大量言語処理におけるエラーと対策

斎藤秀紀，鷺岡昭夫，中野　洋，米田正人

1. 目的

大量言語処理において起るエラーは、小量のデータの処理にくらべて、エラーの質や量、およびその対策において異なる。すなわち、(1)単純作業が多くなるため、ケアレス・ミスをなくせない。(2)データの管理にエラーがおこることがある。(3)メンバーや時期によって、了解違いがおこることがある。(4)やりなおしがきかない。などがその主な点である。

以下に、我々の経験と調査によって得られた、エラーの種類とその対策について報告し、今後の大量言語処理の参考に供したい。

2. 語彙調査システムにおいて起るエラーの種類・原因・対策

語彙調査において起るエラーの種類とその原因・対策について、水谷静夫氏の方法(「語彙調査に生ずる狂いの種類・原因・対策」国語研究所年報5 1954)により、表1のように整理した。

表1. 教科書調査システム(案)におけるエラーの種類と原因(一部)

作業段階	番号	狂いの種類	発見の困難さ	発生頻度	原因	対策
単位切り	041	単位切りミス	やや困難	多	作業者の指導，作業者の不注意	検査
単位切り検	051	検査もれ	困難	多	作業者の不注意	整理表
査・修正	052	修正もれ	〃	少	〃，システム	志気の高揚
	053	修正エラー	〃	中	〃，作業の指導	規則の整備，教育
清書	0611	清書ミス	容易	多	〃，〃	校正

一般に、語彙調査システムは以下のような作業工程をもっている。
(1)対象の指定 (2)単位切り (3)清書 (4)情報つけ(読みがな・代表形・語種・品詞・活用・単位・出典など) (5)パンチ (6)機械処理 (7)出力整理

人間だけの作業において起るエラーの種類は、「まちがい・脱落・重複」の3種類に分けることができる。

また、その原因として、次の7種類が考えられる。
(1)作業者の不注意 (2)作業の指導が不徹底 (3)作業負担が大きい (4)システムの不備(作業規則があいまい，データがうまく流れない) (5)オペレートシステムの不備 (6)管理不充分(作業者や仕事の管理，データの管理) (7)志気の低下

したがって、それぞれの工程において、次のような対策が必要である。
(1)チェックシステムの確立 (2)作業者の教育 (3)作業者の負担の軽減 (4)作業の整理表・受け渡し表を充実する。(5)オペレートジョブ・ランブックを充実する。(6)志気の高揚をはかる。(調査の目的・意義，受け持ち作業の分担などを確認する。

しかし、このような対策をほどこしても、エラーを完全になくすことは不可能である。重要なことは、いかにエラーを発見するかである。調査データを用いたKWICの作成や各種の分析により、かなり小さなエラーも発見できる。(以上 文責中野)

I　文脈つき索引作成システムにおけるエラーデータとその校正状況

国立国語研究所言語計量部第一研究室では、「漱石・鷗外の用語の研究」の中で、文学作品の文脈つき用語総索引を計算機により作成した。このシステムでは、新聞語彙調査システムと異なり、パンチ用原稿の校正，印字シートの校正を省き、機械チェックによるエラー情報を利用した一度の校正ですませた。その他はミニ・KWICによる校正だけである。現在、漢字プリンタによる校正シートの出力を組み込んでおり、機械チェックのためだけのランはなくなっている。

以下に述べるのは、機械によるチェックの精度をはかるための調査である。

【索引システムの作業工程】

(1) 調査対象の指定　(2) 単位切り　(3) 清書　(4) 読みがなつけ　(5) 単位情報つけ　(6) 語種・品詞・活用情報つけ　(7) パンチ　(8) 機械チェック　(9) 漢テレ印字　(10) 校正　(11) 修正パンチ　(12) 機械処理　(13) 出力整理

【チェックの性質】

今回の機械によるチェックは、明らかなエラーとともに、エラーかどうかわからないがあやしいもの，珍しいデータもエラー情報をつけた。この判断は、校正作業における人間にゆだねられる。

