日本語コーパスを援用した，タブレットPC用新文字体の開発

－ 第一報 平仮名文字の筆記効率について －

堀 内 泰 輔*

Development of the New Character Style for Tablet PC using the Japanese Corpus - (1) Writing Efficiency of Hiragana Character -

HORIUCHI Taisuke

In recent years, ubiquitous computer society has come with the high performance and low-pricing of the personal computer and expansion of the availability of the Internet environment. In this research, the tablet PC as a future PC style is observed. Especially we aim at development of the system which a Japanese input can input efficiently.

As a result of the fundamental research, this paper reports the analysis result about the writing efficiency of a hiragana character in which the Japanese corpus was used.

キーワード：コーパス，タブレットPC，文字の歴史，文字認識，Nグラム分析

1．ま え が き

近年，インターネット環境の広まりやパソコン（以下，PCと略す）の高性能・低価格化に伴い，ユビキタスなコンピュータ社会が到来している．

人と PC とのインターフェースについて考察すると，以前のいわゆるCUI環境，つまりキーボードを入力機器とする利用形態から，昨今はマウスに代表されるGUI環境へと移行してきた．しかし，文字入力の際にはキーボードが未だ中心である．音声入力方式も開発され認識精度を向上させてはいるが，疲労やPCの利用環境の観点からは，歓迎されるものには至っていない．

ところが，昨今の情報機器の使われ方を見ると，これまでのPC一本やりの利用方法から，PDAや携帯電話といった，フルキーボードやマウスを持たない製品の利用への共有が多くなってきている．特にPDAの入力手段はタブレットへのペン入力であり，アメリカ製品などではキーボード入力に匹敵する効率的な入力が可能になっている．また，タブレット入力に特化したPCやOSなども登場してきており，いわゆる「タブレットPC」のシェアが今後飛躍的に拡大することが予想される．

本研究においては，これからのPCスタイルとしてのタブレットPCに注目し，特に日本語入力が効率的に入力できるようなシステムの開発を目指している．

本論文では，その基礎的研究の成果として，日本語コーパスを用いた平仮名文字の筆記効率に関する分析成果について報告する．

2．日本語における文字記述の実態

現在の日本においては，文字の記述には，漢字，平仮名，片仮名，英字，各種記号が用いられる．このうち，漢字以外の文字については画数が少なく複雑さも低いので省略することなく筆記されるのに対して，漢字の筆記では画数が多く頻度の高い文字については伝統的な省略記法が用いられる場合がある．たとえば，「権」を木偏に又，「曜」を日偏に玉などとする．仏教の分野でも古来から独特の省略法が見られる．この傾向は中国においては公に徹底されており，従来の漢字（繁体字）を簡略化した簡体字が，1977年12月の「新簡体字表」が公布されて以来用いられている．このような傾向は，情報内容の多様化と複雑化に伴って，古今東西を問わない一般的なものと考えられる．

よって，「文化の発展に伴って，文字の筆記効率も高くなるように文字の形状そのものも変移する．」という仮説を設定できよう．本研究ではこの仮説を日本語において実証することをひとつの目的としたい．

＊ 一般科助教授

原稿受付 2003年5月19日

3．文字の簡略化の歴史

日本では４世紀ごろまでは文字を持っていなかったが，その後中国から漢字が輸入されて日本人が用いる最初の文字となった．当初は漢字を表音文字として用い，ここに万葉仮名が成立していった．しかし，これは漢字の原型をそのまま利用しているため，画数の少ない漢字を選ぶにしても，長文の記述には不向きであった．

そこで登場した手法が草化と省文である．前者は漢字の草書体を簡単な字形に崩す方法でありこれは平仮名文字を生み出すことにつながった．また，後者は漢字の一部のみを利用する方法であり，片仮名文字を生み出した．

この両者が分化・統一の道程の末に，現在の仮名文字に定着したわけであるが，明治時代以降には，さらに筆記効率をあげるべく，速記文字が創出され一部の分野では現在も用いられている．

