日本語の漢字比率と平均情報量について

Kanji Ratio and Linguistic Entropy of Japanese

横　原　恭　士

はじめに

文字によって情報を表現する場合、表音文字、表意文字、表音文字と表意文字の組合せなど言語によっていくつかの型がある。情報量の点から言うと、その言語で使用可能な文字数（基本文字数と言う）が多いほど、1文字あたりの平均情報量は多くなる。英語など表音文字であるアルファベットだけで情報を表す言語では、基本文字数は数十個と少なく、したがって1文字当りの平均情報量も少ない。表意文字だけの中国語では数千の漢字を使うので1文字当りの平均情報量は高い。日本語は71個の仮名と約2千個の漢字を使うので、平均情報量は中国語よりは低いが英語よりは高い。

本稿では表意文字と表音文字を混用する言語である日本語についての考察を、仮名と漢字の混在比率や平均情報量の点から行う。

1．基本漢字数と漢字・仮名・空白の混在比率

日本語では仮名と常用漢字1945個を基本文字数として文章を書く。この時のある内容の文章中の全文字数に対する漢字の文字数の比率（漢字の混在比率＝漢字比と言う）は、いろいろな資料から得ることが出来る。この漢字比は常用漢字が全て使用可能である通常の文章ではある範囲内に入るであろうが、使用可能な漢字の数（基本漢字数と言う）を常用漢字数以下に制限すると、その制限された基本漢字数により影響を受けるであろう。

このことを確かめるため、基本漢字数が常用漢字数以下に制限されている［小学校・中学校の教科書］の65の文章の基本漢字数と漢字比、および常用漢字が全て使用できる時の［天声人語 '93春］の4つの文章の漢字比を調べた。また同じ69の文章の基本漢字数と仮名の混在比率（仮名比と言う）、基本漢字数と空白の混在比率（空白比と言う）との関係も調べた。

1．1　基本漢字数と漢字比

基本漢字数と漢字比との関係は基本漢字数が0から650付近までは正比例しているが、基本漢字数が650以上ではその関係が変化する。

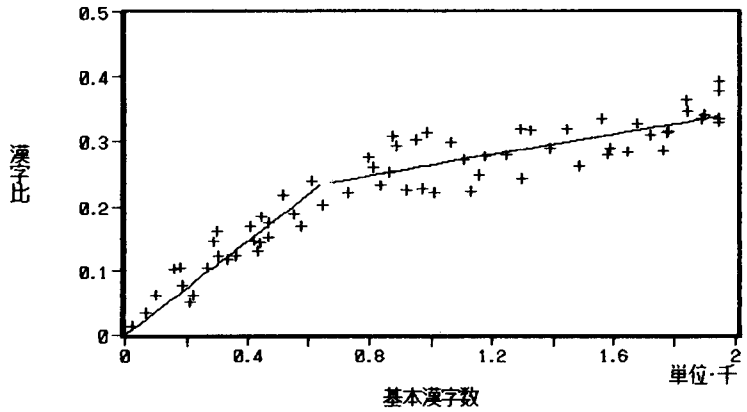

図1　基本漢字数と漢字比

〔基本漢字数が650までの基本漢字数と漢字比の関係〕

漢字比＝0.000367×基本漢字数　　(1)

〔基本漢字数が650～1945での基本漢字数と漢字比の関係〕

漢字比＝0.0000799×基本漢字数＋0.184　　(2)

これらの式から、漢字の平均の使用確率は基本漢字数が650までは0.0367％で一定であるが、650以上では0.0367％から徐々に減少し、常用漢字数を全て対象にする通常の文章では0.0175％となることが分かる。

基本漢字数と漢字比の間に（1）式と（2）式の2つの関係があることは、次のように考えられる。

①（1）式の関係

ある内容のかなり長い文章や文書中で使用対象となる漢字数（使用漢字数と言う）と漢字比の関係が（1）式で与えられる。ある内容の長い文章や文書中で使用する漢字は常用漢字を全て使用するのではなく、内容に関係のある漢字だけが使用対象となるのである。この使用対象となる漢字の平均の使用確率は常に0.0367％である。漢字比が与えられておればその時の使用漢字数は（1）式を変形した次の式から得られる。

使用漢字数＝漢字比／0.000367　　(3)

