日本語における漢字の役割について

——英語・中国語との情報量による比較から——

A Role of Kanji in Japanese Language

——A Comparison of Linguistic Entropy among Chinese, English and Japanese——

横　原　恭　士

はじめに

人間は五感によって体外からの情報の収集を行っている。視・聴・嗅・味・触の五感の中で、視覚による情報量が1秒当り300万ビットで最も大きい。耳から入る情報量は1秒当り数万ビットで視覚による情報量の百分の一程度である。人間同士はお互いに情報の受渡しを言語を用いて視覚と聴覚により行っているのである。人間が最大の情報収集能力を発揮しうる視覚による情報伝達においては情報を文字の形で表し、耳からの情報伝達では情報を言葉で表す。本稿では日本語を英語や中国語と比較することによって、情報が日本語の文字で表された場合の情報伝達についての定量的な評価を行う。はじめに日本語と英語という二つの言語の情報伝達上の比較を、対訳文の文字数の差という量的また視覚的に把握できる量によって行う。次に平均情報量および一定の情報を表す時の文字数の差からの検討を、英語・中国語・日本語を比較することによって行う。

日本語と英語の文章を比較した場合、訳者の文章表現力が原文に劣らない場合の対訳文のように内容が同じと見なされる場合でも、以下で検討するように使用する文字数にかなりの差がでてくる。

日本語と英語にはいろいろな違いがある。それは同じ内容の文章を表現する時の文字数であり、アルファベットと仮名・漢字など基本文字数の差である。1文字当りの平均情報量にも差があり、表音文字のみの言語と表意文字と表音文字の混在した言語という違いもある。中国語は漢字と空白だけで情報を表す言語であり、漢字は日本語の起源でもあるので比較対象に加えた。

言語間の違いに着目し、またコンピュータの発達を念頭において三つの言語の比較検討を行った。異なる言語の比較を行ってみることは、情報化社会の中での言語のあり方、役

割、有効性を見通すうえで重要なことであると考えるからである。

この英語と日本語の比較においては、表意文字である漢字を起源とする日本語と表音文字だけで構成される英語の差を、文字数の点から検討するだけではなく、言語としての基本的な考え方や情報化社会といわれる時代背景からの検討も必要である。

情報化時代が始まる前から、人間は文字や言葉で情報を表現してきた。文字や言葉による情報伝達は記号を一次元または時系列的に並べて情報を表現しようとするものである。この一次元的情報表現の簡潔な形として表音文字が発達してきたのである。その背景には、情報を手で書き表すための時間をできるだけ早めたいという動機があったのであろう。

情報処理機械であるコンピュータの発達した情報化社会では、日本語の場合も手書きによる文字表現や言葉による情報表現に要する時間と同じ程度かより短い時間で、コンピュータによる情報表現が可能である。情報化以前には存在した言語間の手書きのための時間の差はコンピュータによってなくなったのである。現在では、情報を表現するための時間の点から、どの言語にも優劣はつけがたい言語平等の時代なのである。このようにコンピュータやワープロにより情報を入力するという点では言語間に優劣はなくなったが、情報を表示するという点ではどうであろうか。視野は限られているが、1秒当り300万ビットの情報を入力できる眼を持っている人間にとって、一定面積の視野の中にたくさんの情報が入っている方が情報伝達上有利である。情報処理機械を用いる情報入力に言語間の優劣がなくなってくると、一定視野内への情報表示の有効性が将来の言語の方向を左右するのではないだろうか。こういう点から時系列的な文字配列の中にも、二次元（面）的な情報表現力をもつ漢字を含む日本語の有効性が活かされるのではないか。これからは、一次元すなわち表音文字に加えるに二次元すなわち表意文字や画像による情報伝達の時代なのではないだろうか。コンピュータによる情報化は視覚の情報把握能力を有効に使う、すなわち視野に入る情報量をできるだけ多くしようとする二次元情報伝達かへの流れでもある。こういう流れを見据えつつ、文字としての面的情報表現力すなわち二次元性をもつ漢字を含む日本語と、文字としても言葉としても簡潔で一次元性の言語である英語を定量的に比較し日本語の中の漢字の役割を評価してみた。さらに日本語の起源である中国語についても、そのモデルを日本語との比較対象とした。