4．平仮名文字の複雑さとその測定

このように，平仮名文字は１０世紀初頭以来連綿として用いられてきたが，その形状はほとんど変化していない．とすれば，情報環境が全く異なる現代において，果たして平仮名文字の筆記効率は優れている，といえるのだろうか．ここでは，平仮名文字の複雑さについて検討してみたい．

この関連の先行研究としては文献 1)と 2)がある（以降，「樺島研究」と略す．樺島研究においては，平仮名文字の複雑さを，筆数(H)，交点の数(K)，単純化した曲がり(M)の数の３つのファクターで得られる関数として記述している．これらから得られる複雑さの尺度 r は各平均値(Hh,Kh,Mh)と各標準偏差(Hs,Ks,Ms)から以下の式で求められる．

$$R=(H-Hh)/Hs+(K-Kh)/Ks+(M-Mh)/Ms$$

これを，複雑さの昇順に**表１**に示す．

5．平仮名文字の筆記効率について

上記の複雑さは，文字の形態の尺度としては妥当なものと考えられるが，実際の筆記では，筆が宙に浮いている（これを「空筆」と呼ぶことにする）時間を無視することはできない．その意味で実際の筆記時間や筆の長さなど測定する必要があろうと考え，PC 用タブレットを用いた筆記文字の距離と時間を測定するためのプログラムを試作し，筆者自身によるデータ収集を行った（**図１**）．

表１ 平仮名文字の複雑さ（樺島）

文字	H	K	M	r
く	1	0	1	-3.3
し	1	0	1	-3.3
つ	1	0	1	-3.3
へ	1	0	1	-3.3
い	2	0	0	-2.8
こ	2	0	0	-2.8
り	2	0	0	-2.8
て	1	0	2	-2.6
う	2	0	1	-2.2
ひ	1	0	3	-1.9
ろ	1	0	3	-1.9
ん	1	0	3	-1.9
に	3	0	0	-1.7
ら	2	0	2	-1.5
そ	1	0	4	-1.3
と	2	1	1	-1.3
の	1	1	3	-1.0
け	3	1	0	-0.8
さ	3	1	0	-0.8
ち	2	1	2	-0.6
え	2	0	4	-0.2
か	3	1	1	-0.2
や	3	1	1	-0.2
ふ	4	0	1	0.0
た	4	1	0	0.3
よ	2	2	2	0.3
る	1	1	5	0.3
せ	3	2	1	0.7
も	3	2	1	0.7
す	2	2	3	1.0
み	2	2	3	1.0
ゆ	2	2	3	1.0
き	4	2	0	1.2
は	3	2	2	1.4
を	3	2	2	1.4
め	2	3	3	1.9
わ	2	3	3	1.9
お	3	2	3	2.1
ま	3	3	2	2.3
な	4	2	2	2.5
れ	2	3	4	2.5
ほ	4	3	2	3.4
む	3	2	5	3.4
あ	3	4	3	4.0
ぬ	2	4	5	4.2
ね	2	4	5	4.2

