ニュース用語の分類と誤入力訂正への適用†

相沢輝昭††* 栗田泰市郎†††

我々は先に単語中の1字の置換誤りを訂正する手法を提案し良い結果を得たが[2]，使用する用語辞書の語数と訂正能力との関係が明確でなかった．そこで本論文では，放送ニュース用語約3万を代表的なニュース分野に分類し，それによって得られた1,500～3万語の大小8種の辞書による訂正処理実験を行い，辞書の諸性質と訂正能力との関係を定量的に調べた．その結果，特に，訂正能力は辞書の語数の対数にほぼ比例して低下するが，語数が2万以下ならば90%以上の訂正率が確保できることが分かった．放送ニュース用語の場合には，政治，経済，社会，市民生活，事件・事故，文化，科学，スポーツ，国際，という代表的な9分野への分類によって，分野別辞書の語数を14,000～2万に抑えることが可能であり，したがって，これら分野別辞書を切り換えて使うことにより，放送用語全体にわたって高い訂正能力を達成できる見通しが得られた．

1. まえがき

音声入力やOCR入力の際に生じる認識誤りを訂正して総合的な認識率を高める研究が古くから行われてきた．文献1)，2)には1983年ごろまでの内外の関連研究が一通り概観されているが，それ以降も活発な研究が続けられている[3)~10)]．特に最近は'音声入力ワードプロセッサ'の実現を具体目標にしたものが目につく．

我々も先に，音声入力の後処理用を主目的とした，日本語単語の置換誤りの訂正法を提案した[2)]．その主な特徴は次のとおりである．

1) 誤字検出には，日本語のカナ2文字連続の頻度分布の偏り（特に頻度0の2文字連続）を利用する．すなわち，入力文字列に頻度0の2文字連続がある場合，それらを誤字候補とする．そのような2文字が無い場合は全文字を誤字候補とする．このような誤字の2段階検出は，誤字検出にとって極めて有効であることが分かっている[2)]．なお，2文字連続の頻度分布は，使用する用語辞書の全用語のカナ見出しを用いて前もって作っておく．

2) 誤字訂正には'正字候補表'を用いる．この表は，各文字ごとに，それが誤字候補となった場合の正字の候補を，入力装置の誤りの傾向等を加味して前もって作っておくものである．どの正字から置換していくかは，入力装置の誤字発生確率と上記の2文字連続の頻度分布に基づく'評価関数'に従って決める．この正字候補表を差し換えることによって，音声入力や鍵盤入力など様々のタイプの入力に簡単に対応できるところが本方式の利点の一つである．

3) 処理系全体はコンパクトであり，処理速度も比較的高速である．

このような誤入力訂正処理において，用語辞書は種種の形で使われる．まず，2文字連続の頻度分布は，用語辞書のカナ見出しに基づいて作成される．また，入力文に誤りがあるか否かのチェックや，誤字訂正が成功したか否かのチェックも，辞書のカナ見出しとの照合によって行われる．

このように，用語辞書は処理全般にわたって基本的役割を果たすわけであるが，先の報告[2)]では5,704語から成る比較的小規模のもの1種類だけによっていた．その限りでは約90%という高い訂正率が得られていたが，実用上からはこの程度の辞書では小さすぎる．しかし，辞書を大きくすると，表記上で似通った用語が増えたり，我々の訂正方式にとって重要な役割を果たす'頻度0の2文字連続'（上記1）参照）が減ったりして，状況は不利になる．実用的な広範囲の文章に適応できて，しかも訂正能力を劣化させない一つの方法は，用語全体を分野によって分類し，コンパクトな分野別辞書を切り換えて使うことであろう．

本論文では，先に我々が提案した誤入力訂正処理手法[2)]が，このような考え方の下での実用規模の辞書に対しても有効性を失わないことを，放送ニュース文を対象として実験的に確認する．すなわち，まず2章で約3万のニュース用語のニュース分野別分類の方法と一般的な分類結果を述べる．次に3章で，代表的な分野別辞書の性質を訂正処理の立場から調べ，その上で

† Classification of News Words with an Application to Correcting Errors on Japanese Words Input by TERUAKI AIZAWA (Video Engineering and Data Processing Research Division, NHK Science and Technical Research Laboratories) and TAIICHIRO KURITA (Advanced Television Systems Research Division, NHK Science and Technical Research Laboratories).