1．対訳文による英語と日本語の比較

同じ内容の文章について、日本語と英語で表現した場合の違いについて字数をもとに検討した。対象として、日本語で400字程度の文である対訳文「Bertrand Russell's Best」[(1)]から最初の5篇の文章を選んだ。この対訳文を対象としたのは、訳文である日本語の文章が冗長でもなく省略しすぎもなく文章表現力が原文と同じ程度の能力だと思われるからで

ある。各々の文章は原文の英文字数が500～800字、訳文の日本語で300～400字の短文である。比較は内容や意味には立ち入らず定量的に行った。英語については、アルファベットの文字数と空白（スペース）の数を数えた。空白は単語と単語の間および文と文の間に1個とし、ハイフォンや括弧などの記号は除外した。日本語については、漢字、仮名、空白の数を数えた。片仮名は仮名に含め、漢字については仮名読みの字数もかぞえ併記した。英文字の合計はアルファベットの文字数に空白を加えたもの、日本語の合計は漢字、仮名、空白を加えたものである。また、漢字を仮名読みした場合の合計も括弧内に記載した。

1．1　文字数の比較

対訳文の字数を表1に示す。

表1　対訳文の字数

標　題	英字数（原文）			日本語字数（訳文）			
	英字[a]	空白[b]	計[A]	漢字[c]（よみかな[d]）	仮名[e]	空白[f]	計[B]（計[C]）
① APPEARANCE AND REALITY	802	197	999	113（204）	324	27	464（555）
② THE EXISTENCE OF MATTER	590	122	712	102（197）	203	16	321（416）
③ THE EXISTENCE OF MATTER (continued)	492	114	606	91（176）	231	18	340（425）
④ ON INDUCTION	588	132	720	86（138）	248	20	354（406）
⑤ REMEMBERING	654	133	787	125（232）	250	18	393（500）
合　計	3,126	698	3,824	517（947）	1,256	99	1,872(2,302)

＊（　　）は漢字の仮名読みの値

表2　字数の比

標　題	$\frac{A}{B}$	$\frac{A}{C}$	$E=C+c$	$\frac{A}{E}$	$F=C+\frac{c}{2}$	$\frac{A}{F}$
①	2.15	1.80	668	1.50	611.5	1.63
②	2.22	1.71	518	1.37	467	1.52
③	1.78	1.43	516	1.17	470.5	1.29
④	2.03	1.77	492	1.46	449	1.60
⑤	2.00	1.57	625	1.26	562.5	1.40
平均(合計)	2.04	1.66	(2,819)	1.36	(2,560.5)	1.49

表2は表1をもとに、英文字数の合計と日本語字数の合計、漢字を仮名読みした時の合計、およびそれに漢字字数と同数あるいは半数の空白の数を加えた合計との比を計算したものである。仮名のみの日本語の場合に空白を加えたのは、単語の間に空白を加えることによって仮名だけの日本語はより分かりやすくなるということを考慮したからである。

（1）空白を含めた全字数

英文字数が日本語字数の1.78倍から2.22倍、平均2.04倍の字数である。

（2）漢字の仮名読みの字数と漢字の字数

平均1.83倍である。

（3）英文字数と仮名日本語の字数

英文字数が平均1.66倍の字数である。

（4）英文字数と空白補正を加えた仮名日本語の字数

日本語を全て仮名読みした場合の文字数に、漢字1個につき1ないし2分の1個の空白を加えた場合も計算した。空白1個の場合には英文字数と日本語文字数の比は、1.36で2分の1個の場合は1.49であった。