【機械チェックの内容】

(1) 桁ずれチェック　(2) フォーマットチェック　(3) 品詞連続チェック　(4) 語種・品詞組み合わせチェック　(5) 活用情報チェック　(6) 語形チェック　(7) 単位情報チェック　(8) 頁・行情報チェック

【機械チェックと校正状況】

前述工程8におけるエラー情報は、フォーマットチェック・品詞連続チェック　語種品詞組み合わせチェック・語形チェックの4種類である。以下に、森鷗外の作品「興山拓得」（延べ4066語）の校正状況を示す。

表1.　工程9における校正状況

校正された箇所	チェック情報 あり	チェック情報 なし	計
フォーマット	38	12	50
単位情報	11	34	45
見出し語	22	31	53
よみ	7	25	32
語種情報	2	12	14
品詞情報	15	55	70
活用情報	10	37	47
計	105	206	311

表2.　エラーの原因と種類

種類＼原因	フォーマット パンチ	清書	計	単位 パンチ	清書	計	見出し パンチ	清書	計
おち	4	0	4	1	0	1	24	15	39
ばけ	41	3	44	7	37	44	0	0	0
いり	1	1	2	0	0	0	0	14	14
	よみ パンチ	清書	計	語種 パンチ	清書	計	品詞 パンチ	清書	計
おち	8	7	15	0	0	0	3	1	4
ばけ	5	9	14	6	7	13	28	37	65
いり	2	1	3	1	0	1	1	0	1

表3.　ばけのエラーと機械チェック

	チェック情報 あり	なし
単位エラー	34	10
語種エラー	12	1
品詞エラー	50	15
計	96	26

	活用 パンチ	清書	計
おち	2	2	4
ばけ	7	8	15
いり	21	7	28

表1に示すように、校正された箇所311のうち、その約1/3の105箇所に機械によるチェック情報がついていた。一応の効果はあったと考えられる。

フォーマットエラーの「ばけ」がパンチミスで41もあるのは、漢テレ・シフトキーの押しまちがいがほとんどだと思われる。見出しや読みのエラーが多いのは単なるミスとともに、規則の不徹底のせいもある。活用情報のいりが多いのは、前の調査の影響である。これらのエラーは作業者の訓練である程度なくせるが、単位・語種・品詞で「ばけ」が多いのは、情報つけ自身のむつかしさによる。(須賀)

Ⅱ・高等学校の教科書の用語用字調査における，プレエディットでの作業ミス

0 調査におけるプレエディットのあらまし

調査対象の文章抽出→単位切り→清書→付加情報記入（→ＰＴパンチ）

1-1 調査の対象

- 社会・理科・数学の三教科のうちの各科目につき一冊ずつ。すなわち，「倫理社会」「地理Ｂ」「日本史」「世界史」「政治経済」「生物Ⅰ」「化学Ⅰ」「物理Ⅰ」「地学Ⅰ」「数学Ⅰ」の十科目。
- 見出し，および本文。図・表・注・問・写真を除く。
- 全数調査。（推定のべ語数：40万Ｗ単位／60万Ｍ単位）

1-2 対象部分抽出のミス

- ミスは比較的少なく，ケアレスミスがほとんど。
- 発見は容易。

2-1 この調査に用いた単位

(0) この調査では，文の最小の構成要素としての長い単位（Ｗ単位）と，語を構成する要素としての短い単位（Ｍ単位）の二種類を用いた。作業は，まず赤鉛筆でＷ単位に分割し，その作業が済んだものに，黒鉛筆でＭ単位の切れ目を入れるという二段階方式を採用した。（本プリントでは，Ｗ単位の切れ目を／で，Ｍ単位の切れ目を／で示す）

(1) Ｗ単位のあらまし

Ⓐ 独立形態（→Ｂ）に付属して，文法的な関係を示す語（助辞）は一語一Ｗ単位とする。（助辞は別途作成のリストによる）

ex. が　は　を　のに　から　だ（だった，だろう）　です　etc.