†† NHK 放送技術研究所画像研究部
††† NHK 放送技術研究所テレビ方式研究部
* 現在 (株)ATR 自動翻訳電話研究所

これら種々の辞書による訂正実験を行って辞書の影響を定量的に調べる．結果的に，我々の訂正方式は現実的規模の辞書に対しても高い訂正率を持つことが示される．

2. ニュース用語の選定と分類付け

我々は放送ニュース文を主対象としているので，それに適用できる実用規模のニュース分野別辞書の作成を検討した．

放送用ニュース用語の選定のために用いた基本データは NHK 編『新用字用語辞典』[11] である．これに収容されている語は片カナ表記の外来語を除く約3万語で，序文によれば「見出し語の取捨選択，表記の原則の決定にあたっては，放送のことばの向上を目ざして毎月開いている NHK 放送用語委員会の取り決めや意見を十分に反映させるとともに，広く放送の現場部門に意見を求めた」ものとなっている．第1段階として，これら全用語を採用した．

最終的に設定したニュース分野は**表1**に示す 16 分野である．主要なニュース分野は1～9で，他の分野は用語の分類作業をやりやすくするために設定したものである．これら分野の設定に当っては，NHK 放送総局資料部が新聞の切抜資料の分類に活用している『報道資料分類表』[12] を基本文献とし，それに，文献 11) の編者である NHK 放送文化調査研究所の放送用語の専門家の意見を取り入れて改訂を施した．

用語分類の基本ルールは次の2点である．

ア）各用語について特によく使用される分野1～Fを3分野まで選んで指定する．

イ）4分野以上にわたって使用される用語は一般分野0に入れる．

実際の分類作業は，NHK でのニュース編集及び用語研究の双方に従事した経験を持つ元職員1名の手を煩わした．作業に要した時間は延べ3か月，その間，

表1 設定したニュース分野
Table 1 News categories to classify words.

ニュース分野	主な内容（一部用語例を「 」で囲んで示した）
0 一般	分野にかかわらず使われる用語 (例)「受付」「消す」「無い」…
1 政治	一般，国会，選挙，政党，行政，地方自治，司法・警察，財政，外交，軍事，皇室
2 経済	一般，金融・証券，対外経済・貿易・為替，産業，農林水産，鉱業，工業，運輸・通信，土木・建設・建築，商業
3 社会	一般，報道・新聞・通信，放送，出版，福祉厚生，医事衛生，公害，労働一般，労働組合・運動・争議，調停機関，雇用
4 市民生活	一般，生活，婦人・こども・老人，趣味・娯楽，観光・旅行，行事・記念日，世相・風俗・習慣，季節の移り変り，日常語・俗語
5 事件・事故	一般，政治的事件，経済的事件，詐欺，殺人・傷害，強窃盗，自殺・家出・誘拐，火事・爆発，事故，自然災害，裁判
6 文化	一般，教育，宗教，文学，美術・工芸，芸能，映画，演劇，音楽，歴史的なもの・しきたり
7 科学	一般，自然科学，天文・地学，宇宙開発，原子力・原子物理学，生物，人文科学，社会科学，先端技術
8 スポーツ	一般，総合競技，陸上，水上，冬季スポーツ，球技，野球，武道（相撲を含む），その他のスポーツ，射幸的競技
9 国際	国際一般，国際連合，各国，両極一般，国際機関・団体・組織・条約・協定，問題・紛争・戦争
A 天気予報	主として天気予報に使われる用語 (例)「空もよう」「梅雨明け」「等圧線」「晴れ」…
B 動植物名	動物または植物の名前 (例)「牛」「竹」…
C 感情・心の動き	感情・気持・心の動きを表すもの（一部，性格的なものも含む） (例)「愛情」「居たたまれない」「怒る」「快活」…
D 職業・役職・地位	(例)「米屋」「社長」「元締め」…
E 動植物関連	動植物の一部・雌雄・総称・特色のあるもの (例)「尾ひれ」「花粉」「雌牛」「常緑樹」「冷血」…
F 地理・地形・自然の風景	(例)「太平洋」「地峡」「白砂青松」「連峰」…

表2 ニュース用語の分野別分類結果
Table 2 Results of classifying news words.