日本語では、ある一つの内容の文章や文書がいかに長くても常用漢字を全て使うことは

なく、内容に関係する範囲からの漢字を使う。この時に使用対象となる漢字数と漢字比の関係が（3）式である。

②（2）式の関係

かなり長い文章や文書でも内容が多種多様なものでなければ（3）式で得られる使用漢字数で足りるが、多種多様な文章を表現するためにはより多くの漢字を用意しておく必要がある。その時の漢字数すなわち基本漢字数と漢字比の関係が（2）式である。基本漢字数が多いほど多種多様な文章を表現することが可能となる。日本語にとって常用漢字数1945は多分野の多種多様な文章を表現するために必要な数である。（2）式から基本漢字数が常用漢字1945の時の平均的な漢字比は0.339となる。

以上のことから、日本語のかなり長い一つの文章や文書中で使用対象となる漢字数は、（2）式に基本漢字数として常用漢字数1945を代入したときの漢字比0.339を（3）式に当てはめて得られる924字となる。すなわちどんなに長い文章や文書でも内容が一つと考えられるものであれば使用対象となる漢字（種類）数は約900字である。教育漢字数996は900よりもやや多く、日常的な文章を表現するための最小限の基本漢字数だと言える。

1．2　基本漢字数と仮名比

仮名比は基本漢字数の増加とともにほぼ次式に従って直線的に減少する。

$$仮名比 = -0.000096 \times 基本漢字数 + 0.766 \qquad (4)$$

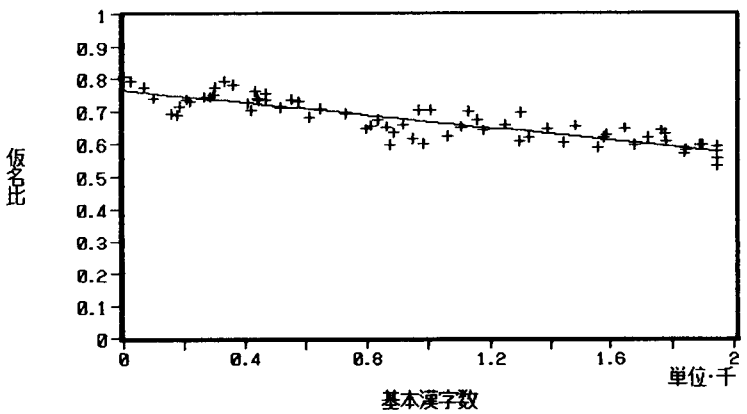

図2　基本漢字数と仮名比

1．3　基本漢字数と空白比

空白比は基本漢字数が250以下ではほぼ0.2で一定である。基本漢字数が300以上では、基本漢字数の増加にともない緩やかに減少している。

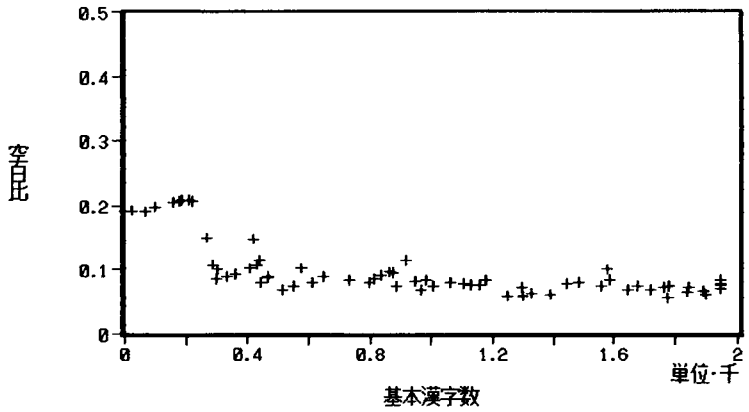

図3　基本漢字数と空白比

1．4　中国語への拡張

(2) (3) 式が中国語にも適用できるとする。中国語の漢字比を0.9として (2) 式から中国語の基本漢字数を推定すると約9000字となる。すなわち中国語では多様な情報を表すために必要な基本漢字数は9000字と推定できる。[中国語辞典] によると漢字の総数は40000余りであるが1932年に制定された「国語常用字彙」にはおよそ9000字を収録したとあり、(2) 式から推定した9000字と同じ値である。また (3) 式からは一つの内容のかなり長い文章や文書中で使用対象となる漢字数が約2500字と推定できる。この値は [中日辞典] の説明にある「現代漢語常用字表」の中国語の常用字数2500字と一致している。これらのことから、(2) (3) 式は中国語にも適用できるのではなかろうか。

2．表音文字の連続長さ

アルファベットや仮名など文字の配列によって意味を持つ表音文字の実際の平均情報量は文字の連続長さに依存するであろう。英語ではアルファベットの連続長さは空白と空白で区切られる単語の長さであり、日本語では空白や漢字で区切られるまでの仮名の連続長さである。

次節での平均情報量の検討のため、前節と同じ69の文章の基本漢字数と仮名の連続長さとの関係と、対訳文である [天声人語 '93春] の20の文章の漢字比と仮名の連続長さとの関係および英語の連続長さを調べた。

