統計数理 (1992)
第40巻 第2号 131-151

中国語高頻度単語の拼音表記の統計的特性

総合研究大学院大学　金　明哲*
統計数理研究所・総合研究大学院大学　村上　征勝

(1992年12月　受付)

ま え が き

中国語（普通語）の発音を，ローマ字で表記したものを拼音（Pin yin, ピンイン）という．中国では拼音は1958年に正式に導入されたが，今日に至るまで，漢字学習のための漢字の音の表記としての使用に留まり，拼音に対する研究はほとんど行なわれていない．しかし，中国における対外開放政策と情報処理機器の普及により，拼音の研究は次第に注目されてきている．現在中国では，中国語をキーボードから計算機に入力するのに10種類以上の方法が用いられている．その中で最も広範に用いられている方法が，ローマ字で入力し，それを漢字に変換する拼音入力法である．小学校では拼音による漢字教育が導入されており，学校教育を受けた人であれば，誰でも特別な訓練を受けずに計算機に拼音で漢字入力ができるという状況である．したがって，拼音入力法は，多くの専門家から中国では一番将来性のあるキータッチ入力法と目されているといってよい（猛 (1990)).

本格的な自然言語処理システムの開発のためには，あらかじめ，自然言語の本質的な特徴である表現形式や意味の多様性に十分配慮する必要がある．そのためには，言語学，心理学，数理統計学，工学などさまざまな立場から言語を分析し，有効な情報を整理しておくことが必要であるが，残念ながら中国語の場合は，機械処理を目的とした言語の計量的な研究は十分には行なわれていない．特に，中国語の拼音表記に関する計量分析の研究論文はこれまでにない．中国語の拼音表記の研究と，その機械処理の適応性の検証及びそれに役立つ基礎的な情報を提供するのが本論文の目的である．

自然言語の記号列（文字，音素，音節，単語，品詞）は，思考を主とした意味論の観点や，統計的な観点，あるいは構文的な観点などの種々の観点から解析されてきた．このうち統計的な観点からの解析は，計算機の発展に従い急速に進歩し，綴誤りの検出・訂正や自然言語の理解と処理にも応用されるようになった．

自然言語の文字列の誤りについては，その80%以上は，(1) 1つの記号の置換，(2) 1つの記号の挿入，(3) 1つの記号の脱落，(4) 隣接記号の互換によって生じ，これらの置換，挿入，脱落，互換に対して，(1), (2), (3), (4) の逆変換と辞書を併用したアルゴリズムによって，約95%の割合でもとの正しい綴に復元できるとDamerau (1964) は報告している．現在，知られている誤り検出と訂正の方法は，(i) 辞書に登録されている全単語との参照を基本とする方法と，(ii) n 個の文字の組である多字組（n-gram）の出現頻度等を利用する方法に分けられる．本論文では前者を辞書法，後者を統計法と呼ぶことにする．辞書法では，任意の文字綴同士の類似度を数量的にとらえる方法が中心になっている．文字綴同士の類似度の数量的なとらえかたに

* 数物科学研究科　統計科学専攻：〒106　東京都港区南麻布4-6-7.

関して，Levenshtein (1965) は文字綴間の距離を測り，最も小さい距離値を持つ文字綴を訂正の第一候補にする方法を提案した．その後，Okuta et al. (1976) は脱落，挿入，置換に重みを付けて距離を測る方法に発展させた．統計法では，多字組による誤りの検出と訂正が中心になっており，2字組 (digram)，3字組 (trigram) の頻度を利用した誤りの検出と訂正法がよく用いられている．多字組の頻度を利用して英単語の誤り訂正を試みた例としては，Cornew (1968) をはじめ，Harmon (1972), Riseman and Hanson (1974), Hanson et al. (1976), Hull and Srihari (1982), Shinghal (1983) などがある．また，機械処理に役立つ情報を提供することを主な目的とした研究としては，Suen (1979) が代表的であり，この中で彼は1959年から1978年までの英語の文字列の多字組に関する計量分析の論文10編の統計データをまとめている．日本語の統計法による誤り検出と誤り訂正に関しては，池原・白井 (1984)，栗田・相沢 (1984) などがあり，2字組に関しては，今栄 (1960) が心理学の立場から分析している．多字組 (*n*-gram) の統計データを利用して単語の音声認識における誤り訂正を試みた研究としては，Jelinek (1976)，鹿野 (1987)，丸山 (1989) がある．また品詞を単位とした記号列の統計データを学習し，ニューラルネットワークによる単語品詞列を予測するという研究も報告されている (中村・鹿野 (1991))．このように，言語の統計的性質が自然言語の理解と処理のために注目されてきている．

本論文では，拼音表記の機械処理に関連する高頻度単語の単語長，近距離単語数，声調や品詞に関する情報の有無と近距離単語数との関係，置換対，エントロピー，多字組の頻度などについて計量分析を試みた．分析に用いたのはSuen (1986) の論文に示された中国語高頻度単語6,321語であり，品詞分類は香坂 (1989) を参照して，12品詞に分類して行なった．この高頻度単語6,321語は，Suenが調べた新聞，雑誌，学校のテキスト，校外読み物など879,300語の90%をカバーしており，中国語語彙の特徴を十分反映しているものと考えられる．

1. 拼音表記について

拼音は大きく声母と韻母に分けられる．声母は音節の頭の子音で，韻母は音節の声母を除く要素 (すなわち母音および音節末の子音) である．例えば，/ma/のmは声母で，aは韻母である．日本語と同様，中国語にも/ai/, /an/のような音節の頭の子音がない音節があるが，音韻論ではこの場合「ゼロ声母」が存在するとみなす．ゼロ声母を含めて，声母は22種類ある．しかし，本論文はローマ字表記の計量分析が中心であるので，声母と言った場合は，特別の説明がない限りゼロ声母を除いた21種類を指すことにする．21種類の声母の中で，一個のローマ字で表記されるものが18種類，二個のローマ字で表記されるものが3種類ある．韻母はローマ字の形で分類すると36種類で，音声で分類すると38種類ある (金 他 (1992))．これは韻母/i/が三種類の発音を用いるためである．ローマ字の形で分類した36種類の韻母を記号列の長さから見ると，最も短いものは一個のローマ字により表記され，最も長いものは四個のローマ字により表記される．韻母を表記するローマ字は，声母を表記するローマ字n, g, rと，韻母だけを表記するa, o, e, i, u, ü, y, wに分けられる．中国語は，この声母 (ゼロ声母も考慮する) と韻母を組み合わせた，つまり「声母＋韻母」の型の音節が410個ある (特別な音，例えば/hm/などは除く (北京語言学院 編 (1986)))．中国語の漢字の発音は，音節だけではなく，音節と音節の高低変化を表す「声調」を加えて表される．声調には4つの基本声調と軽声の5種類があり，同じ音節でも，漢字が異なると声調もまた異なる (北京語言学院 編 (1986)，松本 (1986))．表1(a) に拼音をローマ字の字数で分類した結果を，表1(b) に漢字と拼音，声調のいくつかを例示した．

表1(a). 中国語拼音体系とローマ字表記.

拼音					
声母		韻母			
1文字	2文字	1文字	2文字	3文字	4文字
b, p, m, f, d,	zh, ch, sh	a, o, e,	ai, ao, an, en,	ang, eng,	iang,
t, n, l, g, k,		i, u, ü	ei, ia, ie, in,	ong, ing,	uang,
h, j, q, x, z,			iu, ou. ua. uo,	ian, iao,	ueng,
c, s, r			un, ui, üe, ün,	uai, uan,	iong
			er	üan	

表1(b). 中国語漢字と拼音，声調の例.