ニュース分野	採用語数	無重複	重複1	重複2	重複率
0 一般	11,967	11,967	0	0	0.0
1 政治	1,531	598	612	321	60.9
2 経済	2,290	1,076	858	356	53.0
3 社会	2,567	1,222	971	374	52.4
4 市民生活	8,587	6,049	2,204	334	29.6
5 事件・事故	2,015	902	969	144	55.2
6 文化	2,561	1,281	1,081	199	50.0
7 科学	900	460	388	52	48.9
8 スポーツ	511	240	204	67	53.0
9 国際	218	36	127	55	83.5
A 天気予報	229	71	150	8	69.0
B 動植物名	476	383	85	8	19.5
C 感情・心の動き	1,030	957	65	8	7.1
D 職業・役職・地位	466	115	302	49	75.3
E 動植物関連	281	183	86	12	34.9
F 地理・地形・自然の風景	273	206	56	11	24.5

異なり語数 30,480
採用語数の合計 35,902
無重複……他分野と重複しないもの
重複1……他の1分野と重複するもの
重複2……他の2分野と重複するもの
重複率……(重複1＋重複2)/採用語数(%)

当初は分野0～Bのみで出発したものが，作業の進展とともに特殊分野C～Fを分離していき，最終的に表1に示す16分野に落ち着いた．

分類結果は**表2**に示すとおりである．今回の分類作業に関する一般的結果として次の点が注目される．誤入力訂正処理に関連する結果については3.2節で述べる．

1)「4 市民生活」の分野に予想外に多数の用語が分類された．ここは，用途によってはさらに細分の必要があろう．

2) 主要な分野 1～9 の他分野との重複率はおよそ50% である．さらに，ここには示さなかったが，分野1～6間の重複が比較的多く，逆に分野A～F間の重複はほとんど無いという結果も得られている．

3) 以下のようなやや特殊な用語が見られた．

- 一般用語ではあるが，特定の分野で専門語として使われるもの：解散，肩透かし等の 32 語
- 「4 市民生活」中の日常語・俗語：腹れっ面，めちゃくちゃ等の 274 語
- 「6 文化」中の歴史的な用語：印ろう，代官等の 117 語

3. 誤入力訂正への適用

本章では大小様々のニュース分野別辞書を，先に我我が提案した誤入力訂正処理[2]に適用して，実用規模の辞書に対しても我々の訂正手法が有効であることを示す．

ここで誤入力訂正処理とは，音声入力や OCR 入力の結果である入力誤りを含む文を受け取って，それに訂正処理を施し，最終的に正しい文を出力することを言う．我々の訂正法は，単語中の1字の置換誤りを対象としている．これは誤入力訂正の基礎になるだけでなく，それ自体，キーワード検索において少なくないキーワード入力誤り[13]の自動訂正等にも活用できる．

3.1 誤り訂正法と用語辞書の役割[2]

今回用いた訂正手法そのものは，先に我々が提案したものと同じである．詳細は文献2）に譲って，ここでは以下の説明に必要なことだけを簡単に述べる．

まず用語辞書の役割であるが，第1に，入力された単語あるいは訂正を施された単語が，正しい単語であるか否かのチェックに使われる．すなわち，入力単語が辞書に無ければ誤りがあると判定され，また，訂正単語が辞書にあれば訂正が成功したと判定される．

第2に，誤字検出等の基本データとなる，以下のような2種類の2字組の出現頻度を，辞書中の全用語について足し上げて求めるのに用いられる．

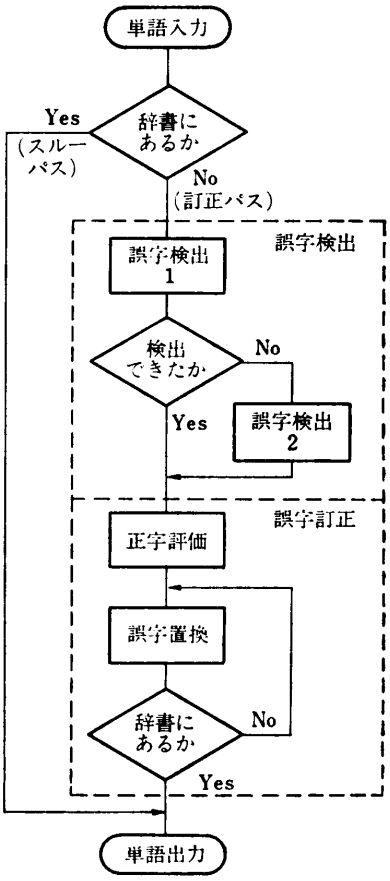

図1 誤り訂正の流れ
Fig. 1 Flow of the correction of a substitution error per word.