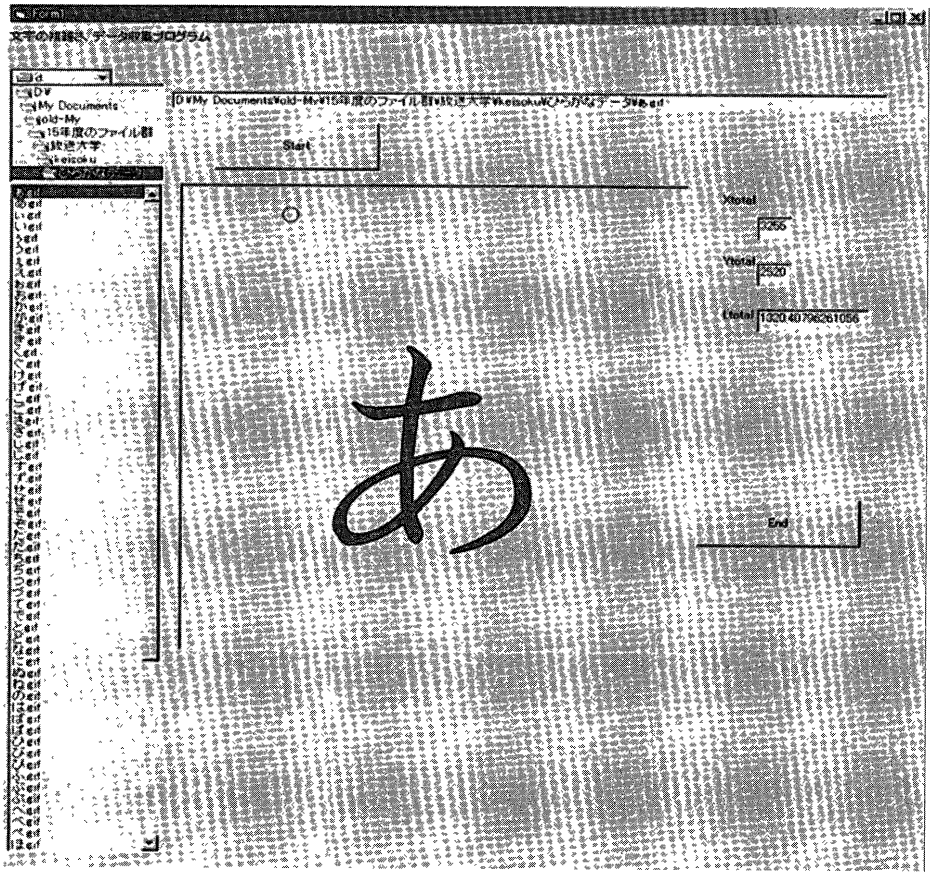

図１　筆記測定画面例

文字を選択すると教科書体の平仮名文字が表示されるので，Start ボタンをクリック後タブレット上をドラッグ（筆記）することで，各筆の長さと時間が計測できる．なお，筆から筆への空中での動き（空筆）の長さと時間も重要と考え，これも同時に測定可能とした．筆者自身による計測結果を**表２**に示す．本表では，筆記所要時間が長かった順に示した．

次に，樺島データと筆者データ，両者の比較を行った．この結果をグラフとして，**図２**に示すが，この図において，平均順位とは，筆の長さ（実筆＋空筆）の順位と所要時間の順位を単純平均したものである．これによると，「か」，「さ」，「た」，「え」　などは樺島研究では順位が低くなっている．これにより，幾何学的な複雑さは低くても，実際に筆記する際の空筆にかかる時間を考慮すると複雑さが上昇する文字が存在する，ということがわかる．今回の調査目的は，文字の筆記効率を問題にしているから，筆記時間のデータ採取は有意義であったと考えられる．

６．日本語コーパスによる，筆記効率の評価

以上により，平仮名文字の複雑さの尺度が凡そではあるが得られたことになる．それでは，現代日本語において，平仮名の出現頻度と複雑さとの関係が適正なものであるかが問題となる．そこで，この評価を行うために，日本語コーパスを利用して頻度データを得ることとした．現在，インターネット上には電子化されたテキストが豊富に存在するが，このうち，日本語の平均的な記述を行っていると考えられる新聞記事を対象に日本語データを収集することにした．

日本語の新聞に関しては，各紙とも有料の文献検索サービスを行っているが，日本語コーパスとしての大量のデータを収集するには不向きである．しかし，地方新聞ではあるが佐賀新聞が，ここ約１０年分のほとんどの記事を無料で検索できる Web ページがあることがわかった．これをコーパス化すれば，新しい日本語表現を豊富に含むコーパスを利用できることになる．今回は，これを達成すべく，下記のような順序で作業を行った．