2．1　基本漢字数と仮名の連続長さ

仮名の連続長さは基本漢字数の増加につれて減少している。69の文章から、仮名の連続

長さと基本漢字数Xの関係を求めると次式となる。

$$仮名の連続長さ = 10^{-0.000086X+0.611} \qquad (5)$$

（X：基本漢字数）

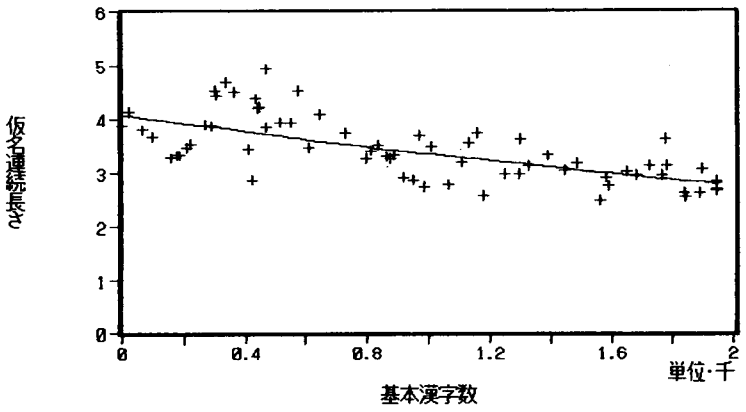

図4　基本漢字数と仮名連続長さ

2．2　漢字比と仮名の連続長さ

対訳文である［天声人語 '93春］の20のデータの漢字比と仮名の連続長さとの関係から、基本漢字数が1945の時の仮名の連続長さを漢字比Xから求める式が得られる。

$$仮名の連続長さ = 10^{-0.716X+0.656} \qquad (6)$$

（X：漢字比）

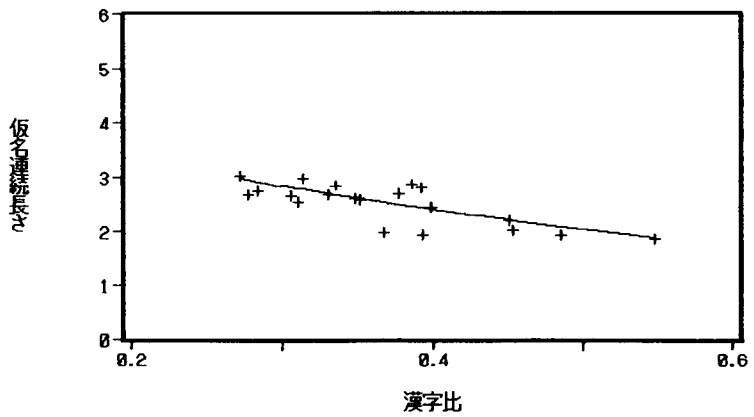

図5　漢字比と仮名連続長さ

2．3　英語の連続長さ

英語のアルファベットの連続長さは単語の長さに等しい。[天声人語 '93春] の20の文章のアルファベットの連続長さの変動は少なくその平均は4.66である。

3．文字の平均情報量

英語・仮名・漢字の平均情報量について検討する。

3．1　英語

アルファベット26文字が等確率で現れるとしたときの英語の平均情報量は（$\log_2 26=$）4.70ビットである。実際の英語では、[横原1992] より平均情報量は1ビット程度である。この値は4.70を英語のアルファベットの連続長さ4.65で割った値にほぼ等しい。本稿では実際の英語の平均情報量を1ビットとする。

3．2　仮名

基本漢字数と仮名の連続長さとの関係式（5）から、基本漢字数が0すなわち仮名だけの日本語の文章での仮名の連続長さは4.1となる。仮名だけの文章で仮名71文字が等確率で現れるときの平均情報量は（$\log_2 71=$）6.15ビットで、これをこの仮名の連続長さ4.1で割れば1.5 ビットを得る。[横原1992] からこの値は仮名だけで表された実際の日本語の平均情報量にほぼ等しい。

アルファベットや仮名のようにかなり長い実際の文章の文字数（数千から数万）に比べて基本文字数が少なく、文字のつながりによって情報を表す言語では、実際の平均情報量は基本文字が等確率で出現するとした時の平均情報量を文字の連続長さで割ればよいと言える。

日本語では仮名と漢字が混在している。日本語中での仮名の数は全文字数の7割程度あり使用文字数は基本文字数71よりはるかに多いので、実際の日本語すなわち漢字と仮名を混用した時の仮名の平均情報量も6.15を仮名の連続長さで割った値としてよい。