長さ n の単語 $a_1a_2\cdots a_n$ に対して，

1° 隣接2字組：a_ia_{i+1} $(i=1\sim n-1)$

2° 語尾語頭2字組：a_na_1

以下，1° に関する頻度表を T_1，2° に関するそれを T_2 と呼ぶ．T_1 と T_2 は，単語を構成するコード種別数を m とする時，m^2 個のエントリーを持つ．

訂正処理の大まかな流れは**図1**に示すとおりである．図1の「誤字検出1」では，T_1 と T_2 を用いて，誤入力単語中の頻度0の2字組を探す．そのような2字組が見つかれば，その両方を誤字候補とし，見つからなければ「誤字検出2」に入り，すべての文字を誤字候補として処理する．

T_1, T_2 は「正字評価」でも使われる．すなわち，入力単語中の文字 a_i が誤字候補と判定された時，文字 c が a_i の正字である可能性の評価値 $g(c, a_i)$ として

次の関数を用いる.

$$g(c, a_i)=K^2(c, a_i)\times f(a_{i-1}, c)\times f(c, a_{i+1}) \quad (1)$$

ここで $K(c, a_i)$ は c を a_i と誤入力する頻度で,'正字候補表' として前もって作成しておく. また, $f(x, y)$ は2字組 xy の頻度であり T_1 または T_2 より求める*. $g(c, a_i)$ をすべての正字候補 c について求め, その値の大きい順を正字候補の順位とする. この順に誤字を正字候補で置換し, それが辞書に登録されたものであればそれを訂正結果として出力する.

3.2 誤り訂正から見たニュース用語辞書の性質

本節では我々の誤り訂正手法の観点から, ニュース用語辞書の諸性質を調べることにする. そのために, それに先立って, 用語の表記系と辞書の選択について述べる.

入力機器の現状から考えて, 用語の表記系としては JIS カナコード, 68音節コード, あるいは102音節コード等が一般的であろう. 我々の訂正法は, 必要なテーブルさえ用意すれば, どのコード系にも適用できるが, コード種別が少なくなるにつれて上述の頻度表 T_1, T_2 中の頻度0の2字組の割合が減少する[2]. これが図1の「誤字検出1」を難しくし, 結果的に訂正率を低下させる. そこで以下の実験では, 一般的なコード系の中でコード種別最少の (したがって最も条件の厳しい) JIS カナコードを用いることにする. これは通常のカナ46字, 小さいカナ9字, それに長音符号, 濁点, 半濁点を加えた合計58種のコードから成る**.

実験には8種類の辞書を取り上げた. 予備検討の結果から辞書の収容語数が訂正率に大きく寄与することが分かっていたので (このことは図2にも明らかである), 表2の結果を参考にして, まず収容語数の多い

J0: 一般

J4: 市民生活

を取り上げた. 次に語数2,000前後の辞書の中から, 語数最小のもの及び最大のものとして,

J1: 政治

J3: 社会

を選んだ. このうち, 真の分野別辞書 J1, J3, J4 は, 実用上は一般分野別辞書 J0 と合併して用いられるべきなので, これら合併辞書を含むさらに4種類の辞書を加えた.

K0: ニュース用語辞書の全体

K1: J0+J1

K3: J0+J3

K4: J0+J4

誤入力訂正処理への実用という点からはK系の辞書のみが意味を持つが, 辞書の収容語数が訂正率等へ及ぼす影響を定量的に調べるために, あえてJ系の辞書も調査及び実験の対象に加えた.

基本分野別辞書 J0〜J4 の単語長の統計を**表3**に示す.

調査結果は**表4**にまとめた. 以下, 調査項目について説明する.

(1) ゼロ頻度率: t_1, t_2

これは辞書の見出しから作られる2文字連続頻度表 T_1, T_2 の全エントリー (総数は 58^2) のうち, 頻度値が0になっているものの割合(%)である. この値が大きいほど, 誤字検出がやりやすくなる. t_1 は隣接2字組の表 T_1 に対するもの, t_2 は語尾語頭2字組の表 T_2 に対するものである.

なお T_1 は長さ3以上の用語に対して, T_2 は長さ2以上の用語に対して作成した. 頻度値が0でないものも, 大部分は255以下である. 256以上の値を持つエントリーは, 辞書 J0 の T_1 で 13, 辞書 K0 の T_1 で 70 である. 最高値は K0 の T_1 中の2字組「ョウ」に対応する 2,373 であった*.

表3 用語長の統計
Table 3 Statistics of word length.

辞書		J0	J1	J3	J4
用語数		11,967	1,531	2,567	8,587
語長	1	424	1	2	4
	2	1,849	17	79	333
	3	2,253	213	349	1,282
	4	3,298	552	854	2,551
	5	2,564	463	721	2,467
	6	1,105	182	340	1,278
	7	318	54	136	440
	8	102	23	55	153
	9	39	12	21	52
	10	11	11	5	17
	11	2	2	3	4
	12	2	1	2	6
平均語長		3.9	4.6	4.7	4.6

* T_2 が用いられるのは誤字候補 a_i が語頭または語尾にある時である.