（１）全記事のダウンロード

（２）テキストファイル化

（３）文字のモノグラムの作成

６－１　全記事のダウンロード

佐賀新聞記事データベースは**図３**のような画面構成

表２　平仮名文字の計測結果

ひらがな	画数	長さ順位 合計順位	長さ順位 実筆順位	長さ順位 空筆順位	時間順位 合計順位	時間順位 実筆順位	時間順位 空筆順位
あ	3	16	5	41	1	1	21
ぎ	6	6	31	4	2	11	1
が	5	2	22	1	3	3	4
ぼ	6	1	3	3	4	15	3
ぽ	5	3	4	9	5	8	10
ば	5	8	8	13	6	9	8
ず	4	4	15	6	7	10	9
げ	5	5	23	5	8	25	7
ぁ	3	43	35	54	9	2	35
ぱ	4	9	7	15	10	6	20
だ	6	7	39	2	11	47	2
お	3	17	21	19	12	5	25
ざ	5	13	50	7	13	37	5
ぜ	5	15	20	11	14	17	13
ぢ	4	20	30	18	15	16	15
ぷ	5	29	48	17	16	29	11
ぶ	6	23	56	10	17	39	6
か	3	12	38	8	18	19	17
ほ	4	14	12	16	19	27	19
ぬ	2	19	6	46	20	4	61
を	3	31	24	33	21	12	43
ね	2	10	1	32	22	7	56
き	4	32	44	20	23	45	12
ぉ	3	49	54	42	24	24	29
ど	4	28	52	14	25	38	16
な	4	34	36	28	26	32	24
え	2	47	40	60	27	13	51
は	3	21	10	37	29	22	42
ぞ	3	25	18	30	28	26	40
び	3	39	25	45	30	30	44
れ	2	11	2	34	31	14	68
ぇ	2	74	72	64	32	28	50
む	3	22	14	31	33	41	26
ゆ	2	33	13	50	35	23	54
す	2	40	27	49	34	20	55
ま	3	26	28	26	36	34	38
け	3	27	19	36	37	35	37
わ	2	36	16	52	38	21	60
ぴ	2	35	11	62	40	18	65
せ	3	38	34	35	39	46	28
た	4	18	42	12	41	68	14
ふ	4	50	70	23	42	60	18
ご	4	41	59	22	43	59	22
め	2	24	9	48	44	33	63
で	3	48	51	43	45	48	45
ぺ	2	63	61	55	46	53	39
る	1	45	17	78	47	31	78
も	3	37	46	21	48	61	33
ゃ	3	60	64	40	49	72	23
よ	2	53	45	59	50	43	57
づ	3	44	58	24	52	64	32
ぐ	3	58	65	38	51	58	41
や	3	30	29	29	54	62	36
さ	3	52	62	39	53	66	30
わ	2	64	55	65	55	42	64
じ	3	61	66	44	56	69	31
い	2	73	74	53	57	57	47
ゅ	2	57	49	58	58	49	58
ち	2	55	43	67	59	51	62
そ	1	56	37	72	61	36	72
べ	3	62	76	27	62	74	27
ぃ	2	78	80	57	60	65	46
み	2	46	26	69	63	50	69
の	1	51	32	75	64	40	75
と	2	67	67	61	65	63	52
ょ	2	68	69	63	66	54	66
ら	2	66	57	68	67	55	67
ひ	1	54	33	76	68	44	76
に	3	42	53	25	69	75	34
う	2	69	73	51	70	70	49
り	2	70	63	66	71	67	59
ろ	1	59	41	79	73	52	79
ぅ	2	76	79	47	72	76	48
ん	1	65	47	80	74	56	80
こ	2	71	71	56	75	78	53
く	2	79	77	70	76	71	70
て	1	72	60	74	77	73	74
つ	1	75	68	73	78	77	73
し	1	77	75	71	79	79	71
へ	1	80	78	77	80	80	77