Ⓑ 文構成上の単位（文節）の中で，意味的・形態的に独立している部分。すなわち，文節から助辞（→Ａ）を除いた部分。

ex. 花　私　静か　美しい　少し　行く　行った　食べない　見て　行けば　行くなら　行こう　etc.

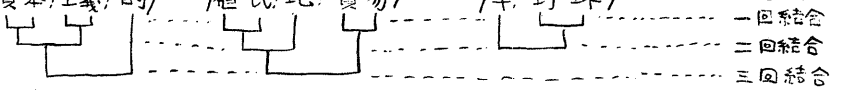
(2) Ｍ単位のあらまし

Ⓐ 助辞は一語一単位。すなわち助辞はＷ単位＝Ｍ単位。ただし，「だった」「だろう」「でした」等は，「た」「う」等を切り離す。

Ⓑ 助辞以外の語（Ｗ単位）は，つぎのように細分する。

(イ) 漢語は二要素（漢字二字）の結合，すなわち一回結合をしているものを一単位とする。他の要素と結合した混種語中に漢語要素どうしが結合しているものもこの原則をあてはめる。

ex. /資本/主義/的/　/植民/地/貿易/　/常/野球/　etc.



(ロ) 和語・外来語・固有名詞（人名，地名，国名，地形名など）は，一要素を一Ｍ単位とする。混種語の中でも同様に処理する

ex. /結び/ついて/　/お/父/さん/　/ウーマン/リブ/　/フランス/人/　etc.

(ハ) 一字漢字に「する」「す」「ずる」「じる」，外来語要素に「る」等が付いてできた語は切り離さない。

ex. /愛する/　/愛した/　/信ずる/　/信じない/　/デモる/　etc.

(C) 数は一字一単位とする（数を表わすことに主眼のない数字，たとえば「一体全体」「十字」の「一」「十」などは(B)で処理）。

ex. /1/9/5/7年/　　/二/百/十/日/　　etc.

(3)単位切りの例

・/ピルビン/酸/は/脱/炭酸/され/て/,/活性/化/され/た/酢酸/に/なっ/た/のち/,/まず/オキサロ/酢酸/と/反応/し/て/クエン/酸/と/なり/,/脱/水素/酵素/の/はたらき/で/水素/を/失い/,/脱/炭酸/酵素/の/はたらき/で/二/酸化/炭素/を/失う/反応/を/くり/かえす/うち/,/図/4.2/の/よう/に/オキサロ/酢酸/に/もどる/。/

・/新/航路/の/探求/で/主役/を/演じ/た/の/は/,/イスラム/教徒/と/戦い/ながら/国内/統一/を/強化/し/て/い/た/ポルトガル/と/イスパニア/で/あっ/た/。/西国/とも/国王/の/援助/の/もと/に/探険/が/続け/られ/た/。/

2-2 単位切りミス（カッコ内は正しい切り方）

(1) W単位処理でのミス

Ⓐ 切りすぎ

- サ変動詞語幹を切ったもの

ex. /固定/し/て/（/固定/し/て/）　　/区別/できる/（/区別/できる/）

- 助辞でない助詞を切ったもの

ex. /細胞/ごと/に/（/細胞/ごと/に/）　　/戦い/ながら/（/戦い/ながら/）

- 一語の助辞を分割したもの

ex. /できる/の/で/,/（/できる/ので/,/）　[cf. /できる/の/で/ある/]

- その他

ex. /バルト/海/-/北海/貿易/（/バルト/海/-/北海/貿易/）

Ⓑ 切り忘れ

- 助辞を切り忘れたもの

ex. /今や/（/今/や/）　/精一杯だっ/た/（/精一杯/だっ/た/）　/立場/から/（/立場/から/）　/攻撃/など/（/攻撃/など/）

- 接続詞を切り忘れたもの

ex /前足/(/つまり/手/)/（/前/足/(/つまり/手/)/）

- 修飾語と被修飾語を切り忘れたもの

ex. /高い/方/は/（/高い/方/は/）

- 二語の助辞を一語とまちがえたもの

ex. /作ら/れる/ので/ある/（/作ら/れる/の/で/ある/）[cf. /作ら/れる/ので/,/]