中国語	声母	韻母	声調記号と名	意味
duō 多	d	uo	— 一声	多い
ào 奥	ゼロ	ao	＼ 四声	奥深い
mā 妈	m	a	— 一声	お母さん
má 麻	m	a	／ 二声	麻
mǎ 马	m	a	∨ 三声	馬
mà 骂	m	a	＼ 四声	ののしる
ma 吗	m	a	なし 軽声	～か？

2. 単語の長さの分布

Suen (1986) の 6,321 語に関して，漢字，ローマ字を単位として測った単語の長さの分布を図 1(a) に，その累積度数分布を図 1(b) に示した．ローマ字を単位とした場合，6 文字の単語が一番多く全体の約 19% を占め，3～8 文字の単語は全体の約 85%，2～9 文字の単語は全体の約 97% を占める．品詞ごとに調べても，名詞と動詞はほぼ同じ傾向を示す．

漢字を単位として測った場合，一漢字単語が 1,762 語で全体の約 28% を占め，二漢字単語が 4,238 語で全体の約 67%，三漢字単語が 283 語で全体の約 4.4%，四漢字単語が 38 語で全体の約 0.6% を占める．漢字を単位とした場合の平均単語長は 1.78 で，ローマ字を単位とした場合の平均単語長は 5.67 である．名詞と動詞では，漢字を単位とした場合の平均単語長はそれぞれ 1.85 と 1.66 で，ローマ字を単位とした場合の平均単語長はそれぞれ 5.78 と 5.40 である．

表 2 は，中国語，英語 (Shinghal and Toussaint (1979), Hull and Srihari (1982))，日本語 (栗田・相沢 (1984)) の単語の長さの統計である．日本語はかな文字を単位とした統計であるので，仮に一つのかな文字をローマ字で 2 文字として計算すると，中国語は英語，日本語より平均単語長が短いということになる．単語長が短いと，単語に含まれる情報が相対的に少ない

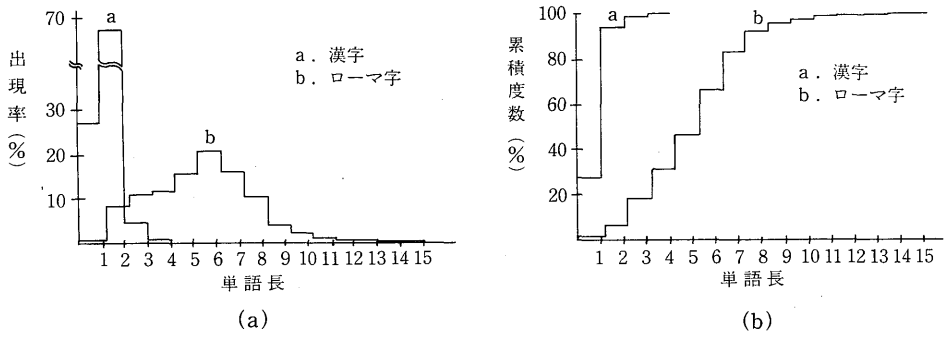

図1. (a) 漢字を測定単位とした場合とローマ字を測定単位とした場合の高頻度単語6,321語の単語長の分布. (b) 漢字を測定単位とした場合とローマ字を測定単位とした場合の高頻度単語6,321語の単語長の累積度数.

表2. 単語長の分布. A: Shinghal, R. and Toussaint, G.T. B: Hull, J.J. and Srihari, S.N.

言語	中国語	英語	
調査者	筆者	A	B
単語長(ローマ字)	単語数		
1	7	0	22
2	467	108	89
3	652	247	397
4	795	467	1,007
5	1,043	571	1,384
6	1,216	667	1,773
7	1,041	756	1,886
8	640	624	1,666
9	274	516	1,364
10以上	186	897	2,061
合計	6,321	4,835	11,603
平均	5.6	7.0	7.3

言語	日本語
調査者	栗田
単語長(仮名)	単語数
1	139
2	617
3	1,288
4	2,209
5	1,531
6	555
7	96
8	22
9	3
10以上	0
合計	6,460
平均	4.0

ため，誤り訂正には不利である．したがって，品詞情報，多字組の特徴等の情報を有効に利用することが英語や日本語より重要となる．

3. 近距離単語

類似記号列の検索や記号列の誤り訂正のために広範に使われている辞書法は，入力記号列と辞書内の記号列との間の距離を測る方法である．この方法で類似記号列を検索する場合には，ある距離範囲内の単語を検索対象として出力する．誤りを訂正する場合には，最も距離の近い記号列の単語を訂正の第一候補とする．

この場合に用いられる距離としては，ハミング距離（Hamming distance）とレーベンシュタイン距離（Levenshtein distance）がよく知られている（Levenshtein (1965), Okuta et al. (1976), 田中・菊地 (1980)）．ハミング距離は比較する記号列の長さが等しい場合に用いられ，レーベンシュタイン距離は比較する記号列の長さが必ずしも等しいとは言えない場合に用いられる．例えば，次の二つの記号列，もとの記号列 $X=a_1a_2\cdots a_s$ と入力された記号列 $Y=b_1b_2\cdots b_t$ を考えてみる．記号列 X から Y を得るのに，k 個の記号の置換，m 個の記号の挿入，n 個の記号の脱落があったとする．このような組 (k, m, n) は無限に多く存在するため，次の量

$$D(X, Y)=\min(w*k+q*m+r*n)$$

を記号列 X と記号列 Y の間の重み付きレーベンシュタイン距離と呼ぶことにする．ここで w, q, r は正の数で置換，挿入，脱落に付ける重みである．ただし，本論文では $w=q=r=1$ とし，それをレーベンシュタイン距離と呼び，記号列の長さが $s=t$ のときハミング距離と呼ぶ．

ここで置換，挿入，脱落の誤りの具体例を示す．例えば，単語の記号列 abcd を R と呼ぶ．この R について

置換による誤りの例として，T_1=abed, T_2=aefd,
挿入による誤りの例として，T_3=abced,
脱落による誤りの例として，T_4=ab,

をあげることができる．上の例で単語間の距離は $D(R, T_1)=1$, $D(R, T_2)=2$, $D(R, T_3)=1$, $D(R, T_4)=2$ である．ところで，T_3 は R の c と d の間に e が挿入されたものとみなすこともできるが，R の d が e に置換され，d が挿入されたものとみなすこともできる．したがって，混乱をさけるため前述のように記号列 X と Y の最小の距離を X と Y 間の距離と定義している．ところで，ハミング距離はレーベンシュタイン距離の部分集合と考えてよい．つまり，ある記号列 X からハミング距離が計算できる記号列というのは，かならず，X の記号列のいずれかの記号が置換されたものであるので，X からレーベンシュタイン距離 l の記号列の中に占めるハミング距離 l の記号列の割合を求めると，全体の置換，挿入，脱落の誤りの中で置換の誤りが占める割合が求められ，置換による誤りの訂正がどの程度重要かを知ることができる．

3.1 品詞の分布

中国語において，品詞の情報を用いた場合に，誤りの訂正がどの程度容易になるか，つまり近距離単語数をどの程度減少させることができるかを考察するため，まず本節では品詞の分布などについて調べてみる．中国語の単語にも，一つの単語が一つの品詞の機能のみを果たす単品詞単語の他に，一つの単語が幾つかの品詞の機能を果たす多品詞単語が存在する．例えば，「操作」という単語の場合，「操作机械」（機械を操作する）の時の操作は動詞であり，「操作方便」(操作は便利です）の時の操作は名詞である．

Suen の 6,321 語を，香坂（1989）の基準に従って，主な 12 品詞と多品詞，その他に分類した結果を表 3 に示した．単品詞単語と多品詞単語で全体の約 94% を占め，残りの約 6% がその他，すなわち，感嘆詞，接頭語，接尾語などである．単品詞の単語の割合をみると，名詞が全体の約 43% を占めて最も多く，次に動詞が約 20%，形容詞が 5.5%，副詞が約 2% で，これらの 4 つの品詞で全体の約 70% を占める．これ以外では最も多い代詞でも約 1% に過ぎない．

多品詞単語は 1,320 語で全体の約 21% を占める．この中で一漢字単語が約 39%，二漢字単語が 61%，三漢字単語は 1 語のみで，四漢字以上からなる単語はない．いくつかの品詞の機能を果たす多品詞単語について調べてみると，その種類は 78 ほどある．出現頻度の多い多品詞上位

表3. 6,321語の品詞の分布（単位は％）.