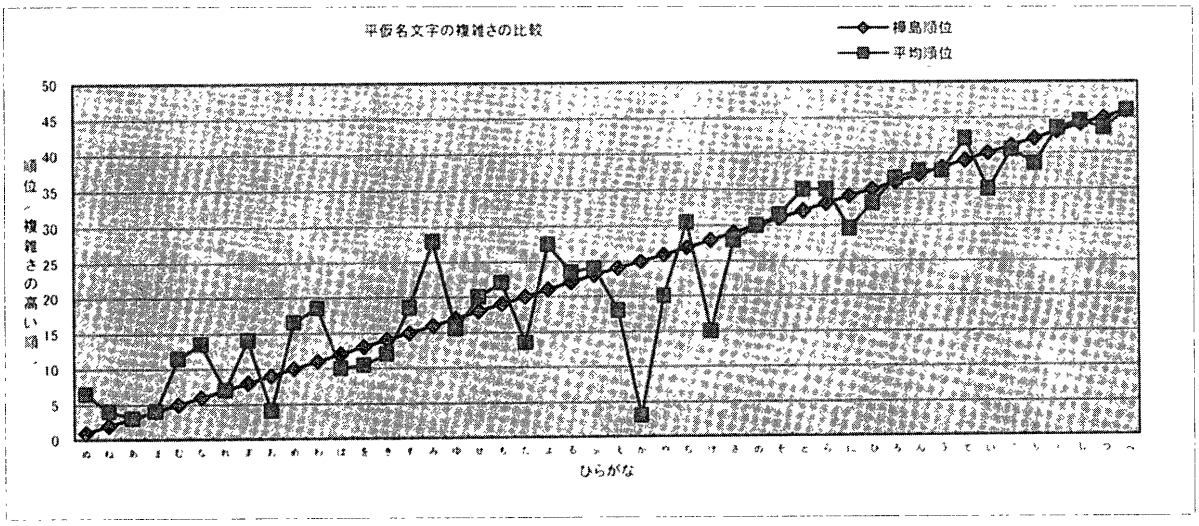

図２　ひらがな複雑さの比較

図 3 佐賀新聞記事検索 Web 画面（http://www.saga-s.co.jp/pubt2002/ShinDB/）

になっており，外国通信社配信の一部を除きすべて蓄積しており，１９９４年以降の佐賀新聞掲載記事を簡単な操作で検索できる．ただし，利用方法はこのような会話式の利用に限られ，該当記事の表示は１０件単位でしか行えない．

そこで，キーワードに全記事に共通するであろう句点（「.」）を入れ，検索対象を「全年指定」に，記事のジャンルを「全面指定」にして検索したところ，437,181 件の合致を得た．

次に，各記事をすべてダウンロードするための苦肉の策として，検索結果の URL を分析した．その結果，次々に１０件単位で表示されるページの各 URL にはほぼ一定の規則があることがわかり，すべての URL（約 43,000 個）を Excel で自動生成し，これに UNIX の wget コマンドをオプションとともに付加することで，ひとつのシェルスクリプトを得た．

これを実行した結果，約 100 時間を経てすべての記事をダウンロードすることに成功した．これは，ファイル数（記事数）約４２万個，フォルダ数 1,286 個，全容量 495MB（ディスク上では 1.52GB）という巨大なものである．

６－２　テキストファイル化

ダウンロードしたのは当然 HTML ファイルであるから，HTML タグを削除して完全なテキストファイルにする必要がある．この HTML ファイルの title タグには，掲載年月日とジャンルが記載されているため，この情報をファイル名に反映させることにした．こうすることで，年月単位やジャンル単位の記事検索がファイル名で行えるため，今後の分析に有利となる．

今回ダウンロードしたものは，1994 年 1 月 1 日から 2003 年 4 月 12 日までの記事であるが，1999 年 12 月 28 日を境にして，ファイルの管理方法（ディレクトリ構造）が異なることが判明した．古い記事は年のディレクトリ下に月のディレクトリを作りそこにその月の全記事を収納していたが，新しい記事では，月の下にさらに日のディレクトリを作り，その日の記事を入れるようになっており，ディレクトリの細分化が行われていた．