** JIS C6220 のカナコード系には, このほかに句読点, かっこ等も含まれているが, ここでは単語表記に用いられるものに限った.

* 誤り訂正処理の実行に際しては, 256以上の値は全て255で打ち切り, T_1 と T_2 の全エントリーを1バイトに圧縮して使用した. このような圧縮による訂正率の低下は K0 の場合で 0.3% 程度であった.

表 4 誤り訂正処理から見た辞書の性質
Table 4 Properties of lexicons for error correction.

辞書		J0	J1	J3	J4	K1	K3	K4	K0
用語数		11,967	1,531	2,567	8,587	13,498	14,534	20,554	30,480
ゼロ頻度率(%)	t_1	46	79	65	43	45	44	38	35
	t_2	63	88	78	58	61	59	52	49
エントロピー	h_1	5.2	4.8	5.0	5.2	5.1	5.1	5.2	5.1
	h_2	5.2	3.2	3.7	4.2	4.0	4.0	4.2	4.1
	h_3	2.7	2.1	2.1	2.8	2.8	2.8	3.0	3.1
一部異なり語数(平均値)	n_1	5.3	1.2	0.8	1.1	5.7	5.5	5.0	7.3
	n_2	48.0	9.4	8.2	16.7	53.9	53.7	59.2	87.8
	n_3	334.2	44.5	55.7	158.7	369.6	381.7	481.3	690.8

（2） エントロピー：h_1, h_2, h_3

辞書中の用語を構成している文字（コード）の出現頻度は大きく偏っており，それによって誤字訂正が可能になっているわけである．上記のゼロ頻度率も，2文字連続の頻度分布の偏りの一面を表現しているが，分布全体の偏り具合を定量的に表現するものにエントロピーがある．ここでは，1文字，2文字連続，3文字連続という3種の頻度分布のエントロピー h_1, h_2, h_3 を調べた：

$$h_1 = -\sum_i p(i) \log_2 p(i),$$

$$h_2 = -\sum_{i,j} p(i, j) \log_2 p(i, j),$$

$$h_3 = -\sum_{i,j,k} p(i, j, k) \log_2 p(i, j, k).$$

ここで例えば $p(i, j)$ は2文字連続 i, j の頻度の割合である．

h_i が最大となるのは一様分布の時で，参考までにその値を求めると次のようになる．

$$\max\ h_1 = \log 58 = 5.86$$

$$\max\ h_2 = \log 58^2 = 11.7$$

$$\max\ h_3 = \log 58^3 = 17.6$$

（3） 一部異なり語の平均個数：n_1, n_2, n_3

見出し表記上で似通った用語が辞書中にどの程度あるかを示す量である．同じ語長の用語を比較した時それらの間の異なり文字数 d を用語間距離と定義すると，語数 N の辞書において距離 d にある用語対の数 N_d の N に対する比 $n_d = N_d/N$ は，一つの用語に対しそれと距離 d にある用語の平均個数を表す[14]．この n_d の値が大きいことは誤り訂正において誤訂正が多くなりやすいことにつながり好ましくない．n_d は長さ3以上の用語に対して求めた．

表4から次のことが分かる．

1) 2文字連続の頻度分布は，h_2 と $\max h_2$ の値の比較から分かるように分類された辞書においては相当に偏っており，ゼロ頻度率 t_1, t_2 も大きいので，高い誤字検出力が期待できる．

2) 一方 n_1 の値から分かるように，特にK系の辞書に対して1文字だけしか異なっていない用語が結構あり，下手に誤字訂正を行うと別の用語に化けてしまう危険性も高い．しかし n_1 は，K0のものと他のものとの比較から分かるように，辞書を分類することによりある程度値を小さくできる．

3.3 誤入力訂正実験

我々の訂正処理の基本対象は単語中の1字の置換誤りである．それは音声入力やOCR入力の際に生じる誤りの型であるが，その訂正手法は様々の型の誤入力訂正の基礎になるとも考えられる．また，我々の訂正手法の能力をできるだけ厳しく評価するためには，前節でも述べたようになるべく種別の少ないコード系を使った方がよいが，手近なコード系の中ではJISカナコードが最少である．

このような点を考慮して，次のような訂正実験を行った．

ア）辞書からランダムに100語を選定する．

イ）選定した単語に対して，JISカナ鍵盤を想定して誤りデータを作成する．すなわち，各キーの周辺キーに均等に誤るとして* 可能な限り誤字と誤字位置を変え，ちょうど1文字の置換誤りを含む，合計約2,000の誤入力サンプル単語を作成する．

ウ）これに訂正処理を施して，もとの単語に復元で

* この場合，誤字は1字につき平均4.4個程度であった．周辺キーだけでなく鍵盤全体に誤るとした場合でも訂正率は大きく低下しないことが表6に示される．

表 5 誤入力訂正実験の結果
Table 5 Results of error correction experiments.