品詞	一漢字	二漢字	三漢字	四漢字	合計
名詞	7.40	31.40	3.48	0.43	42.71
動詞	6.61	13.06	0.19	0.02	19.88
助動詞	0.03	0.05			0.08
形容詞	0.79	4.65	0.08		5.52
数詞	0.17	0.20	0.02		0.39
助数詞	0.46	0.13	0.03		0.62
数量詞	0.03	0.06			0.09
代詞	0.20	0.62	0.03		0.85
副詞	0.27	1.55	0.06		1.88
介詞	0.02	0.14			0.16
接詞	0.03	0.47			0.55
助詞	0.06	0.05			0.11
多品詞	8.11	2.75	0.02		20.88
その他	3.67	1.93	0.52	0.16	6.28
合計	27.85	67.06	4.48	0.61	100.

表4. 多品詞のうち上位9位までの出現率（単位は％）.

品詞	一漢字	二漢字	三漢字	合計
名，動	10.68	43.49	0.77	54.24
名，形	2.80	5.91		8.71
動，形	2.35	5.00		7.35
名，助*	4.24	0.22		4.46
名，動，形	1.97	1.44		3.41
名，動，助	2.35	0.07		2.42
名，副	0.98	1.06		2.04
動，助	1.67	0.07		1.74
動，副	1.14	0.07		1.21
その他	10.69	3.73		14.42
合計	38.87	61.06	0.77	100

助*は助詞，助は助数詞を示す.

9種類で多品詞単語全体の約86%を占める．それらを表4に示した．名詞，動詞，形容詞，副詞の機能を果たす多品詞の単語は全多品詞単語の約77%を占め，その中の約96%が二品詞の機能を果たす単語で，最も多いのは名詞と動詞の機能を果たす多品詞単語で，全多品詞単語の約54%を占める．

多品詞単語を含めて機能別の分布をみると，全単語の約89%が名詞，動詞，形容詞，副詞である．表5に多品詞を含めた場合の中国語，日本語（板橋 他（1971），牧野・城戸（1979））お

表5.　中国語と日本語及び英語の品詞の割合（単位は%）.

調査者	中国語	日本語		英語
品詞	筆者	板橋秀一	牧野正三	牧野正三
名詞	50.78	73.9	70.66	51
動詞	30.59	15.0	16.38	20
形容詞	8.83	2.2		14
副詞	3.31	3.7		5
その他	6.48	5.2	12.96	10

表6.　拼音表記した場合の L_R 距離0~3の単語の分布（単位は語）.

	単語長（漢字）	声調 × 品詞 ×	声調 ○ 品詞 ×	声調 × 品詞 ○	声調 ○ 品詞 ○
距	1	7.78	2.34	3.34	1.05
	2	0.17	0.05	0.11	0.02
離	3	0.00	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.00
0	平均	2.28*	0.69	1.01	0.31
距	1	79.00	21.23	31.61	8.66
	2	2.83	0.61	1.62	0.35
離	3	0.05	0.01	0.00	0.00
	4	0.00	0.00	0.00	0.00
1	平均	23.92	6.33	9.90	2.65
距	1	379.50	100.00	149.20	39.98
	2	23.25	2.59	12.55	1.54
離	3	0.14	0.04	0.00	0.00
	4	0.05	0.05	0.00	0.00
2	平均	121.39	29.61	50.01	12.18
距	1	702.80	183.10	291.80	77.27
	2	121.70	11.59	63.66	6.39
離	3	0.40	0.06	0.00	0.00
	4	0.00	0.00	0.00	0.00
3	平均	277.50	58.81	124.02	25.82

*[7.78×1762+0.17×4238+0.00×283+0.00×38]／6321＝2.28.
1762は一漢字単語数，4238は二漢字単語数，283は三漢字単語数，38は四漢字単語数，6321は総単語数である．

表7. 拼音表記した場合の H_R 距離1～3の単語の分布（単位は語）.

	単語長（漢字）	声調 × 品詞 ×	声調 ○ 品詞 ×	声調 × 品詞 ○	声調 ○ 品詞 ○
距離1	1	64.08	16.99	25.26	6.80
	2	2.41	0.34	1.34	0.21
	3	0.03	0.01	0.01	0.01
	4	0.00	0.00	0.00	0.00
	平均	19.47	4.96	7.94	2.03
距離2	1	185.80	48.49	71.69	19.09
	2	13.92	1.49	7.44	0.90
	3	0.08	0.02	0.04	0.02
	4	0.00	0.00	0.00	0.00
	平均	61.13	14.51	24.97	5.92
距離3	1	193.80	50.58	83.34	22.01
	2	53.13	4.85	27.37	2.65
	3	0.15	0.01	0.04	0.00
	4	0.00	0.00	0.00	0.00
	平均	89.65	17.35	41.58	7.91

よび英語（牧野・脇田（1986））の主な品詞割合を示した．中国語の名詞の割合は，日本語の名詞の割合より約20ポイント少なく，動詞の割合は日本語の場合より約14ポイント多い．

3.2 近距離単語数

中国語を拼音表記した場合に辞書法による誤り訂正が有効か否かを検証するため，同じ漢字数から成る単語群に関して，単語のローマ字記号列の間の距離（レーベンシュタイン距離（L_R 距離）とハミング距離（H_R 距離））を求めてみた．

声調，品詞の情報を用いた場合と用いない場合それぞれについて，1単語あたりの L_R 距離0～3の平均単語数を表6に，H_R 距離1～3の平均単語数を表7に示した．表の中の記号○は声調，品詞の情報を用いた場合，記号×は用いない場合を示す．一漢字単語では，声調と品詞の情報を用いない場合には，1単語あたり L_R 距離が0～2の単語はそれぞれ平均7.78語，79語，379.5語あるが，声調と品詞の情報を用いた場合には1.05語，8.66語，39.98語となり，近距離単語数は減少する．

一漢字単語の中では，声調と品詞情報を用いてもレーベンシュタイン距離1～2の単語がそれぞれ約9語，40語あり，誤り訂正は極めて難しいといえる．しかし二漢字単語の場合には声調，品詞情報を用いるとレーベンシュタイン距離1～2の単語がそれぞれ約0.35語，1.54語であるため，声調，品詞の情報を有効に利用するとSuenの6,321語の中の二漢字単語に関しては二つまでのローマ字の誤りの訂正は可能であることがわかる．

声調の情報を用いた場合，L_R 距離，H_R 距離で測った近距離単語数の減少状況はほぼ同じである．一漢字単語の場合についてみると，1単語あたり近距離単語の減少は，声調の情報を用いた場合は用いない場合の約26%～27%で，二漢字単語の場合には，声調の情報を用いた場合は用いない場合の9%～22%である（距離0を除く）．

品詞の情報を用いた場合の，L_R 距離，H_R 距離で測った近距離単語数の減少状況もほぼ同じで，一漢字単語の場合，品詞の情報を用いた場合は品詞の情報を用いない場合の38%～44%で，二漢字単語の場合は52%～62%である．品詞の情報を利用しても，L_R 距離1～3の単語数が約

表 8. 拼音表記した場合の品詞別の L_R 距離 1～2 の単語の分布（単位は語）.