そこでこれに対処し，さらにファイル操作を容易にするために，全記事をひとつのディレクトリに収納す

ることとした．しかし，42万ものファイルをひとつのフォルダで管理することは，現行の諸OSでは論理的には可能ではあったが，実際にアクセスをしてみると，数分から数十分のアクセスタイムを要した．これでは現実の利用は無理なため，前述の古い記事の収納方法に習い，月単位でのファイル管理を行うこととした．

６－３　文字のモノグラムの作成

次にモノグラム分析を行って，各文字単位の出現頻度を求める工程となる．ただ，１０年間の各単年別，そして，ジャンル別の分析も行えるので，モノグラム分析に先立ち，1994～2003 年の各単年ごとにその全記事を統合する作業を行った．ジャンルに関しては，計 13 種類の統合ファイルを作成した．この他に，全記事を統合したファイルも用意した．

以上でモノグラムにかける準備ができた．しかし当初入手したプログラムではあまりにも時間がかかりすぎることが判明した（試算では全体の実行に約２００時間を要する）．このため，このプログラムを分析したところ，処理のほとんどをPerlインタプリタで行っていることがわかり，他のフリーソフトを探索した結果，Windows環境ではあるが，C++で書き換えを行って数十倍の高速化を行ったものを入手することができた．これ利用して，所要約１０時間でモノグラム処理を終えた．

６－４　文字モノグラムの分析とその結果

（１）文字種の統計

最初に，文字種の統計を**表３**に掲げる．新聞記事のほとんどは全角であるものの，半角では英字，数字，カタカナも使われていること，微量のロシア文字は英字のミス入力であること，などがわかった．

数字は半角，全角とも使われている．各数字ではほぼ半々になっているが，「０」だけは，344,127（半角），677,695（全角）と全角が２倍近い．

（２）文字種別のモノグラム

次に，平仮名，片仮名，漢字ごとの４０位までの順位表を**表４**に示す．これによると，平仮名では「の」が群を抜いて多いことがわかる．片仮名については外来語に多いと思われる「ン」がこれも断然トップとなった．漢字では比較的画数の少ない，複雑さが低い文字が上位を占めることを予想していたが，ほぼその通りになったと思われる．

このほか，年次別の文字種割合の推移，年次別文字種別順位の推移，ジャンル別の，文字種割合の比較なども行った．

表３　文字種別文字数統計

種類	出現度数	%
全角	273,213,754	94.892
半角	14,707,549	5.108
総計	287,921,303	100.000
文字種		
半角カナ	3,822	0.001
全角カナ	19,978,484	6.939
かな	97,145,550	33.740
ギリシャ文字	278	0.000
ロシア文字	36	0.000
英小文字	125,893	0.044
英大文字	1,265,262	0.439
漢字1	124,141,388	43.116
漢字2	187,813	0.065
漢字3	19,760	0.007
記号	37,780,207	13.122
罫線記号	22	0.000
数字	7,272,788	2.526
総計	287,921,303	100.000

７．平仮名文字の筆記効率の分析

以上で，平仮名文字の複雑さのデータと，出現度数データの両者が得られたので，これらをつき合わせ，現在の平仮名文字の効率性について分析した．

図４は，横軸に佐賀新聞での平仮名文字の出現順位を取り，筆者データの長さ順位（●）と時間順位（■）との関係をグラフ化したものである．もし，平仮名文字の筆記効率が高いと仮定すると，このグラフは右上がりになるはずである．

しかし，この結果に見るようにほぼ全体にばらついていることがわかる．相関係数を算出したところ，長さ順位では 0.06，時間順位では 0.26 となり，ほとんど相関は見られなかった．

つまり，「た」（出現頻度第３位）のような出現頻度が高いにもかかわらず複雑さが大きかったり，反対に，「ひ」（出現頻度第５５位）に代表されるあまり使われない文字の字形が単純であったりするのである．このため，現代日本語の文章に表出する平仮名文字は，その筆記効率からみてかなり非効率的であることが伺われる．