仮名の平均情報量＝6.15／仮名の連続長さ　　(7)

基本漢字数をXとすると仮名の平均情報量は（5）（7）式より次式で与えられる。

仮名の平均情報量＝$6.15/10^{-0.000086X+0.611}$　　(8)

（X：基本漢字数）

漢字比をXとすると、仮名の平均情報量は（6）（7）式より次式で与えられる。

仮名の平均情報量＝$6.15/10^{-0.716X+0.656}$　　(9)

（X：漢字比）

3．3　漢字の平均情報量

漢字は表意文字でありアルファベットや仮名と異なり一字一字の文字の独立性が高い。漢字の平均情報量は仮名と英語の平均情報量が分かれば対訳文の総情報量が等しい（[横原1993] 参照）と言うことから求めことができる。[天声人語 '93春] の20の文章の英語と日本語の総情報量が等しいと言うことから、この20の文章について次の式が成り立つ。

英語の総情報量＝仮名の総情報量＋漢字の総情報量

（空白は英語及び仮名に含まれるとする）

英語のアルファベットと空白の平均情報量は1.0ビット、仮名の平均情報量を（7）式により求め、空白の平均情報量も仮名と同じとすると、20の文章の漢字の平均情報量は次式から求めることができる。

漢字の平均情報量＝(1.0×英語・空白字数－
仮名の平均情報量×仮名・空白字数)÷漢字数

この漢字の平均情報量と漢字比の関係を示すと、図6となる。

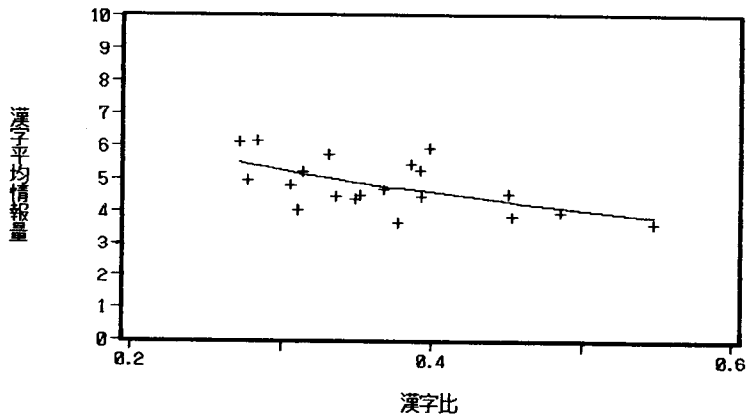

図6　漢字比と平均情報量

この図から漢字の平均情報量は漢字比が増加すると減少する傾向にあることが分かる。漢字比の増加とともに漢字相互間の関連が増すからである。漢字の平均情報量は文字の出現確率の2を底とする対数に比例する考えられるので、これを考慮し漢字比と平均情報量の関係式を求めると次式となる。

〔常用漢字数での漢字比と漢字の平均情報量の関係式〕

$$\text{漢字の平均情報量} = 1.67 \times \log_2(1945 / \text{漢字比}) - 15.9 \qquad (10)$$

基本漢字数が1945でないときにも次の式が成り立つと考えられ、

$$\text{漢字の平均情報量} = 1.67 \times \log_2(\text{基本漢字数} / \text{漢字比}) - 15.9$$

(1) (2) 式をこの式に代入すると次式が得られる。

〔基本漢字数と漢字の平均情報量の関係式〕

$$H=1.67\times\log_2(1/0.000367)-15.9=3.16\text{ビット} \qquad 0\leqq X<650$$

$$H=1.67\times\log_2(X/(0.0000799X+0.184))-15.9 \qquad 650\leqq X \qquad (11)$$

（H：平均情報量，X：基本漢字数）

〔漢字比と漢字の平均情報量の関係式〕

$$H=1.67\times\log_2(1/0.000367)-15.9=3.16\text{ビット} \qquad 0\leqq X<0.24$$

$$H=1.67\times\log_2((X-0.184)/(0.0000799X))-15.9 \qquad 0.24\leqq X \qquad (12)$$

（H：平均情報量，X：漢字比）

4．日本語の平均情報量

前節で求めた仮名と漢字の平均情報量から日本語の平均情報量を求める。

4．1　基本漢字数と平均情報量

(8) 式の基本漢字数と仮名の平均情報量の関係と、(11) 式の基本漢字数と漢字の平均情報量の関係を図7aに示す。漢字の平均情報量は基本漢字数が650までは3.16ビットと一定でその後急に増加し、常用漢字数の1945で4.95ビットとなる。しかし基本漢字数2000以上では増加率は鈍ってくる。仮名の平均情報量は基本漢字数が0の時すなわち仮名だけの日本語の時の1.51ビットから徐々に増加し基本漢字数が1945の時に2.22ビットとなる。仮名の平均情報量が基本漢字数とともに増える理由は日本語では漢字と仮名を混用しており、基本漢字数が増加し漢字比が増えると仮名の出現確率が減少するからである。