品詞	単語長（漢字）	L_R 距離 1		L_R 距離 2	
		声調 ×	声調 ○	声調 ×	声調 ○
名詞	1	38.11	10.43	176.40	46.65
	2	1.97	0.41	15.47	1.87
	3	0.05	0.01	0.14	0.02
動詞	1	30.85	8.54	148.30	40.46
	2	1.02	0.24	8.02	1.06
	3				
形容詞	1	7.80	1.99	34.18	9.03
	2	0.37	0.11	2.74	0.33
	3	0.40			
副詞	1	4.06	1.17	19.71	5.27
	2	0.23	0.04	1.73	0.30
	3				
その他	1	13.45	2.35	67.99	10.16
	2	0.37	0.03	2.33	0.23
	3				

表 9. 拼音表記した場合の品詞別の H_R 距離 1～2 の単語の分布（単位は語）.

品詞	単語長（漢字）	H_R 距離 1		H_R 距離 2	
		声調 ×	声調 ○	声調 ×	声調 ○
名詞	1	30.31	8.07	86.10	22.75
	2	1.70	0.25	9.27	1.11
	3	0.02	0.01	0.08	0.02
動詞	1	24.84	6.82	69.44	18.73
	2	0.83	0.14	4.54	0.58
	3				
形容詞	1	6.23	1.57	15.37	3.98
	2	0.31	0.09	1.52	0.17
	3				
副詞	1	3.08	0.81	9.40	2.44
	2	0.21	0.04	0.91	0.13
	3				
その他	1	11.20	1.90	33.99	4.71
	2	0.35	0.03	1.46	0.18
	3				

38%～62% しか減少しないのは，名詞と動詞が全体の約 81% を占め，近距離単語が同じ品詞内に多く存在するからである．

名詞，動詞，形容詞，副詞の品詞別の L_R 距離 1～2, H_R 距離 1～2 の分布をそれぞれ表 8，表 9 に示した．動詞の近距離単語の分布は，Suen の 6,321 語において品詞を用いた場合の近距離単語の分布とほぼ同じである．したがって，ローマ字を単位とした近距離単語の頻度分布では，動詞が全体の傾向を代表していると考えることができる．

単語の品詞の情報を利用して誤り訂正の候補単語数を更に減らすためには，品詞の細分類，特に名詞，動詞の細分類が必要である．置換，脱落，挿入が合計 1～2 ローマ字あっても，比較対象内の単語間に混同が生じないようにするための細分類の品詞数は，表 5～8 のデータから容易に求められる．品詞の種類を無限に多くするのは不可能であるが，名詞と動詞は何れも細分類することは可能である（Xu (1989)，坂本 他 (1988)，橋本 訳 (1987)）．Xu (1989) は中国語の動詞を 9 種類に細分類して処理を試みているし，また OKI 電気の英日機械翻訳システムでは，名詞を 10 種類に細分類している．仮に細分類した各品詞の単語数が等しいとすると，動詞を 9 種類，名詞を 11 種類に細分類し（表 8 を参照），声調と品詞の情報を用いると，Suen の 6,321 語の中の名詞と動詞に関しては 1 単語内に 1 ローマ字が置換，脱落，挿入されても単語間に混同は生じない．

3.3 置 換

3.3.1 置換率と置換位置の関係

置換による誤りの訂正は挿入，脱落による誤りの訂正より容易である．そこで，置換，脱落，挿入によって生じる近距離単語の中で，置換のみによって生じる単語の割合を求めてみる．この割合は H_R 距離と L_R 距離の比で求められる．例えば，表 6 と表 7 の距離 1 の単語の場合には，声調と品詞の情報を用いない場合，声調情報のみを用いた場合，品詞情報のみを用いた場合，声調と品詞の両方の情報を用いた場合の H_R 距離 1 と L_R 距離 1 の比はそれぞれ 81.40%, 78.35%, 80.20%, 76.60% で，その平均は約 78% である．L_R 距離 2 の場合では同様な平均は約 50% で，L_R 距離 3 の場合では同様な平均は約 32% である．

図 2(a) と図 3(a) は，それぞれ一漢字単語と二漢字単語を拼音表記した場合に H_R 距離が 1 の単語間で，どの位置でどの程度のローマ字の置換が生じるかを示したものである．また，位置別置換の累積度数と声母を表すローマ字による置換の累積度数をそれぞれ図 2(b) と図 3(b) に示した．一漢字単語では，単語の頭から一番目の文字の置換による距離 1 の単語が H_R 距離 1 の単語の約 80% を占め，その 73% が声母を表すローマ字同士による置換である．また，H_R 距離 1 の単語の約 60% が声母を表すローマ字同士による置換である．二漢字単語でも，単語の最初の文字の置換による距離 1 の単語が最も多く，二漢字単語の H_R 距離 1 の単語の約 38% を占める．次に多いのは，頭から四番目の文字の置換による距離 1 の単語で，二漢字単語の H_R 距離 1 の単語の約 19% を占める．二漢字単語の場合，H_R 距離 1 の単語の約 69% が，声母を表すローマ字同士による置換である．これらの分析結果から，単語内に置換，挿入，脱落によって 1～2 ローマ字の誤りが生じた場合には，置換によって生じた誤りの訂正がもっとも重要であり，その際韻母を表す文字同士による置換の誤り訂正よりも声母を表す文字同士の置換による誤り訂正の方が重要である．

3.3.2 置換対

拼音表記された中国語を機械処理する際，ローマ字の置換による単語間の混同を防ぐためには，置換対の情報を用いる必要がある．

語彙 D に含まれる単語を記号列（ローマ字，音素表記など）で表記し，単語 X 中のある表記

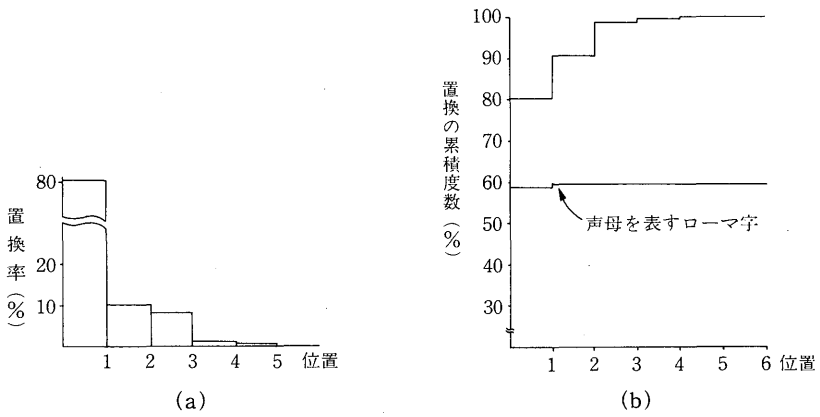

図2. (a) ローマ字表記した場合の一漢字単語における位置別の置換率. (b) ローマ字表記した場合の一漢字単語における位置別の置換の累積度数と声母を表すローマ字の置換の累積度数.

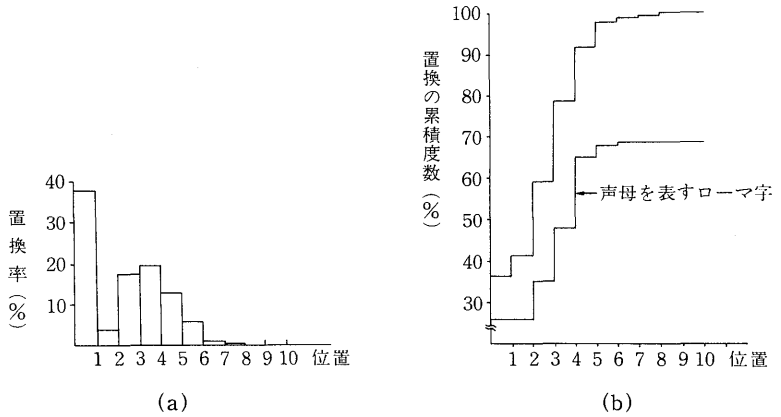

図3. (a) ローマ字表記した場合の二漢字単語における位置別の置換率. (b) ローマ字表記した場合の二漢字単語における位置別の置換の累積度数と声母を表すローマ字の置換の累積度数.