1．2　平均情報量

同じ内容の文章を表す日本語と英語の文字数の差に関する検討を平均情報量の観点からも行った。英語はAからZと空白の27文字で、その基本文字数は常に一定である。英語で

は26文字の組合せによって単語を作り、さらにその単語の組合せによって文を作り情報を表現しているのである。一方日本語では、仮名の清・濁音と空白の72文字と数千個の漢字および片仮名を使用する。日本語では、漢字によってだけで情報を表すのではなく、より容易に情報を表現するため仮名や片仮名を作りだしたのではないか。英語が26個の少ない基本文字数を持ちその組合せで情報を表現する基本文字数最小型であるのに対して、日本語は多様な情報を少ない文字数で表現できる漢字と、使い易さと柔軟性を可能とする仮名との混合型なのである。ここで検討の対象とする日本語は当用漢字と仮名および空白から構成されているとした。片仮名は仮名とみなした。

表3は、既に計算されてある英語と日本語の平均情報量の値を表にしたものである。

表3　日本語と英語の平均情報量[(2)(3)]

	全ての文字が同一確率	各文字の独立出現確率	単純マルコフ情報源	2重マルコフ情報源	実際（推定）
英語　27文字（A～Z，空白）	4.76ビット (1.0)	4.03	3.32	3.1	…… 0.6～1.3
日本語　72文字（平仮名，空白）	6.17ビット (1.3)	5.5 (1.4)	4.9 (1.5)	4.5 (1.5)	…… 1.5～2.0

＊（　）内は日本語の英語に対する比

対訳文の比較において英語の字数が日本語の字数の2.04倍であった。また、英語の字数と日本語をすべて仮名した場合の字数とを比較してみると、表2に示すように英文が1.66倍の字数である。しかし、日本語を全部仮名で表した時には空白を加えたほうが意味がよく分かるので、空白を加えた場合の計算も行うと、漢字1個に対し空白1個を加える補正を行った場合には、英文字数が日本文の1.36倍となり、空白を0.5個加える補正を行った場合には、1.49倍となる。この値は表3の仮名日本語と英語の平均情報量の比の値に近い値である。実際の日本語と英語では、平均情報量の比は1.5程度と推定できるので、空白0.5個を加えた場合の値とよく一致する。すなわち仮名だけの日本語では、漢字を仮名に代えるだけでなく漢字数に対応した空白を加えると、全情報量である「文字数×平均情報量」が英語のそれとほぼ等しくなることが分かる。

1．3　字数と平均情報量からの考察

対訳文において英語および仮名日本語について全文字同一確率のときの全情報量を計算してみると、英語については次のようになる。

英語の全情報量＝3824×4.76＝18202ビット

仮名日本語については、表3から実際の仮名日本語と英語の平均情報量の比は1.5程度と推定できるので、これを考慮すると次の値が得られる。

仮名日本語の全情報量＝2302×4.76×1.5＝16436ビット

また漢字混じりの日本語全情報量を漢字と仮名（空白を含む）の平均情報量を別々として計算すると次のようになる。ただし、仮名の平均情報量は実際の場合の英語との比1.5を考慮した。また、漢字の基本字数は1024字と1850字の2ケースとした。

日本語の全情報量（漢字1024字）

$$=1355\times\log_2\left(\frac{1872}{1355}\times\text{基本仮名字数}\right)+517\times\log_2\left(\frac{1872}{517}\times\text{基本漢字数}\right)=16436\text{ビット}$$

日本語の全情報量（漢字1850字）＝16877ビット

仮名日本語の全情報量と漢字混じりの日本語の全情報量はほぼ同じである。日本語を全て仮名で表し空白を追加しなければ、その時の情報量は漢字混じりの日本語の情報量と差がないことになる。

一方、英語と日本語の全情報量には差がある。両言語の文字による情報の表現が適切であれば、全情報量の差は英語と日本語の情報伝達効率の差、言い替えると符号としての言語の効率の差によるということになる。ここでの計算では、日本語の方が英語より約10％効率がよいことになる。これは何によるのであろうか。仮名日本語が通常の日本語より空白が多くなることから、このことが空白の数の差によると考える。英語の情報量を基準に考えると、英語の全情報量と仮名日本語の全情報量の差1766ビットは仮名・空白の247字分に相当する。これは仮名日本語の字数の10.7％に相当する。これを全て空白で補うとすると、元の空白を含めて空白の比率は14％に相当する。日本語では漢字がこの効果を生んでいると考えられ、漢字1字当り0.48個の空白を増やしたことと同じとなる。このことは、仮名日本語では漢字を仮名に代えるだけではなく単語の間に空白を入れないと読みにくいことからも理解できる。また、1．1節の表2の結果ともよく合致する。