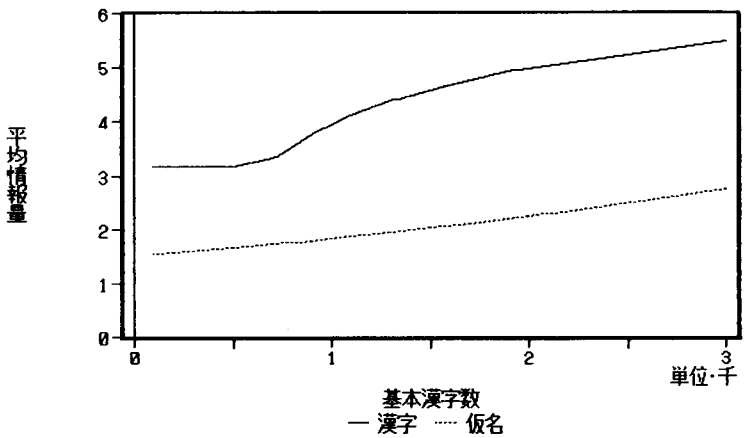

図7a　基本漢字数と平均情報量

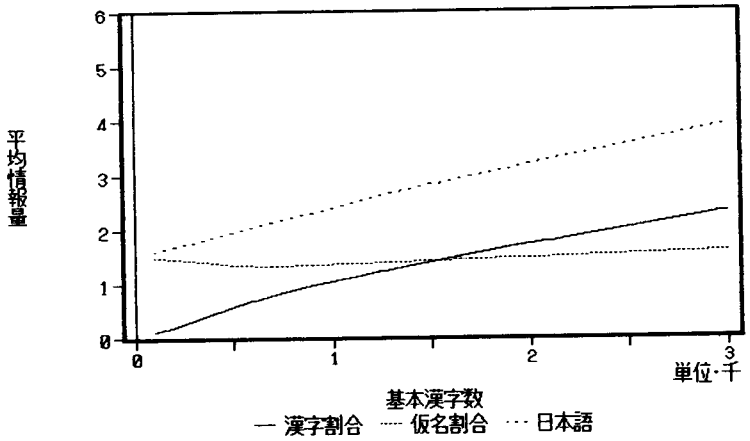

図7b　基本漢字数と平均情報量

日本語の平均情報量は次式から得られる。

日本語のH＝漢字のH×漢字比＋仮名のH×(1－漢字比)

(H：平均情報量，空白は仮名に含めた)

漢字と仮名の平均情報量が分かると、(1) (2) 式で漢字比は基本漢字数の関数なので基本漢字数と日本語の平均情報量の関係が求まる。日本語の平均情報量とその中での漢字の平均情報量の割合と仮名の平均情報量の割合を示すと図7bとなる。日本語の平均情報量は仮名だけの時の1.51ビットから基本漢字数の増加と共に増加しており、常用漢字1945個が全て使用対象となる時は3.15ビットとなる。

基本漢字数が1600個付近で漢字の平均情報量の割合が仮名の平均情報量の割合を上回る。漢字と仮名を混用する日本語では、基本漢字数が1600個以上で漢字を使う価値がより高くなることが分かる。常用漢字数は1945個であり、基本漢字数としてのこの条件を満たしている。基本漢字数が1945個すなわち常用漢字が全て使える条件での日本語の平均情報量は3.15ビットで、その中の仮名の平均情報量の割合は1.47ビット漢字の平均情報量の割合は1.68ビットである。日本語では漢字の平均情報量と仮名の平均情報量の割合は均衡がとれていると言える。またこの時の漢字の平均情報量は4.95ビット、仮名の平均情報量は2.22ビットである。

基本漢字数を増せば日本語の平均情報量は増加するが、漢字を覚えるための負担が増える。たとえば、基本漢字数が7000字の時には漢字比が0.74になり、日本語の平均情報量は6.12ビットにもなる。

常用漢字数が1945である理由をまとめると次のようになる。

基本漢字数が1945の近くでは

①仮名と漢字の平均情報量の割合のバランスがよい。

②漢字の平均情報量の増加率が鈍り始める。

③漢字を覚えるための負担がそれほど大きくない。

4．2　漢字比と平均情報量〔基本漢字数を固定しない時〕

(12) 式は、基本漢字数を1945個に固定しない時の漢字の平均情報量を漢字比の関数で表したものである。(9) 式の仮名の平均情報量と共に示すと図8aとなる。漢字の平均情報量は漢字比が0.24未満では3.16ビットと一定であり、0.24～0.4では増加率が大きく、0.4以上では増加率が鈍っている。