記号を別の表記記号に置き換えて単語 Y を作成した場合に，Y が語彙 D の中に含まれているなら，もとの表記記号と置き換えた表記記号の対を置換対と定義する.

例えば，ある語 X の表記列が abcde で，語彙 D に含まれているとする. この時，単語 X の表記列の d を e に置き換え，単語 Y＝abcee を作った場合に，単語 Y が語彙 D に含まれているなら d と e は置換対をなす. 以下で議論するのは，単語のローマ字綴の間のハミング距離 1 の場合である.

表 10 に，声調と品詞の情報を用いない場合の，各置換対の出現頻度を示す. ただし，四漢字単語 28 語は除いた. 各欄の数値は，その欄を含む行と列に対応する置換対が高頻度単語 6,283 語中に出現した数である. 例えば，a 行と b 列欄の数値 62 は，6,283 語の中で，文字 a（あるいは b）を含んだ単語の表記文字列の a（あるいは b）を b（あるいは a）に置き換えたとき，そのような語が 6,283 語中に 62 語あることを示す.

表 10. 高頻度単語 6,283 語における声調と品詞の情報を用いない場合のローマ字置換対の出現頻度（単位は対）.

	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	ü	w	x	y	z	合計
a	62	27	40	1261	3		301	1072	118		63	28	220	413	33	74	6	46	26	822	235	12	84	105	29	5080
	b	203	399		336	198	173		712	92	443	361	196	13	285	375	95	248	364			362	478	635	206	6236
		c	220	1	116	190	156		185	99	188	117	87	4	101	113	64	964	168			90	128	319	1017	4557
			d	1	262	330	283		540	149	376	227	177	12	179	319	116	294	347			170	373	598	288	5700
				e			25	823					290	231				1		723	192				1	3549
					f	340	330			126	212	294	123	43	169		171	215	264			629		245	178	4056
						g	457		18	261	176	153	111	22	100	52	163	229	258			265	18	417	261	4019
							h		14	228	159	147	79	105	90	35	128	211	214	23		238	14	338	219	3967
								i			1		608	428					1	1557	1454					5944
									j	4	979	291	313		326	1205	43	323	418				1549	1304	211	8553
										k	80	81	47	6	48	10	76	85	110			92	4	233	107	1938
											l	238	192	8	206	550	85	254	310			229	679	807	227	6462
												m	113	6	207	141	77	106	198	1		330	214	298	118	3746
													n	978	95	166	55	114	141	119		90	206	258	84	4862
														o	4		6	15	13	127		27			15	2476
															p	172	47	106	160			192	226	285	95	3126
																q	30	207	233				801	902	138	5523
																	r	80	93	1		140	24	142	93	1735
																		s	230			149	231	436	1896	6440
																			t			199	295	511	226	4779
																				u	23	1		2		3399
																					ü					1904
																						w		182	147	3544
																							x	781	137	6242
																								y	348	9146
																									z	6041

全文字対の1%以上を占める上位から10番目までの置換対を，出現頻度の高い順に並べるとs-z, i-u, j-x, i-ü, j-y, a-e, j-q, a-i, c-z, j-lとなる．

表10でわかるように，26種のローマ字のあらゆる2文字のペアの中で出現頻度がゼロ（置換対にならない）であるものは78対で，これは総2文字ペアの約23%を占める．出現頻度が10に満たない置換対は20対で，総置換対の約6%を占める．

中国語専用のキーボードを設計する際，これらのような情報を考慮すると入力ミスによる単語間の混同を減らすことが可能である．

4. 多字組とエントロピー

英語においてはeの出現頻度が最も高く，また，qの後ろにはuが続くというような規則性がある．現代日本語においては「あ」の次に「る」が続くことが多いが，「あ」の次に「よ」が続くことは少ないことが知られている（栗田・相沢（1984））．このように，ふだん人間が用いている自然言語には，文字（或いは音素）の並びにある種の偏りがある．したがって，自然言語表記

表11. 位置別にみた各ローマ字の出現率（単位は%）.

位置	1	2	3	4	5	6	7	8	9	10	その他	合計
a	.0606	4.7411	3.9976	.6129	.9806	.9829	.5560	.1721	.0300	.0335		12.2178
b	.9942		.1505	.1308	.0754	.0211	.0129	.0066	.0048	.0019		1.4013
c	1.1190		.1127	.1957	.1553	.0548	.0045	.0061	.0047	.0021		1.6582
d	1.6732		.2197	.4349	.1737	.0871	.0181	.0030	.0131	.0027		2.6315
e	.0442	3.1881	1.3202	.5069	.3901	.2797	.1966	.0286	.0149	.0126		5.9946
f	.6964		.0946	.1283	.0825	.0667	.0028	.0020	.0019	.0013		1.0786
g	1.2337	.0020	.1480	1.6086	.8965	.0906	.0329	.0288	.0217	.0086		4.0967
h	1.1931	6.3915	.1342	.8040	.8117	.6138	.1383	.0456	.0212	.0169		10.1955
i		6.4663	1.3752	.7005	.9025	.6483	.1854	.0360	.0247	.0321		10.3992
j	1.9826		.2664	.3928	.2453	.0624	.0190	.0089	.0056	.0033		2.9918
k	.5170		.0440	.0746	.0521	.0095	.0050	.0013	.0009	.0005		.7064
l	.8944	.0025	.2076	.2665	.1679	.0483	.0045	.0059	.0059	.0032		1.6119
m	.6927		.3691	.1877	.1109	.0314	.0193	.0017	.0134	.0013		1.4304
n	.9957	.0307	4.6728	3.1075	.5127	.6587	.6242	.3599	.1032	.0257		11.1523
o	.0123	2.3417	1.4298	.4775	.3746	.2114	.0854	.0339	.0082	.0125		4.9989
p	.2287		.0361	.0300	.0292	.0051	.0021	.0006	.0016	.0006		.3347
q	.6900		.1591	.1535	.1329	.0295	.0048	.0051	.0026	.0033		1.1834
r	.6920	.0408	.1270	.1276	.0722	.0218	.0103	.0085	.0074	.0017		1.1116
s	3.6732		.3055	.4871	.3013	.0583	.0192	.0099	.0085	.0068		4.8772
t	.7868		.1001	.1378	.1021	.0220	.0080	.0039	.0019	.0016		1.1667
u		3.7918	.9659	.4393	.5369	.3214	.1714	.0439	.0392	.0149		6.3589
ü		.0422		.0026	.0034	.0013		.0001	.0001			.0497
w	1.0526		.0867	.2070	.0735	.0611	.0044	.0048	.0027	.0019		1.4973
x	1.4179		.2577	.2768	.2182	.0498	.0069	.0070	.0043	.0035		2.2471
y	2.4981	.0001	.3192	.5144	.2449	.0834	.0123	.0156	.0080	.0126		3.7167
z	3.8904		.2501	.3013	.3353	.0653	.0211	.0089	.0072	.0051		4.8915
合計	27.0387	27.0387	17.1497	12.3064	7.9819	4.5853	2.1653	.8489	.3575	.2101	.3175	100

列の偏りを上手に利用すると，その表記列の誤りの検索と訂正を効率的に行なうことができる (Cornew (1968), Harmon (1972), Riseman and Hanson (1974), Hanson et al. (1976), Hull and Srihari (1982), Shinghal (1983), Suen (1979), 池原・白井 (1984)). 以下では中国語の拼音表記列の中での各ローマ字の位置別出現率，2字組の出現率，各音節の出現率，エントロピーについて分析する．

4.1 各ローマ字の位置別出現率

漢字を拼音で表記した場合，各ローマ字の位置別の出現率を表11に示した．表からわかるように，各ローマ字の位置別の出現率にはかなり偏りがある．たとえば，位置1で出現しない文字はi, u, üの3文字だけである．また，位置2ではa, e, h, i, o, uの6文字の出現率を合わせると約99%にもなり，一方，文字b, c, d, f, j, k, m, p, q, s, t, w, x, y, zは位置2では出現しない．