(2) M単位処理でのミス

Ⓐ 切りすぎ

- 全体で副詞となっているものを切ったもの

ex. /はじめ/て/（/はじめて/）　/かね/て/（/かねて/）　/つい/で/（/ついで/）

- 活用語尾またはそれ相当部分を切ったもの

ex. /わかっ/て/いる/。/（/わかって/いる/。/）　/現/れ/（/現れ/）

- 数でないものを数とみたもの

ex. /精/一杯/（/精一杯/）　/不/十分/（/不十分/）

- その他

ex. /広/がる/（/広がる/）[cf. /寒/がる/,　/広/まる/]

Ⓑ 切り忘れ

- 要素を分け忘れたもの

ex. /前足/ (/前/足/)　/立場/(/立/場/)　/しくみ/(/し/くみ/)
/ひだ状/(/ひだ/状/)　/まもなく/(/まも/たく/)　/結ばれる/(/結ば/れる/)　/作られる/(/作ら/れる/)

● 漢語の語構成を理解していないもの
ex. /三部会/(/三/部会/)　/潜水艦戦/(/潜水/艦/戦/)　/連合国軍/(/連合/国/軍/)　/炭水化物/(/炭水/化/物/)

(C) 切る位置のちがうもの
● 和語系のもの
ex. /めざして/(/めざし/て/)　/つくられ/(/つく/られ/)
● 漢語系のもの
ex. /膝蓋腱/(/膝蓋/腱/)　/同盟国内部/(/同盟/国/内部/)

(3) 単位切りミスの出現状況
〔世界史の例〕

	W 単位			M 単位		
ページ	総語数	エラー	エラー率	総語数	エラー	エラー率
2	201	2	0.99	250	2	0.80
3	245	2	0.81	325	5	1.53
5	276	1	0.36	396	7	1.76
㋐(2·3·5)	(722)	(5)	(0.69)	(971)	(14)	(1.44)
101	226	0	0.00	305	1	0.32
102	254	1	0.39	346	0	0.00
103	168	0	0.00	218	0	0.00
㋑(101~103)	(648)	(1)	(0.15)	(869)	(1)	(0.11)
202	186	0	0.00	260	2	0.76
203	216	1	0.46	313	2	0.63
204	235	4	1.70	336	7	2.08
㋒(201~204)	(637)	(5)	(0.78)	(909)	(11)	(1.21)
《㋐·㋑·㋒》	《2007》	《11》	《0.54》	《2687》	《23》	《0.85》

3-1　清書と清書ミス
- 清書は一行(20字用)ごとに一M単位を記入する。
- 清書ミス……誤字，脱字，行とばし等
- ミスの発見，修正は容易。

4-1　付加情報(清書済原稿用紙に記入)
(1) 単位情報…W単位先頭に立つもの(M単位)にW，それ以外はMと記入。
(2) 助辞情報…助辞には，J を付ける。
(3) ルビ情報…本文に読みを示すルビのある場合，R を付ける。
(4) 読みがな情報…漢字を含む語にはひらがなで読み方を示す。(平がなで示す。)
(5) 代表形情報…語形の変化した語(活用語など)は，もとの形を～印のあとに
○ (1)は語の先頭，(2)～(5)は＜　＞内に入れる。(2)～(4)の間は／で区切る。

4-2　付加情報のミスの一例(代表形よみがな情報のミス)
/死んで/　/いりくんだ/ の「で」「だ」に代表形情報 <へて><へた> みたいミス。
(担当　鷲田)