表4　佐賀新聞記事に表出する，平仮名，片仮名，漢字別上位４０文字

文字	出現度数	全体順位	文字	出現度数	全体順位	文字	出現度数	全体順位
の	8,970,801	3	ン	1,600,148	32	日	1,767,341	28
に	5,201,172	6	ス	1,073,346	49	一	1,324,021	36
た	4,980,390	7	ル	899,119	53	十	1,287,398	38
を	4,806,745	8	ト	841,229	55	人	1,145,569	41
い	4,583,986	9	イ	725,218	64	会	1,142,612	42
は	4,423,680	10	ラ	659,625	69	年	1,123,970	43
と	4,140,482	11	ア	619,559	72	大	1,111,313	45
し	4,088,848	12	リ	608,042	73	二	1,077,250	47
る	4,077,711	13	ッ	572,932	78	国	816,909	56
が	4,027,102	14	ク	568,311	80	中	814,024	57
で	3,843,283	15	タ	451,370	107	三	806,804	58
な	3,378,106	16	シ	417,032	116	本	793,482	59
て	3,290,035	17	カ	396,008	124	市	743,655	61
か	2,431,483	19	ド	385,367	130	長	674,970	68
っ	2,058,769	21	プ	351,590	143	同	651,281	71
ら	2,027,440	22	フ	345,087	145	出	605,430	74
れ	1,978,163	23	ロ	340,472	150	事	599,640	75
り	1,783,781	26	レ	334,234	152	者	589,438	76
も	1,777,983	27	ジ	324,904	161	子	586,473	77
す	1,606,070	31	テ	323,555	162	五	572,563	79
う	1,439,648	33	ム	318,464	165	時	566,753	81
さ	1,420,008	34	コ	305,903	173	生	557,026	85
ま	1,347,178	35	マ	305,042	174	月	550,581	86
こ	1,300,667	37	チ	284,574	188	上	546,701	88
き	1,183,371	39	バ	280,752	191	自	539,237	90
く	1,146,050	40	サ	272,424	197	九	535,809	91
ん	1,123,553	44	グ	262,083	204	合	530,369	92
だ	1,080,465	46	ニ	226,714	233	分	527,434	93
ど	1,076,592	48	メ	216,712	237	佐	526,638	94
め	1,009,145	51	ウ	215,774	239	田	521,230	95
け	941,270	52	オ	213,520	242	町	512,207	97
あ	866,711	54	ナ	211,335	245	行	509,164	98
や	785,742	60	ビ	207,130	252	県	494,284	99
え	732,483	63	パ	205,658	253	業	492,342	100
よ	691,925	65	ブ	202,698	261	四	491,752	101
つ	689,958	66	エ	200,925	267	賀	487,623	102
ち	561,739	84	セ	198,586	268	後	481,987	103
わ	547,232	87	ィ	189,244	281	地	469,810	104
そ	544,592	89	デ	187,821	283	前	462,881	105
み	513,469	96	ャ	184,241	292	学	447,631	108