4.2 多字組の出現率

隣接の2ローマ字組の出現率を表12に示した．2ローマ字組の場合には，単語の先頭と最後には空白があると仮定した．例えば，単語/zhongguo/の記号列はzの前と最後のoの後ろに空白があり，その表記列は/□zhongguo□/であるとした．分析に用いた高頻度単語のローマ字表記列の中で，2ローマ字組の出現率にもかなりの偏りがある．出現率がゼロである2ローマ字組は約55.7%で，出現頻度が高い組から10組並べるとan, i□, e□, g□, n□, u□, sh, □z, □y, ia,

表12. 2ローマ字組の

	Second Letter													
	a	b	c	d	e	f	g	h	i	j	k	l	m	n
a		.0177	.0088	.0158	.0025	.0043	.0131	.0094	1.1695	.0204	.0026	.0316	.0885	3.7647
b	.2375				.1438				.3117					
c	.0942				.0402			.6354	.0914					
d	.6686				1.5676				.3201					
e	.0027	.0194	.0158	.0628	.0029	.0068	.0230	.0266	.7070	.0240	.0012	.0284	.0355	1.9641
f	.3128				.1681									
g	.1577	.0269	.0846	.0794	.3739	.0648	.0767	.0552		.0979	.0217	.0587	.0480	.0170
h	.9313				1.2106				1.7184					
i	1.9981	.0774	.0657	.1366	.3486	.0638	.0693	.1345		.1761	.0236	.0775	.0481	1.3145
j									1.6102					
k	.1404				.1754									
l	.2335				.3320				.5957					
m	.1059				.3938				.2544					
n	.1966	.0430	.0566	.1022	.3343	.0238	4.1195	.0658	.2166	.1159	.0296	.0914	.0496	.0209
o	.0009	.0124	.0284	.0593	.0009	.0141	.0217	.0271		.0624	.0146	.0308	.1046	1.0091
p	.0493				.0329				.0930					
q											.5643			
r	.0995	.0003	.0001	.0020	.3506	.0006	.0001	.0001	.0625	.0001		.0001	.0001	.0008
s	.0798				.0302			2.2553	.1022					
t	.6064				.0190				.2446					
u	.7537	.0227	.0320	.0606	.2541	.0286	.0467	.0092	.4621	.0496	.0089	.0697	.0250	.2186
ü		.0006		.0003	.0071		.0010	.0030		.0001	.0004	.0003		
w	.2117				.4065									
x									1.0491					
y	.3790				.1997				1.1403					
z	.4418				.0730			1.6527	.2126					
□	.0636	.9197	.6591	2.5625	.1148	.4248	.9419	.7939		1.2525	.3440	.8868	.4102	.7090

となる.

英語の場合，26文字に関しての Riseman and Hanson (1974) の統計によると，2文字組では出現頻度がゼロであるのは約35%で，3文字組では出現頻度がゼロであるのは約85%である．Suen の 6,321 語の拼音表記の場合，26ローマ字に関する2～3ローマ字組で出現頻度がゼロであるのはそれぞれ 58.58%，89.95% であり，英語の場合よりやや高いことがわかる．これは中国語の2～3ローマ字組の規則性が英語より強いことを意味している．

4.3 声母，韻母を単位とした音節出現率

各音節の出現率を表13に示した．表において，音節の組み合わせとして本来ありえない部分は空白で示した．また，音節の組み合わせとしては可能であるが，分析で用いた 6,321 語において，出現頻度がゼロである部分を **** で表した．音節の出現頻度の高い順に10位までを並べると de, shi, i, ta, zhi, iu, bu, zai, ji, ren で，出現頻度がゼロである音節は ga, ne, dei, kei, chei, ei, den, reng, pie, chua, shua, rua, zhuai, chuai である．

中国語のこのような特徴を活かし音節の出現頻度を用いた誤り訂正法も考えられる．音節の出現頻度を用いた誤り訂正法と多字組頻度を用いた誤り訂正法の有効性の比較は興味深い課題であろう．

出現率（単位は%）.

						Second Letter						
	o	p	q	r	s	t	u	ü	w	x	y	z	□
	1.3220	.0024	.0092	.0075	.0316	.0132			.0077	.03145	.0327	.0354	1.1227
	.0152						.4319						
	.0598						.0301						.0001
	.1790						.3461						
		.0066	.0101	.1506	.0667	.0081			.0099	.0623	.0738	.0208	3.2534
	.0099						.1410						
	.1802	.0076	.0604	.0352	.0971	.0362	.5207		.0588	.0888	.1265	.1181	2.8210
	.6010						1.2069						.0001
	.0135	.0148	.0628	.0515	.1506	.0469	.2972		.0615	.1062	.1878	.1047	5.2950
							.1887						
	.0512						.0796						
	.0087						.0787	.0266					
	.0306						.0249						
	.0206	.0115	.0588	.0336	.1456	.0655	.0254	.0301	.0636	.0583	.0881	.1415	2.8098
		.0096	.0207	.0160	.0849	.0149	.9882		.0231	.0355	.0802	.0495	1.9939
	.0147						.0106						
							.2573						
	.0286		.0106	.0001	.0001	.0029	.1284		.0002	.0001	.0031	.0039	.1070
	.0264						.2133						
	.1468						.0615						
	1.0250	.0049	.0347	.0351	.0712	.0178			.0154	.0639	.0676	.0660	2.7369
				.0046	.0040	.0003			.0006	.0025		.0020	.0198
	.3150						.1642						
							.2076						
	.5737						.5076						
	.0700						.2699						
	.0105	.1431	.5545	.4680	2.0554	.8727			.8567	.8077	2.1404	2.1779	

表13. 音節（声母＋韻母）の出現率（単位は％）. ＊はゼロ母音を表す.

	a	o	e	ai	ei	ao	ou	an	en	ang	eng	ong	i[l]	i[ ]	er	i[i]	ia	iao	ie	iu
b	.1661	.0564		.2014	.2956	.2144		.2493	.2379	.0518	.0011					.3512		.1342	.0763	
p	.0473	.0546		.0610	.0625	.0192	.0001	.0263	.0091	.0294	.0508					.0602		.0253	****	
m	.1446	.0726	.1788	.0620	.5678	.0626	.0413	.0843	.6872	.0401	.0306					.0751		.0320	.0146	.0003
f	.5045	.0049			.1837		.0318	.1998	.3342	.4586	.1074									
d	.7512		5.6279	.2133	****	.9143	.3263	.3531	****	.2543	.2013	.3393				.6644		.0512	.0098	.0032
t	1.8372		.0678	.2515		.0467	.1451	.0851		.0346	.0029	.4008				.3497		.0982	.0203	
n	.4836		****	.0385	.1415	.0273		.1721	.0013	.0003	1.0625	.0767				.3214		.0114	.0013	.0188
l	.0296		1.1305	.6546	.0875	.1340	.0210	.0282		.0218	.0166	.0114				1.1034	.0036	.0941	.0619	.1738
g	****		1.0987	.1423	.7030	.2073	.0906	.1798	.0892	.0503	.1238	.5794								
k	.0080		.6281	.1521	****	.0601	.0676	.2695	.0222	.0325	.0020	.1229								
h	.0001		.7294	.3647	.0293	.3521	.4691	.0755	.2429	.0355	.0317	.0299								
j																1.4212	.6341	.4129	.6041	.7473
q																.8645	.0042	.0425	.1503	.1239
x																.4616	.3127	.4572	.3331	.0379
z	.0194		.1436	1.4249	.3785	.1533	.1349	.0369	.0690	.0083	.0552	.1255	.7905							
c	.0060		.0514	.2192		.0287	.0013	.0779		.0185	.0979	.2209	.3398							
s	.0068		.0853	.0343		.0062	.0350	.2425	.0221	.0070	.0048	.0634	.3802							
zh	.0125		.9560	.0126		.4865	.1014	.1970	.1503	.1950	.4460	1.1457		1.7453						
ch	.0930		.0779	.0055	****	.0449	.0161	.1054	.0496	.4021	.4933	.0739		.1926						
sh	.0535		.2845	.0038	.0305	.1826	.3988	.1476	.3123	.6934	.6683			4.4522						
r			.0585			.0115	.0209	.3097	1.1894	.0486	.0557	.0855		.2327						
#	.0170	.0064	.0576	.1267	****	.0148	.0379	.1337	.0079	.0015	****				.4924	3.4345	.0808	.5916	.7425	1.736
合計	4.1805	.1949	11.1765	3.9684	1.4724	2.9662	1.9393	2.9737	3.4247	2.3839	3.4522	3.2755	1.5101	6.6231	.4924	9.1073	1.0353	1.9503	2.0144	2.8410