III 高等学校の教科書における用語・用字調査システム

1. 計算機処理手順の概略

漢テレでパンチされた紙テープを読み込み，種々のチェック・ルーチンを通り，修正の為のデータの印字を漢字プリンターで行う。その印字用紙をみて，人手で修正を行った後，正しいデータを作成し，クイック，語彙表等を作成する。

2. データ（1レコード）のフォーマット

代表形インデックス	配列情報	単位	出現形	区切り記号	情報	区切り記号	ルビ情報	区切り記号	出現形よみ	区切り記号	代表形よみ	区切り記号	丁度エラー	フォーマットエラー	データエラー	教科書名	ページ番号	段落番号	文番号	語番号	文種情報	%
12	20	2	30	2	2	2	2	2	30	2	30	2	2	2	2	6	8	6	6	10	2	2

←項目

←バイト数

3. チェックの種類

大別して，漢テレチェック，フォーマットチェック，データチェックに分かれる。各チェックでMT上にエラーフラグをたてるが，その内容の一覧を以下にあげる。

漢テレ・エラー	J	J字チェックにかかった（軸脱落）。	⑦
データ・エラー	D	データ中に(SP)あり。	①
	X	先頭文字エラー。	②
	A	付加情報の先頭が／か＞。	③
	E	クギリなしエラー。	③
	C	←←連続，←の次の10データ以上経て←がある。	④
	F	→の次の10データ経たのに←がない。	④
	G	→→連続。	⑤
	H	よみがない	⑦
	I	単位切りエラー，数字・記号連続。	⑧
	J	よみに何かが入っている。	⑨
	K	漢テレ以外の文字がよみにある。	⑩
	L	よみにひらがな以外の文字がある。	⑩
	M	代表形よみに漢テレ以外の文字がある。	⑪
	N	代表形よみにひらがな以外の文字がある。	⑪
	O	テーブルにあるのに助辞情報がない。	⑫
	P	テーブルにないのに助辞情報がある。	⑬
	Q	ページがだぶっている。	⑭
	R	ページがとんでいる。	⑭
	S	出現形の先頭に一，ー，ゝ，ゞ，っ，が来た。	⑥
フォーマット・エラー	Y	情報データの後に何かが入っている。	①
	Z	ページ，数字三桁の後が＞でない。	①
	B	＜，があるのに＞がない。	②

一つのデータで二つ以上のエラーが起ったとき，漢テレ・エラー，データ・エラー，フォーマット・エラーは共存できるが，同一エラー内では数字の大きいものが優先される。

4. データの印字

人手による校正の為の資料として以下の三つの印字を行う。

(1) ミニクイックの印字

教科書名，単位，代表形よみ，助詞情報，出典情報（ページ番号，段落番号，文番号，語番号），文脈を印字する。

〔 世界史 〕　　M単位

代表形	情報 頁	段落	文	語		出現形
ちゅうか	243	00	00	015	29か国の代表は，	中華 人民共和国政府の提唱する平
ちゅうごく	037	01	01	007	ドシナ半島の東北部	中国 の政治・文化の影響下にあっ
ちゅうごく	037	02	01	012	て，インド東海岸と	中国 の南部沿岸地方を結ぶ貿耕フ
ちゅうごく	039	01	01	013	のうえで確認できる	中国 史上最古の王朝は◇歌安であ
ちゅうごく	187	02	03	015	間に，列強は争って	中国 から鉄道敷設権・鉱山採掘権
ちゅうごく	187	02	05	009	れて，米西戦争以後	中国 市場に深い経済的関心をもつ

(2) 原文イメージの印字

教科書の原文を単位切りしたものを印字する。／はM単位の切れ目を，㊞はW単位の切れ目をあらわす。

● 民主／政治 の 発達● @ 民主／政治 の 実現 を 求める 強い 運動 が 始ま っ／た のは ， 近代／初期 の 絶対／王政 に 対する 人／人 の 激しい 抵抗 が 起こ っ／た 以後 の こと で ある ．井 商業 や 工業 の 発展 に よっ／て ， 資本 や 財産 を ふやし／た 人／人 は ， 政治 に 対する 発言／権 を 強く 要求／し／始め／た ．井 かれ／ら は ， 古い 支配／者 に 向かっ／て ， 人間 が 生まれ／ながら に 有 し／て いる 天賦 の 基本／権 を ， 法 に よっ／て 保障／し ， 君主 と いえ／ども