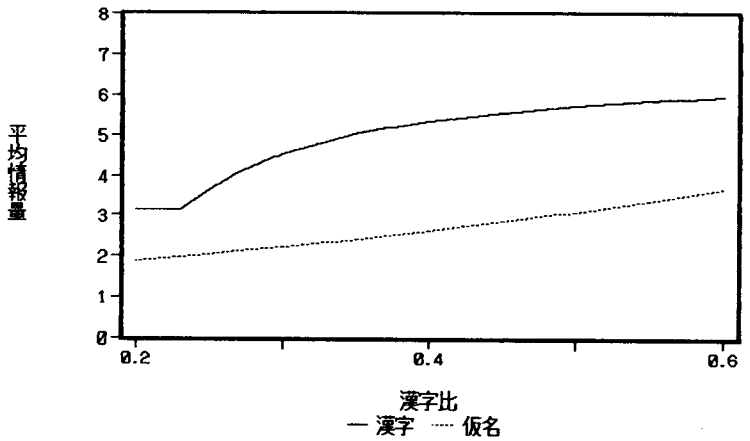

図8a　漢字比と平均情報量

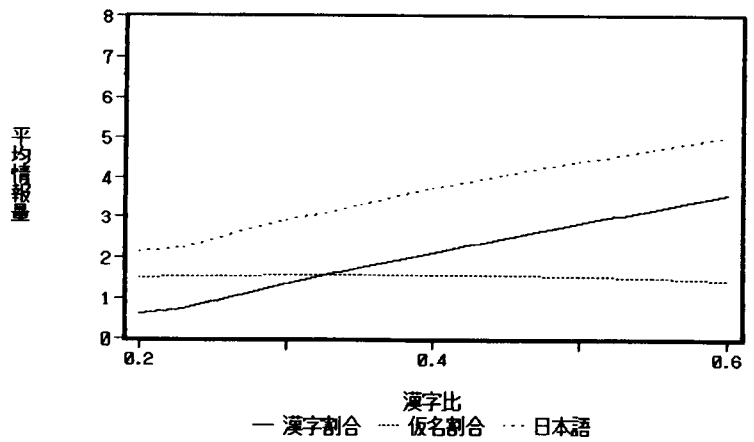

図8b　漢字比と平均情報量

日本語の平均情報量とその中での漢字と仮名の平均情報量の割合を図8bに示す。日本

語の平均情報量は漢字比の増加とともに増加するが、基本漢字数も増加する。漢字の平均情報量の割合は漢字比が0.33で仮名の平均情報量の割合と等しくなる。基本漢字数を1945に固定しないときにはこの0.33の付近で仮名と漢字のバランスが良いと言える。

4．3　漢字比と平均情報量〔常用漢字数の時〕

常用漢字が全て使用可能な通常の日本語において、漢字比を独立変数とした時の漢字の平均情報量は（10）式から、仮名の平均情報量は（9）式から得ることができる。この漢字比と漢字・仮名の平均情報量の関係を図示すると図9aとなる。漢字の平均情報量は漢字比とともに減少し、仮名の平均情報量は漢字比の増加と共に増加する。漢字比が0.6に近いところで仮名の平均情報量が漢字の平均情報量を上回る。