表 13. (つづき)

	ian	in	iang	ing	iong	u	ua	uo	uai	ui	uan	un	uang	ueng	ü	üe	üan	ün	合計
b	.3425	.0082		.2470		1.6061													6.1835
p	.0762	.0552		.1292		.0394													1.1358
m	.3181	.2083		.2978		.0927													4.3533
f						.5244													3.3107
d	.2463			.2153		.2265		.5084		.3631	.1600	.0291							14.4779
t	.2863			.1550		.1278		.0220		.0445	.0309	.0036							4.9513
n	.4193	.0213	.0064	.0053		.0308		.0573			.0062				.1117	.0001			4.1953
l	.1471	.0544	.4265	.1504		.1333		.0388			.0196	.1011			.0840	.0150			6.2478
g						.2237	.0142	1.1126	.0301	.0782	.3639	.0024	.1111						5.5775
k						.0954	.0042	.0220	.0936	.0062	.0201	.0226	.0321						1.9044
h						.1740	.3671	.4000	.0448	.5934	.1628	.0468	.0339						4.5330
j	.7494	.5809	.2826	.5539	.0003										.3346	.1978	.0099	.1596	7.0481
q	.3509	.0873	.1051	.3615	.0081										.5538	.1448	.2326	.0258	2.8561
x	.5321	.3971	.8089	.5240	.0365										.3050	.3398	.0774	.0496	5.5023
z						.1574		.5915		.2378	.0038	.0131							3.962
c						.0207		.0344		.0073	.0002	.0495							1.0241
s						.1555		.4104		.1670	.0440	.0163							1.4713
zh						.4131	.0056	.0289	****	.0181	.0861	.0523	.0933						4.9665
ch						.5934	****	.0004	****	.0174	.1177	.0323	.0476						1.7984
sh						.4089	****	.5213	.0094	.1794	.0009	.0115	.0277						5.6964
r						.3991	****	.0635		.0061	.0066	.0026							1.5086
#	.3676	.4746	.3700	.3314	.3951	.6112	.0069	1.1717	.2030	1.1568	.2597	.3530	.3173	.0017	.9932	.2452	.5369	.1125	7.3094
合計	3.8360	1.8883	1.9997	2.9711	.4400	6.0330	.3980	4.9833	3.8075	2.8752	1.2826	.7359	.6633	.0017	2.3822	.9427	.8569	.3473	100

4.4 エントロピー

条件付きエントロピーの変化から多字組の規則性を考察することが可能である．26個のローマ字において，i 番目のローマ字の出現確率を $p(i)$ $(i=1, 2, \ldots, 26)$ としたとき

$$F_0=-\sum_{i=1}^{26} p(i)\log_2 p(i)$$

を26ローマ字のエントロピーと呼ぶことにする．また，i 番目のローマ字の後に j 番目のローマ字が出現する条件付き確率を $p(j\mid i)$ $(i=1, 2, \ldots, 26;\ j=1, 2, \ldots, 26)$，$i$ 番目のローマ字と j 番目のローマ字が隣接して出現する確率を $p(i, j)$ $(i=1, 2, \ldots, 26;\ j=1, 2, \ldots, 26)$ としたとき

$$\begin{aligned}F_1&=-\sum_{j=1}^{26}\sum_{i=1}^{26} p(j\mid i)\log_2 p(j\mid i)\\&=-\sum_{j=1}^{26}\sum_{i=1}^{26} p(i, j)\log_2 p(i, j)+\sum_{i=1}^{26} p(i)\log_2 p(i)\end{aligned}$$

を1次条件付きエントロピー，同様に

$$\begin{aligned}F_2&=-\sum_{k=1}^{26}\sum_{j=1}^{26}\sum_{i=1}^{26} p(k\mid i, j)\log_2 p(k\mid i, j)\\&=-\sum_{k=1}^{26}\sum_{j=1}^{26}\sum_{i=1}^{26} p(i, j, k)\log_2 p(i, j, k)+\sum_{j=1}^{26}\sum_{i=1}^{26} p(i, j)\log_2 p(i, j)\end{aligned}$$

を2次条件付きエントロピーと定義する．さて，Suenの6,321語を用いて F_0, F_1, F_2 を求めてみる．1ローマ字当たりのエントロピーは

$$F_0=-\sum_{i=1}^{26} p(i)\log_2 p(i)=4.11\ [\text{ビット}\mid\text{文字}]$$

1次条件付きエントロピーは

$$\begin{aligned}F_1&=-\sum_{j=1}^{26}\sum_{i=1}^{26} p(i, j)\log_2 p(i, j)+\sum_{i=1}^{26} p(i)\log_2 p(i)\\&=6.57-4.11=2.46\ [\text{ビット}\mid\text{文字}]\end{aligned}$$

2次条件付きエントロピーは

$$\begin{aligned}F_2&=-\sum_{k=1}^{26}\sum_{j=1}^{26}\sum_{i=1}^{26} p(i, j, k)\log_2 p(i, j, k)+\sum_{j=1}^{26}\sum_{i=1}^{26} p(i, j)\log_2 p(i, j)\\&=8.80-6.57=2.23\ [\text{ビット}\mid\text{文字}]\end{aligned}$$

となる．

英語（Shannon (1951)）の場合には F_0, F_1, F_2 がそれぞれ4.14ビット，3.56ビット，3.3ビットである．中国語を拼音表記した場合のローマ字の各エントロピーを英語と比較して見ると，F_0 はほぼ同じであるが，F_1 と F_2 は中国語の方がそれぞれ約1.1ビット小さい．この情報と4.2節で求めた多字組の出現頻度がゼロとなる割合とを合わせて考えてみると，中国語の場合には，2字組と3字組を用いるなら，訂正の候補文字数を英語より少なく絞ることが可能であることがわかる．

結　　論

中国語を機械処理する際に必要となる拼音表記の文字列に関する情報を得るため，中国語語彙の特徴を十分反映しているものと考えられるSuenの6,321語の拼音表記列を用いて計量分析を行なった．その結果，中国語の拼音表記列の機械処理に必要な統計的な特性を明確にできたと同時に，機械処理に必要な重要な知見が得られた．その結果を簡単にまとめると

(1) 単語の長さの観点からみると中国語の場合は，平均単語長が短いため，中国語の拼音表記列の誤り訂正は英語より不利である．しかし，出現頻度がゼロである2～3字組の割合と条件付きエントロピーの減少から考えると，2～3字組を利用して誤り訂正を行なった場合には，中国語の方が英語より訂正の効率がよいと予測できる．

(2) 一漢字単語の中には，声調と品詞情報を用いてもL_R距離1～2の単語がそれぞれ約9語，40語あり，誤り訂正は極めて難しいといえる．しかし，二漢字単語では声調，品詞情報を用いるとL_R距離1～2の単語はそれぞれ0.35語，1.54語であるため，声調，品詞の情報を有効に利用するとSuenの6,321語の中の二漢字以上からなる単語に関しては二つのローマ字までの誤りの訂正は十分可能である．