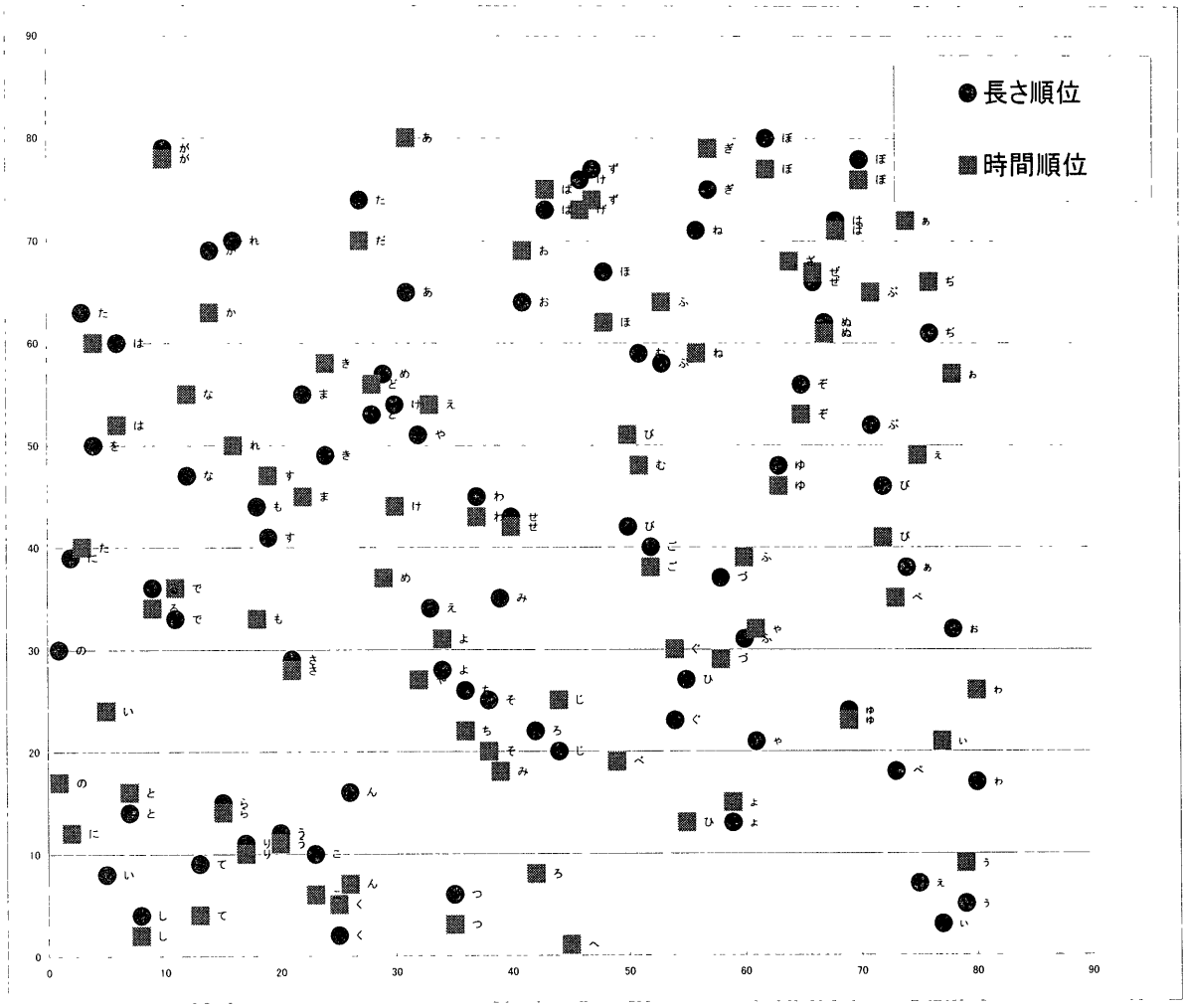

図4 平仮名文字の出現順位（横軸）に対する，長さ・時間順位

8．あとがき

本論では，日本語で用いられている文字のうち平仮名文字に注目し，新聞記事での出現頻度と文字の複雑さとを比較することで筆記効率の分析を行った．その結果，平仮名文字の筆記効率が極めて悪いことが判明した．このことは，手書きで平仮名文字を記述する上での大きな問題点であり，タブレットPCにおいて日本語のうち，現行の平仮名文字をそのままの形で認識していることは，入力効率の面でロスが多いことになる．

本研究の次の段階では，この結果に基づいて，日本人にとって覚えやすく，コンピュータからは誤認識されにくい平仮名文字体の開発を行っていくこととしたい．

参考文献

1) 「日本の文字」，樺島忠夫，岩波新書７５，1979 p. 79
2) 「図説日本語」，林大監修，角川小辞典９，1982 p. 224-225
3) 「文字」，岩波講座日本語８，岩波書店，1977
4) 「文字の歴史」，ジョルジュ・ジャン，創元社，1990
5) 「最新式グレッグ速記法」，小池喜三郎，研究社，1986
6) 「速記と情報社会」，兼子次生，中公新書1476，中央公論社，1999
7) 「日本語百科大事典」，金田一春彦他，大修館書店，1988