(3) 清書イメージの印字

清書原稿のイメージの形式で印字を行う。

単位	出現形	情報 ルビ	読み	代表形	頁	段落	文	語	修正文種	J字検査	データ検査	型検査㊀
W	秩序	〈K.	.ちつ・じょ	・	>001	0102	012		曲			
W	を	〈J.	.	・	>001	0102	013		曲			
W	もっ	〈 .	.	・	>001	0102	014		曲		誤読み文字	
M	た	.	.	・	.001	0102	015		曲			
W	社会	〈K.	.しゃ・かい	・	>001	0102	016		曲			

5. チェックと修正

4.で出力された資料をもとに，(1)により見出し語のチェックを，(2)により各種情報のチェックを，(3)により総合的なチェックを行う。

修正方式は語単位の修正とページ単位の修正の二種類が用意されている。

6. オペレート管理，データ管理

作業の進行表を保存することによって，および，ジョブの各ステップで入力データ数，出力データ数，エラーデータ数，オペレータ名，作業実施年月日等をライン・プリンタに出力し，それを保存することによって，オペレートのとばし，重複が起らない様に注意している。

また，保存用磁気テープのラベルの書式はすべて統一されている。

〔担当．米田〕

Ⅳ. 漢字テレタイプの誤りの傾向について

1. 序

新聞の用語調査で作成されたデータに関し，漢字テレタイプのパンチエラーの傾向を調べることを目的とする。対象としたデータは，昭和40年度発行，朝日新聞朝刊1～6月の半期分である。データの収集方法については，長単位データおよび中間段階で再入力される短単位処理済みデータとの照合不一致となった単語または文字を対象とした。これらのデータは，長単位データの短単位処理で修正されたもの，逆に再入力時点での新らたな誤りデータも含まれる。調査データ総数は2833仆から，カッコ，長音等の記号類，かな文字，単位切りに関する誤りは除外し，漢字データ850仆から，さらに340仆を抽出したものである。データの分類については，字形類似によると思われる誤り，同音または類似した意味による当て字，キー選択時点でのシフトエラーに限定した。なお，データは新聞総量調査一紙一年分の誤りデータ（長・短位ともに統数1117ょ仆）の一部をなしプレ調査としての性格を持つ。

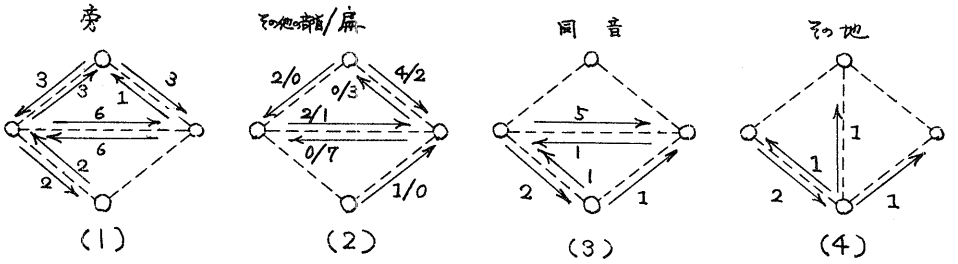
2. 誤りデータの傾向

従来，漢字テレタイプの文字配列の方法として，部首順配列，代表音訓による50音順配列，また市町村名等，ある特定の場合に多く使用される文字ごとに一つのキーにまとめて配当する方式と，種々の方法がとられてきた。しかし，いずれの配列にしても，確定した方式とは言えず，各方式がデータまたはユーザの業種におうじて使用されているのが現状である。本稿では，これら漢字テレタイプの文字配列順序は直接パンチされたデータの順と関係あるものと想定し，回研における漢字処理データをもとに，字形と誤りの関係について概観する。使用した漢字テレタイプは収容漢字数2110字文字配列は50音順である。また，一つのキーには4文字ずつ配列されている。