(3) L_R距離1～2のうち置換によるものがそれぞれ約70%，50%である．この置換のうち一漢字では約60%，二漢字では約69%が声母を表すローマ字同士による置換である．したがって，置換によって生じた誤りの訂正がもっとも重要であり，韻母を表す文字同士による置換の誤り訂正よりも声母を表す文字同士の置換による誤り訂正が重要である．

(4) 26ローマ字からなる文字ペアのうち，78ペアが出現頻度がゼロ（置換対にならない文字ペア）で，これは総ローマ字ペアの23%を占める．

(5) 単語の拼音表記列の頭から2番目の位置におけるa, e, h, i, o, uの6文字の出現率は99%を占め，また文字b, c, d, f, j, k, m, p, q, s, t, w, x, y, zは2番目の位置には出現しない．

(6) 単語の拼音表記列で隣接する2～3ローマ字組で出現頻度がゼロであるのはそれぞれ約59%，90%で英語の場合よりやや高い．

以上のような情報を有効に利用することによって中国語の拼音表記列の誤り訂正を容易にしかも精度良く行なうことが可能となると考える．

本研究の延長としては，中国語の特徴を活かした「声母＋韻母」の音節の出現頻度を用いた誤り訂正法と多字組の出現頻度を用いた訂正法の有効性の比較，単語数が数万語となった場合の機械処理の可能性や統計データを用いた効率の良い誤り訂正のアルゴリズムの研究などが考えられる．

謝　　辞

本研究は，著者の一人金が，神戸大学田中研究室および宇都宮大学徳田研究室に在籍中に行なった，機械処理を目的とした中国語高頻度単語の計量分析の研究をさらに発展させたものである．御指導下さった神戸大学教授田中栄一氏，宇都宮大学教授徳田尚之氏，有益なコメントをくださった東京女子大学名誉教授水谷静夫氏，学習院大学教授田中章夫氏，元日本工業大学教授水野坦氏ならびに二人のレフリーに厚くお礼を申し上げる．なお，本研究を行なうにあたっては，足銀国際交流財団より研究助成金の援助を受けた．同財団に深く感謝する．

参 考 文 献

Cornew, R.W. (1968). A statistical method of spelling correction, *Inform. and Control,* **12**, 79-93.

Damerau, F.J. (1964). A technique for computer detection and correction of spelling errors, *Comm. ACM*, **7**, 171-176.

Hanson, A.R., Riseman, E.M. and Fisher, E. (1976). Context in word recognition, *Pattern Recognition*, **8**, 35-45.

Harmon, L.D. (1972). Automatic recognition of print and script, *Proceedings of the IEEE*, **60**, 1165-1176.

橋本万太郎 訳 (1987). 『中国語動詞の研究』, 白帝社, 東京.

Hull, J.J. and Srihari, S.N. (1982). Experiments in text recognition with binary n-gram and Viterbi algorithms, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **4**, 520-530.

池原 悟, 白井 論 (1984). 単語解析プログラムによる日本文誤字の自動検索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, **25**, 298-305.

今栄国晴 (1960). 日本語の digram の相対頻度とその特性, 心理学評論, **4**, 85-100.

板橋秀一, 鈴木久喜, 城戸健一 (1971). 日本語の高頻度単語の音形上の幾つかの特徴, 電子情報通信学会論文誌, **54-C**, 428-434.

Jelinek, F. (1976). Continuous speech recognition by statistical methods, *Proceedings of the IEEE*, **64**, 532-556.

香坂順一 (1989). 『現代中国語辞典』, 光生館, 東京.

金 明哲, 徳田尚之, 村上征勝, 田中栄一 (1992). 音声学的観点からの中国語高頻度単語の計量分析, 行動計量学, **19**, 49-65.

栗田泰一郎, 相沢輝昭 (1984). 日本語に適した単語の誤入力訂正法とその大語彙単語音声認識への応用, 情報処理学会論文誌, **25**, 831-841.

Levenshtein, V.L. (1965). Binary codes with correction of deletions, insertions and substitutions of symbols, *Dokl. Akad. Nauk SSSR*, **163**, 845-848.

牧野正三, 城戸健一 (1979). 近距離単語間の識別に必要な音素対の性質, 電子情報通信学会論文誌, **62-D**, 507-514.

牧野正三, 脇田 寿 (1986). 英語高頻度単語解析, 電子通信学会音響研究会論文, **86-12**, 45-51.

丸山活輝 (1989). HMM 音韻連結と NETgram を用いた英単語音声の認識, 信学技報, SP89-89.

松本丁俊 (1986). 『中国語音声学概論』, 白帝社, 東京.

猛 子 (1990). 我国計算機漢字入力技術的現状与發展, 人民日報 (海外版) 5月25日.

中村雅巳, 鹿野清宏 (1991). ニューラルネットによる英単語品詞列予測モデル, 電子情報通信学会論文誌, **74-D-II**, 1-7.

Okuta, T., Tanaka, E. and Kasai, T. (1976). A method for the correction of garbled words based on the Levenshtein metric, *IEEE Trans. Comput.,* **C-25**, 172-178.

北京語言学院 編 (1986). 『新中国語』, 中華書店, 北京.

Riseman, E.M. and Hanson, A.R. (1974). A contextual postprocessing system for error correction using binary n-grams, *IEEE Trans. Comput.,* **C-23**, 480-493.

坂本 仁, 大塚正人, 細野直恒, 巽 和弘, 田中 章 (1988). 日英/英日機械翻訳システム PENSEE, bit (機械翻訳), 9月号, 151-157.

Shannon, C.E. (1951). Prediction and entropy of printed English, *Bell Syst. Tech. J.*, **30**, 55-64.

鹿野清宏 (1987). Trigram model による単語音声認識結果の改善, 信学技報, SP87-23.

Shinghal, R. (1983). A hybrid algorithm for contextual text recognition, *Pattern Recognition,* **16**, 261-267.

Shinghal, R. and Toussaint, G.T. (1979). A bottom-up and top-down approach to using context in text recognition, *Int. J. Man-Mach. Stud.,* **11**, 201-212.

Suen, C.Y. (1979). n-gram statistics for natural language understanding and text processing, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **1**, 164-172.

Suen, C.Y. (1986). *Computational Studies of Most Frequent Chinese Words and Sounds*, World Scientific, Singapore.

田中栄一, 菊地良昭 (1980). 図形間の距離, 電子情報通信学会論文誌, **63-D**, 1018-1025.

Xu, Y. (1989). On semanteme and syntax-based analysis system of Chinese language, *Journal of Chinese Information Processing*, **3**(3), 34-42.

Statistical Characteristics of Most Frequently Used Chinese Words as Written in Pin-yin

Mingzhe Jin

(Department of Statistical Science, The Graduate University for Advanced Studies)

Masakatsu Murakami

(The Institute of Statistical Mathematics and
The Graduate University for Advanced Studies)

Representing Chinese words by Roman characters, call pin-yin, is widely done in China. However, there are many uncertainties about the statistical properties of pin-yin to give useful information on handling Chinese words by word processing machines such as typewriters, word processors and computers. Our study was based on the most frequent Chinese words in Suen's paper (1986). The statistical properties of pin-yin we studied are as follows: (1) distribution of word length, (2) distribution of short-distance words based on Roman letters, (3) distribution of parts of speech, (4) frequency distribution of short-distance word based on Roman letters and the effects by parts of speech and tone, (5) primary and secondary conditional entropies of Roman letters, (6) substitution pairs of Roman letters, (7) frequency distribution of unigrams and bigrams based on Roman letters.

Key words: Most frequently used Chinese word, pin-yin, Roman letter, tone, part of speech, machine processing, entropy, short-distance word, error correction, substitution pair.