辞書	J 0	J 1	J 3	J 4	K 1	K 3	K 4	K 0
用語数	11,967	1,531	2,567	8,587	13,498	14,534	20,554	30,480
入力サンプル数	2,237	2,415	2,263	2,525	2,275	2,249	2,420	2,326
訂正率(%) c_1	90.7	96.8	97.2	95.3	92.0	89.1	90.7	84.5
(c_2)	(32.3)	(69.2)	(58.0)	(29.2)	(32.3)	(30.6)	(25.2)	(21.1)
誤訂正 (%)	5.9	2.2	2.2	3.5	5.2	7.4	5.6	9.4
無訂正 (%)	3.4	1.1	0.6	1.0	2.7	3.2	3.0	5.6
辞書探索回数	3.8	2.6	2.9	4.6	3.8	3.9	4.6	3.9
打ち切り回数	1	0	0	6	4	8	16	11

きるかどうかを調べる.

8種の辞書に対する訂正実験の結果は**表5**に示すとおりである. 以下, まず項目ごとに結果を見ていく.

(1) 訂正率: c_1, c_2

本来の訂正率は c_1 で, 辞書 K 0 に対する以外はすべて 90% 以上という良い結果が得られている. 参考データとして, 誤字検出の際に頻度0の2字組のみを手がかりとした場合 (図1中のパス「誤字検出2」を省略した場合) の訂正率を c_2 として示した. この場合は, 誤字検出ができないケースが増えるので, c_2 は大幅に低下している. 辞書を分類したことによる効果については後述する.

(2) 誤訂正, 無訂正

我々の場合, 誤字の2段階検出を行っている結果, 誤字検出不能という状況は起こらない. 訂正が失敗するのは専ら誤訂正と無訂正による. 誤訂正とは, 例えばもとの単語「カイトウ」に対する誤入力「カイテウ」を, 別の単語「カイテイ」に訂正してしまう場合である. 無訂正とは, 例えばもとの単語「テンラン」に対する誤入力「テンラク」を, そのまま正しい単語と見てしまう (「テンラク」は辞書にあるので図1のスルーパスを通ってしまう) 場合である. 単語の文字列の並びの統計的性質のみに基づく訂正法では, このような誤訂正や無訂正は避けられないが, 表5で見る限り, その値は予想外に小さいと言える.

(3) 辞書探索回数, 打ち切り回数

訂正が成功したか否かは, 訂正候補を辞書の見出しと比較することにより行われる. 訂正が成功するまでに平均して何回辞書を引かねばならないか, それを示したのが「辞書探索回数」である. なお, 表5に示したこの回数には, 入力された単語に誤りがあるか否かを発見するための最初の辞書探索1回分をも含めてある.

辞書探索は訂正処理時間の大きな部分を占めるので, この値は小さいほど良い. この値はまた, 訂正候補を作る際の正字評価関数(1)の良さをも示している. 表5から分かるように, 平均して3, 4回の辞書探索で訂正に成功しており, 高速な訂正処理が可能である.

辞書探索回数は平均値であるから, 実際には表5の値を大幅に上回る場合も起こる. 我々の実験では, 訂正候補を 20 回まで作り直しても訂正に成功しない場合, 訂正処理を打ち切ることにした. そのような打ち切りが何回起こったかを示したのが「打ち切り回数」である. 打ち切りによる訂正失敗は誤訂正でも無訂正でもないから, 表5で打ち切り回数の大きいところでは, 訂正率 c_1 と誤訂正と無訂正の値の和が 100% 以下になっている.

なお上述のとおり, 誤入力の発生はキー入力を想定して各キーの周辺に均等に誤るとしたが, その条件を変えた場合の結果を**表6**に示す. ここでは代表的に一

表 6 訂正処理に及ぼす誤り条件の影響
Table 6 Effects of error condition on the error correction processing.