通常，50音順に配列された場合の利点は，漢字に同音文字があるにせよ，比較的容易に文字を捜し出すことが可能である。しかし，反面，機械的に文字を配当していくことは，同一キー内の同音漢字が集中的に配当される確率が高くなる。これは，漢字の性質上当然であるが，同音であると同時に「旁」が同形である確率もまた高くなる。キーパンチャーにとっては，同音でありかつ字形の類似した複数の漢字を選択していくことは二重のハンディを負うことになり，誤りの発生原因の大きな部分となってくる。

同一キー内でシフトエラーとなった文字／その他

艦	鑑	戦鑑 旗鑑	未	末	末経験 末定 八月末
織	職	組職 川島職	自	目	日米目 四度目 丸石目
険	検	候眠 保検	士	土	力土 三好富土彦
壊	懐	半懐し 崩懐し	夫	天	忠天 南◇期夫谷
積	績	面績	丈	文	新井紳ノ文
僥	暁	四百暁円	圧	庄	強庄
稼	嫁	弁護士嫁業			
唱	昌	少年合昌団	簿	薄	電話番号薄 肉簿
噌	曽	味曽	宮	宮	ランブイエ宮大理石
諸	緒	西側緒国	幕	募	三暮 募切れ
砲	疱	大疱	施	旋	布旋明 福利厚生旋設



	同音				異音				その他	計
キー間	旁	扁	その他部首	その他	旁	扁	その他部首	その他		
	55	7	16	51	9	31	39	24	41	273
	20 %	30%	19 %		3 %	11 %	14 %	9 %	15 %	100 %
同一キー	26	13	9	10					5	63
	41 %	21 %	14 %	16 %					8 %	100 %

表及びグラフの内容から，キー間での誤り，同一キー内での誤り，については相互の単純比較は無理としても，同一キー内での「旁」の同形文字に特に高くシフトエラーが発生している。シフトの選択方向としては，左→右シフトの場合が多い。これらの現象は，キー配列が如音順配列の場合の誤りのパターンを示していると思われるが，より明確な説明を与えるためには，他の配列方法による漢字テレタイプで作成されたデータとの比較が必要となろう。部首順に配列された場合，明らかに，同一キー内での同形の「旁」は少ない。また，「扁」単位でブロックされていることは，ブロック単位は如音順の場合より大きくなり，視覚上からも誤り率は少なくなるものと思われる。しかし，キー内の文字位置は少なくても誤りの頻度が高い方向に対しては，旁，扁，共に文字の構成要素には注意することが望ましい。これは，グラフ上で示される誤りの位置関係から，字形に影響を受けたと思われるパターン（1，2）と単なる漢字音によって生じた誤りパターン（3，4）とは明らかに異った型を示しているからである。パンチミスの要因としては，パンチャーの原稿と目の動きの関係，原稿の書き方，データとなる資料の種類等が考えられるが，上記のパターンの場合，字形の類似性から生じたものと考えるのが妥当であろう。

3. 結び

以上，誤りの特徴と傾向について概観してきたが，この種の誤りについては，パターンがほぼ限られているところから，今後，コンピュータによる自動修正への可能性が期待できよう。本稿では，誤りデータの中で，かな文字，記号については省略したが，これらの誤りは総データの半分近くになっている。これらは，漢字データと共に正確なデータ作りには無視できない部分である。また，部首順列の漢字テレタイプの場合の誤り比較も，これら誤りデータの特徴をつかむ上でも重要と思われる。今後，この種の調査について，語彙調査で拾われたエラーデータについて，単位切り，データの中間再入力での誤り率，かな文字，特に拗音，大文字のとりちがえ等についても明らかにしてゆきたい。

（担当 斎藤）