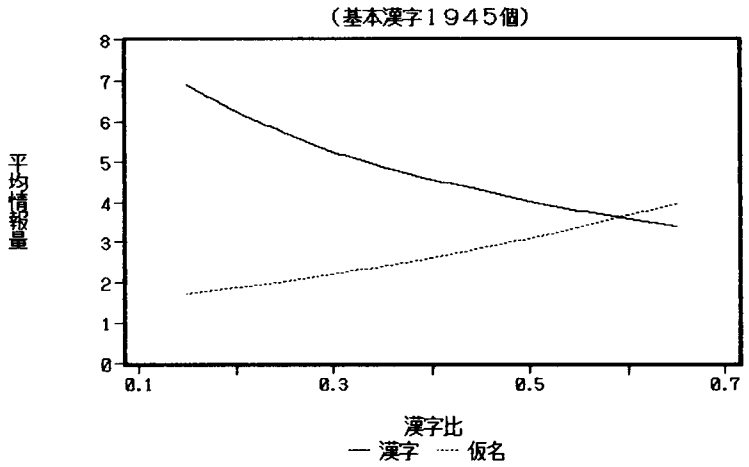

図9a　漢字比と平均情報量

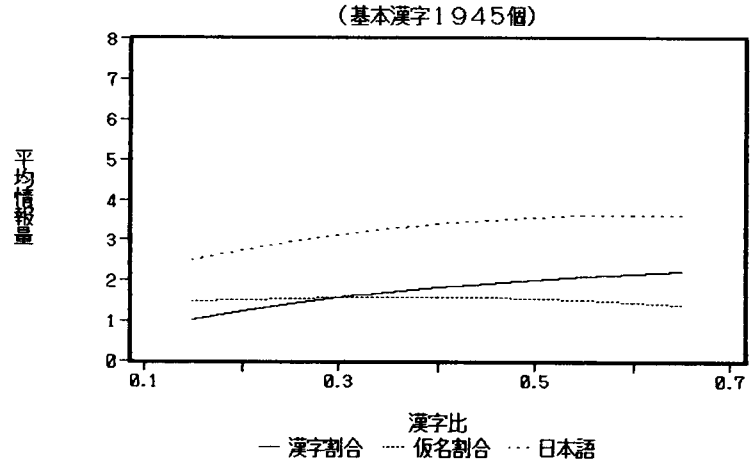

図9b　漢字比と平均情報量

漢字比と日本語の平均情報量、その中での漢字と仮名の平均情報量の割合の関係を示すと図9bとなる。漢字の平均情報量の割合が仮名の平均情報量の割合を上回るのは、漢字比が0.3以上の時である。これらのことから日本語の漢字比の適値は0.3近辺で上限は0.6程度と考えられる。[天声人語 '93春] の20の文章の漢字比の実測値は殆ど0.25～0.55の範囲にあり、この結果と一致する。

一方、(2) 式から基本漢字数が1945のときの漢字比は0.339と計算できる。4.2節での結果も合わせて、0.33付近に日本語の仮名と漢字の均衡点（漢字比の適値）があると考えられる。

本節での日本語の平均情報量の検討から、日本語の常用漢字数が1945個であり、漢字の空白を含めた全文字数に対する比率が0.33近辺にあることが理解できる。

5. 中国語・英語との比較

〔中国語の漢字数との比較〕

日本語ではある内容のかなり長い一つの文章や文書中で使用対象となる漢字種類数は約900字、多様な文章の表現のために用意されている漢字種類数は1945字である。1.4節の結果から中国語ではある内容のかなり長い一つの文章で使用対象となる漢字種類数は2500字、多様な表現のために用意すべき漢字種類数は9000字である。

〔中国語の平均情報量との比較〕

(11) 式を中国語にも適用し、漢字比0.9空白比0.1とする中国語のモデルについて、基本漢字数と平均情報量の関係をを図示すると図10となる。

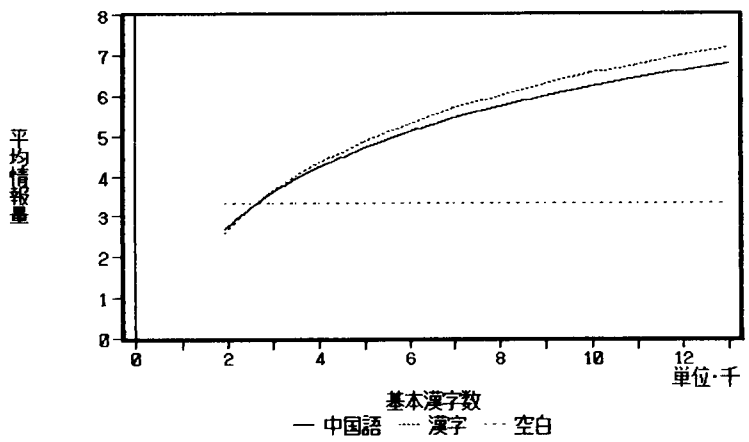

図10 中国語の平均情報量

中国語では基本漢字数が3500の時の平均情報量は3.95ビット、9000の時の平均情報量はは5.99ビットとなる。日本語は漢字比が0.339の時の平均情報量は3.15ビットであった。

漢字の平均情報量だけをみると、図9aより日本語の漢字の平均情報量は漢字比が0.35の時4.89ビットトである。これは中国語では基本漢字数が5000の時の漢字の平均情報量に相当する。

〔各文字の比較〕

本稿で得た結果を含めた英語・日本語・中国語の文字比・基本文字数・使用漢字数・平均情報量の値を表1に示す。

表1　各文字の比較

言語	文字比			基本文字数		一文章中使用対象漢字数	平均情報量(ビット)
	表音	表意	空白	表音	表意		
英　語	0.82	——	0.18	26	——	——	1
日本語	0.4～0.7	0.25～0.55	0.07～0.09	71	1945	900	2.9～3.6
中国語	——	0.9	0.1	——	3500～9000	2500	3.9～6.0