誤り条件						
P_a	周辺に均等	0.57				0.48
P_b		0.33				0.31
P_c	0	0.11				0.21
訂正回数の上限	20	20	40	60	80	80
辞書探索回数	3.8	4.4	4.9	5.1	5.1	6.3
打ち切り回数	1	61	25	12	5	9
訂正率 (%)	90.7	85.9	87.6	88.3	88.7	86.5

P_a: 左右キーに誤る割合
P_b: 上下キー 〃
P_c: 周辺以外のキー 〃
使用辞書: J 0 入力サンプル数 1,768

般用語辞書 J0 に対する結果のみを示した．周辺以外のキーにも誤るとした場合，通常の訂正回数の上限値 20 では打ち切り回数が大きくなり，80 回程度に増やす必要がある．しかし，その場合でも辞書探索の平均回数は余り増えない．表6で見る限り，我々の訂正方式は誤り条件の変化に対して相当に強じんであると言えよう．

表5，表6などから次のような考察が可能である．

1) 全用語辞書 K0 を用いた訂正では 84.5% の訂正率しか得られていないが，分野別辞書 K1〜K4 を用いると 90% 程度の訂正率が得られている．用語分類の効果はあったと言えよう．

さらに表5の訂正率 c_1, c_2 と，表4のゼロ頻度率 t_1, t_2 を辞書の語数の対数でプロットすると，**図2**に示すように，いずれもほぼ直線上に乗る．このことから，我々の訂正手法は使用する辞書の語数の対数にほぼ比例することが分かり，表5のK系の辞書に対する結果を一般化して「辞書の語数が2万以下ならば 90% 以上の訂正率を実現できる」と言えるように思われる．放送ニュース用語の場合には，表2から明らかなように，ニュース分野別分類によって語数を2万程度にすることは可能であり，したがって我々の訂正手法は実用に耐え得ると言えよう．

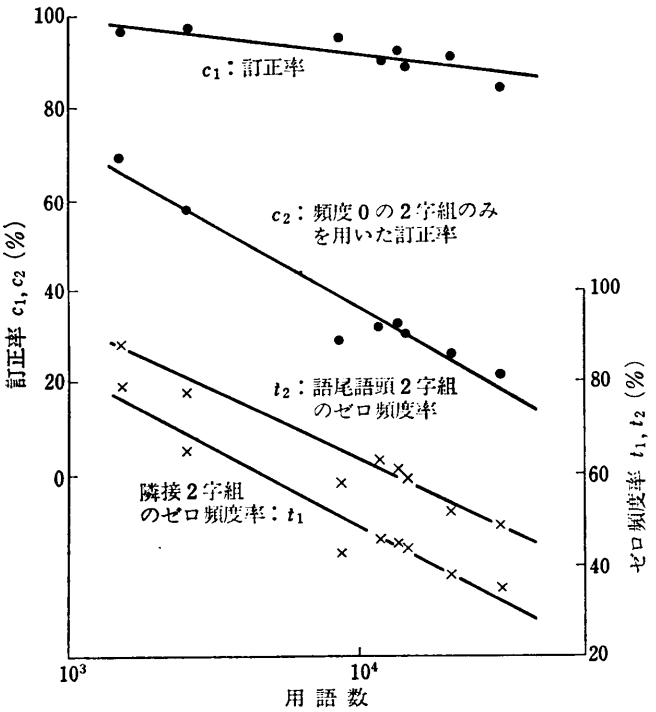

図 2 用語数による訂正率とゼロ頻度率の変化
Fig. 2 Dependences of correction rates and zero rates of digram frequency tables on the number of words.

2) この訂正実験結果を単音節音声認識の現状技術と結びつけて，どの程度の認識率の改善が見込めるか試算してみる．市販の単音節音声認識装置を6人の話者に使ってもらい，認識結果の良い順に2人ずつまとめることにより，次の3種のデータ I〜III を得た*．

	I	II	III
単音節認識率	80	73	63
単　語　〃	68	53	34
1音誤る単語の率	28	38	45
2音　〃　〃	4	8	19

（単位：%）

我々の訂正処理は，このうちの「1音誤る単語の率」を‘訂正率’の分だけ訂正し，結果的に「単語認識率」を改善するように働く．表5の訂正率を用いてその値を計算すると次の結果が得られる．

	I	II	III
K1で訂正した時の単語認識率	94	88	75
K4　〃　〃	93	87	75

極めて単純化した試算結果であるが，これで見る限り相当の訂正効果が期待できる．

3) 1回の訂正に要する辞書探索回数が辞書の語数にかかわらず平均的に5回以下であることも注目すべきことと思われる（表5参照）．これは正字候補評価関数(1)が適切で，むだな訂正候補を余り生成しないことを意味している．辞書探索回数が少ないことは，辞書を外部記憶装置に置く場合に特に有利になる．

4. むすび

先に我々が提案した誤入力訂正手法[2]においては用語辞書が重要な役割を果たしていたが，用語辞書の諸性質（特に収容語数）と訂正能力との関係は必ずしも明確でなかった．殊に先の報告では6,000語程度の辞書のみを用いていたので，それを実用規模の辞書に拡張した時の訂正能力の低下の程度を見極めておく必要があった．

