k}$$

$$,W_i\ =\ W_i^1+W_i^2$$

ここで k は語の個数、W_i は i 番目の語の検出回数、W_i^j は i 番目の語が持つ j 番目の表記の検出回数である。結果を表１に示す。同様に各語について、その語がゆれる割合を計算した。結果を表２に示す。

表1
書籍原稿について求めた
執筆者ごとに見た表記のゆれの割合　単位 %

執筆者	章	全語に対するゆれの割合			
		執筆した各章について		執筆したすべての章について	
		平均	標準偏差	平均	標準偏差
A	前書き	0	0	0	0
	第１章	0	0		
B	第２章	0.8	0.2	0.2	0.1
	第３章	0	0		
	第14章	0	0		
C	第４章	3.5	1.0	8.7	2.1
	第５章	12.9	2.7		
	付録	11.1	3.7		
D	第６章	14.0	2.1	8.7	1.4
	第７章	2.8	0.7		
	第13章	2.7	0.6		
E	第８章	3.2	0.7	5.3	1.2
	第９章	7.1	1.7		
F	第10章	9.2	1.8	7.6	1.3
	第11章	6.8	0.9		
	第12章	4.9	0.6		

表2
書籍原稿について求めた
語ごとに見た表記のゆれの割合
単位 %

語	全執筆者に対するゆれの割合	
	平均	標準偏差
もっとも	14.8	1.8
すべて	7.1	0.6
または	0.6	0.1
たとえば	7.7	0.9
したがって	4.9	0.6
そのとき	8.6	3.2
よい	7.5	1.1
わかる	7.4	0.9
おもに	0	0
および	0	0
おもう	0	0
つぎの	0	0

Quantitative Evaluation of Mis-spellings in Japanese Text by Incorrect-pattern Matching
Jun-ichi AMANO, Yoshio MARUYAMA, Mitarou NAMIKI and Nobumasa TAKAHASHI
Department of Computer Science, Faculty of Technology, Tokyo University of Agriculture and Technology

3．「表記の誤り」に対する実験

表記の誤りの傾向を調べる実験を行った。ゆれに対する実験と同様に、語による差や人間による差を求める。また、人間による校正の前後の文章についても調査したので、人間の校正能力を測定することもできる。

(1) 用いた文章……われわれの学科の最近数年間の卒業論文及び修士論文（以下「卒修論」と呼ぶ）で、執筆者である学生には表記の基準が明示的に与えられている。また、学生は学部3年次に表記についての教育を受けている。卒修論の個数は 20 編で、それぞれ別の学生が執筆している。字数は平均約 4.2 万文字、合計約 83.6 万文字である。また、そのうちの 5 編は、教官 1 名による校正が行われた後の文章についても実験を行った。

(2) 用いた語・表記……与えられている表記の基準の中から、表記の誤りを誤パターンで表しやすい語を 18 個選んだ。誤パターンで表しにくい誤りは引用符の左右の不適合などである。

(3) 実験の手順……一つ一つの卒修論に対して正しい表記と誤った表記の検出回数を数え、誤検出を人手で取り除いた。次に、その結果から各執筆者の文章に現れる語の「誤りの割合」を計算した。それは次の式で求めた。

$$\sum_{i=1}^{k}\left(\frac{W_i^w}{W_i}\times\frac{W_i}{W_1+\cdots+W_k}\right)=\frac{\sum_{i=1}^{k}W_i^w}{W_1+\cdots+W_k}$$

$$,W_i = W_i^c+W_i^w$$

ここでW_i^c は i 番目の語の正しい表記の検出回数、W_i^w は誤った表記の検出回数である。同様に、各語についての、その語がゆれる割合を計算した。校正前の卒修論について、執筆者ごとの誤りの割合を表3に、語ごとの誤りの割合を表4に示す。校正後のものについても同様に表5と表6に示す。

4．考察

今回の実験によって、次のことがわかった。

(1) 「ゆれやすさ」や「誤りやすさ」は語による差が大きい。また、語の文章中での出現頻度とゆれやすさ・誤りやすさの間には特に相関は見られなかった。したがって、「よく用いる語はゆれにくい」というようなことは言えない。

(2) 人間による差も大きい。表記について同じ教育を受けたにもかかわらず、まったく表記誤りが見られない人や、対象とした語を 3 割近く誤る人がいる。

(3) 人間ではすべての誤りは取りきれない。校正者の教官は表記に対する深い見識と豊富な経験を持っているが、それでも取りきれないことがある。

表3
卒修論について求めた執筆者ごとに見た表記の誤りの割合
単位 %

執筆者	全語に対する誤りの割合 平均	標準偏差
A	3.4	0.6
B	12.1	3.7
C	29.7	6.0
D	10.4	2.4
E	5.1	1.0
F	18.5	4.8
G	1.8	0.4
H	9.3	1.6
I	16.2	2.7
J	1.8	0.6
K	0.8	0.2
L	0.6	0.2
M	0.7	0.1
N	0.6	0.2
O	4.2	0.7
P	2.1	0.4
Q	0	0
R	0.3	0.1
S	0.7	0.2
T	19.1	3.1

表4
卒修論について求めた語ごとに見た表記の誤りの割合
単位 %

語	全執筆者に対する誤りの割合 平均	標準偏差
繰返し	100	27.0
繰り返す	21.3	9.2
それぞれ	12.9	2.5
一つ	24.1	2.0
ずつ	6.3	0.9
すべて	9.7	0.7
だけ	10.0	0.8
とおり	25.3	3.5
数十	17.1	7.7
ような	0.6	0.1
わたる	0	0
したがって	3.6	0.4
しやすい	1.0	4.1
できる	0.2	0.0
幅	11.3	3.3
ため	0.0	0.0
ごと	1.2	0.3
終わる	100	38.4

表5
校正後の卒修論について求めた執筆者ごとにみた表記の誤りの割合
単位 %

執筆者	全語に対する誤りの割合 平均	標準偏差
K	0	0
L	0	0
M	0.7	0.1
N	0	0
O	1.1	0.2

表6
校正後の卒修論について求めた語ごとに見た表記の誤りの割合
単位 %

語	全執筆者に対する誤りの割合 平均	標準偏差
繰返し	100	0
繰り返す	0	0
それぞれ	0	0
一つ	0	0
ずつ	0	0
すべて	0	0
だけ	0.3	0.7
とおり	0	0
数十	0	0
ような	0	0
わたる	0	0
したがって	0	0
しやすい	0	0
できる	0	0
幅	0	0
ため	0	0
ごと	0	0
終わる	100	0