＊中国語の文字比の値は概略値である

＊中国語の常用字2500字、[中日辞典]の説明にある「現代漢語常用字表」の次常用字1000字、「国語常用字彙」の9000字から、中国語の基本文字数を3500～9000字とした。

＊日本語の平均情報量に幅があるのは、使用する漢字（表意文字）比に幅があるからである。中国語の平均情報量に幅があるのは基本文字数の幅による。

まとめ

1．日本語ではある内容のかなり長い文章や文書中で使用対象となる漢字数は、その文章や文書中での漢字比が分かっていれば（3）式で与えられる。すなわち一つの文章では常用漢字1945字の中の約900字が使用対象となる。ある内容の一つの文章ではテーマや分野が絞られ常用漢字全てを使うことはないからである。日常生活に関する文章だけであれば996字の教育漢字だけで足りるが、多様な文章を表現するには常用漢字数の漢字が必要となる。

2．一つながりの文字で情報を表す表音文字では、文字の連続長さ（英語なら単語の文字数、仮名なら空白や漢字で区切られる仮名の連続長さ）で基本漢字数が等確率の時の平均情報量を割れば実際の文字の平均情報量が求められる。

英語の平均情報量＝4.70／英字の連続長さ≒1.0ビット

仮名の平均情報量＝6.15／仮名の連続長さ

英字の連続長さはほぼ一定であり、したがって英語の平均情報量は1.0ビットで一定であるが、日本語の仮名の平均情報量は、漢字比が増えると仮名の連続長さが減少するので漢字比の増加とともに増加する。

3．漢字と仮名と日本語の平均情報量は漢字比の関数として求めることができる。

漢字のH＝$1.67\times\log_2(1945/X)-15.9$

仮名のH＝$6.15/10^{-0.716X+0.656}$

日本語のH＝漢字のH×X＋仮名のH×(1－X)

（H：平均情報量，X：漢字比）

4．日本語の常用漢字数が1945字である理由をまとめると次のようになる。

基本漢字数が1945の近くでは

①仮名と漢字の平均情報量の割合のバランスがよい。

②漢字の平均情報量の増加率が鈍り始める。

③漢字を覚えるための負担がそれほど大きくない。

5．日本語中の漢字の混在比率

日本語の漢字比は実測値と平均情報量の検討から0.25～0.55の範囲にある。漢字比の最適値は平均情報量の検討から0.33近辺にあると言える。

6．仮名と漢字は各々が単独で使われる時すなわち仮名だけの日本語や漢字だけの時よりも、仮名と漢字が混在している通常の日本語の中での方が各々の平均情報量が高くなる。仮名にとって漢字が空白と同じ役割をはたすからである（[横原1992] 参照）。また漢字にとっても仮名が空白の役割を果たしている。日本語では仮名と漢字の混用により2000弱の漢字数で漢字学習の負担を軽減し、かつ全体の平均情報量を上げている。

本稿では、文字すなわち書き言葉としての日本語を漢字の混在比率と平均情報量の観点から考察した。言語による情報伝達には書き言葉と話言葉の２つの側面があり、話言葉としての考察は今後の課題である。

参考文献

[小学校・中学校の教科書]

①木下順二他：改訂しょうがくこくご１上、１下，教育出版（1990）

②木下順二他：改訂小学国語２上、２下，教育出版（1991）

③栗原一登他：国語三上、三下，光村図書（1992）
④木下順二他：改訂小学国語４上、４下，教育出版（1990）
⑤木下順二他：改訂小学国語５上、５下，教育出版（1991）
⑥栗原一登他：国語六上、六下，光村図書（1992）
⑦石森延男他：国語１，光村図書（1990）
⑧石森延男他：国語２，光村図書（1991）
⑨石森延男他：国語３，光村図書（1992）
［天声人語 '93春］朝日新聞論説委員室〔編〕＋（株）英文朝日〔訳〕：〔英文対照〕天声人語 '93春，原書房（1993）
［中国語辞典］倉石武四郎：岩波中国語辞典，p13，岩波書店（1990）
［中日辞典］北京・商務印書館，小学館：中日辞典，pⅦ，小学館（1992）
［横原1992］横原恭士：日本語における漢字の役割について，相愛大学研究論集第８号（1992）
［横原1993］横原恭士：日本語と英語の画数による比較研究，相愛大学研究論集第９号（1993）