そこで本論文では，放送ニュース用語約3万を代表的なニュース分野に分類し，それによっ

* その詳細については別途まとめて報告する予定である．

て得られた 1,500～3万語の大小8種の辞書による訂正処理実験を行い，辞書の諸性質と訂正能力との関係を定量的に調べた．その結果，特に，訂正能力は辞書の語数の対数にほぼ比例して低下するが，語数が2万以下ならば 90% 以上の訂正率が確保できることが分かった．しかしこの結果は，訂正手法の点から見て最も厳しい，58 種コードの表記系を用いた場合のものであり，それ以上のコード種を持つ表記系の場合にはさらに良い結果が得られる可能性がある．

放送ニュース用語の場合には，政治，経済，社会等の代表的な9分野への分類によって，分野別辞書の語数を 14,000～2 万に抑えることが可能であり，したがって，これら分野別辞書を切り換えて使うことにより，放送用語全体にわたって高い訂正能力を達成できる見通しが得られたことになる．

我々がこれまでに検討してきた誤入力訂正処理は，単語中の1字の置換誤りの訂正であったが，'音声入力ワードプロセッサ' の実現等を考えると，次の段階として，文節単位の訂正能力が不可欠である．これについても引き続き検討中であり，機会を改めて報告したい．

謝辞 ニュース用語の分類に関して種々の有益なご意見を頂いた NHK 放送文化調査研究所の菅野，井上の両主任研究員（当時）に深謝する．

参考文献

1) 川合：英文綴り検査法，情報処理，Vol. 24, No. 4, pp. 507-513 (1983).
2) 栗田，相沢：日本語に適した単語の誤入力訂正法とその大語い単語音声認識への応用，情報処理学会論文誌，Vol. 25, No. 5, pp. 831-841(1984).
3) 池原，白井：単語解析プログラムによる日本文誤字の自動検出と二次マルコフモデルによる訂正候補の抽出，情報処理学会論文誌，Vol. 25, No. 2, pp. 298-305 (1984).
4) 並木，浜田，中津：音声認識を用いた日本語入力方式，信学論，Vol. J 67-D, No. 4, pp. 405-412 (1984).
5) 長田，牧野，日高：日本語の文脈情報を用いた文字認識，信学論，Vol. J 67-D, No. 4, pp. 520-527 (1984).
6) 新谷，目黒，梅田：認識情報及び単語・文節情報を利用した文字認識後処理，信学論，Vol. J 67-D, No. 11, pp. 1348-1355 (1984).
7) 三橋，八田，平塚：音声入力における誤認識訂正処理，日本文入力方式研究会資料，19-1 (1984).
8) 杉村，斎藤：文字連接情報を用いた読取り不能文字の判定処理，信学論，Vol. J 68-D, No. 1, pp. 64-71 (1985).
9) 小林，小森，白井：大語彙を対象とした文節音声の認識，信学論，Vol. J 68-D, No. 6, pp. 1304-1311 (1985).
10) 中川，義永：誤りを含んだ音素系列からの候補単語の検索，計量国語学，Vol. 14, No. 8, pp. 327-334 (1985).
11) NHK 編：新用字用語辞典（第1版第4刷），559 p.，日本放送出版協会，東京 (1983).
12) NHK 編：報道資料分類表，(部内資料)，318 p. (1970).
13) Blair, D. C. and Maron, M. E.: An Evaluation of Retrieval Effectiveness for a Full-text Document-retrieval System, *Comm. ACM*, Vol. 28, No. 3, pp. 289-299 (1985).
14) 阿部ほか：辞書を利用する文字認識系の能力の評価，信学論，Vol. 52-C, No. 6, pp. 305-312 (1969).

(昭和 60 年 10 月 1 日受付)
(昭和 61 年 5 月 15 日採録)

相沢 輝昭（正会員）

昭和 15 年生．昭和 38 年京都大学工学部電気工学科卒業．同年 NHK に入局．以来，放送技術研究所において日本語処理及び情報検索の研究に従事．昭和 61 年(株) ATR 自動翻訳電話研究所に出向．現在，同所言語処理研究室長．電子通信学会，AVIRG 各会員．

栗田泰市郎

昭和 30 年生．昭和 55 年慶応義塾大学大学院修士課程修了．同年 NHK に入局，長野放送局．放送技術研究所情報処理研究部を経て，現在同研究所テレビ方式研究部に勤務．高品位テレビジョンの研究に従事．電子通信学会，テレビジョン学会各会員．