通常の日本語では、漢字が空白の効果を含み情報伝達上の言語の符号としての効率化が行われているため、仮名日本語に比べ基本字数の差すなわち平均情報量の差を考慮してもかなり少ない字数で情報を表現できるのである。漢字は基本字数により平均情報量を上げるとともに符号としての言語の効率を上げる効果も持っていることになる。英語などの表音文字と情報量を比較する場合に、漢字が持っているこの符号の効率を上げる効果を「漢字効果」と呼ぶことにする。

英語や仮名などの表音文字では、単語などの区別のため空白を多く必要とするが、漢字を含む日本語では漢字と仮名の間には空白は不要なのである。

但し、このことは書言葉すなわち文字について言えるのであって、話言葉の場合には別の観点からの検討が必要である。

2．英語・中国語と日本語

前節で得た結果をもとに、英語、日本語と中国語のモデルについて、字数および平均情報量からの比較を行った。対象とした言語の基本文字は次のものである。

①英　語　　A～Z・空白の27文字
②日本語0　　仮名・空白の72文字
③日本語Ⅰ　　仮名・空白の72文字と漢字1024文字
④日本語Ⅱ　　仮名・空白の72文字と漢字2048文字
⑤日本語Ⅲ　　仮名・空白の72文字と漢字8192文字
⑥中国語Ⅰ　　空白と漢字8192文字
⑦中国語Ⅱ　　空白と漢字32768文字

平均情報量は英語を1.0ビットとし、日本語0は表3の値から1.5ビットとした。また、日本語では空白は仮名の一部とみなしたが、中国語では空白と漢字は別とした。日本語ⅠⅡⅢと中国語ⅠⅡの平均情報量は次の式によって計算した。

（平均情報量）

①英　語　　1.0ビット
②日本語0　　1.5ビット
③日本語ⅠⅡⅢ　　$\text{仮名比}\times 1.5+\text{漢字比}\times\log_2\left(\text{基本漢字数}\times\frac{1}{\text{漢字比}}\right)\times\frac{1}{4.76}$ビット
④中国語ⅠⅡ

$$\left\{\text{空白比}\times\log_2\left(\frac{1}{\text{空白比}}\right)+\text{漢字比}\times\log_2\left(\text{基本漢字数}\times\frac{1}{\text{漢字比}}\right)\right\}\times\frac{1}{4.76}\text{ビット}$$

この値には「漢字効果」が入っていない。この平均情報量に「漢字効果」を加えた。これらの平均情報量と「漢字効果」を加えた値を用いて、1000ビットの情報を各言語で表した時に要する文字数を計算した結果を表4に示す。図1は平均情報量と平均情報量に「漢字効果」を加えた値を図示したものである。

表4　モデル言語の平均情報量と「漢字効果」

言　語		平均情報量	平均情報量 ＋「漢字効果」	1000ビットの情報に 必要な字数
英　語		1.0ビット		1000
日本語0 （仮名・空白72字）		1.5		667
日本語Ⅰ （漢字1024字）	仮名50% 仮名75%	1.91 1.76	2.27 1.94	441 515
日本語Ⅱ （漢字2048字）	仮名50% 仮名75%	2.01 1.81	2.37 1.99	422 503
日本語Ⅲ （漢字8192字）	仮名50% 仮名75%	2.22 1.92	2.58 2.10	388 476
中国語Ⅰ （漢字8192字）	空白 5% 10% 25%	2.65 2.56 2.22	3.33 3.21 2.76	300 312 362
中国語 （漢字32768字）	空白 5% 10% 25%	3.08 2.93 2.54	3.76 3.58 3.08	266 279 325

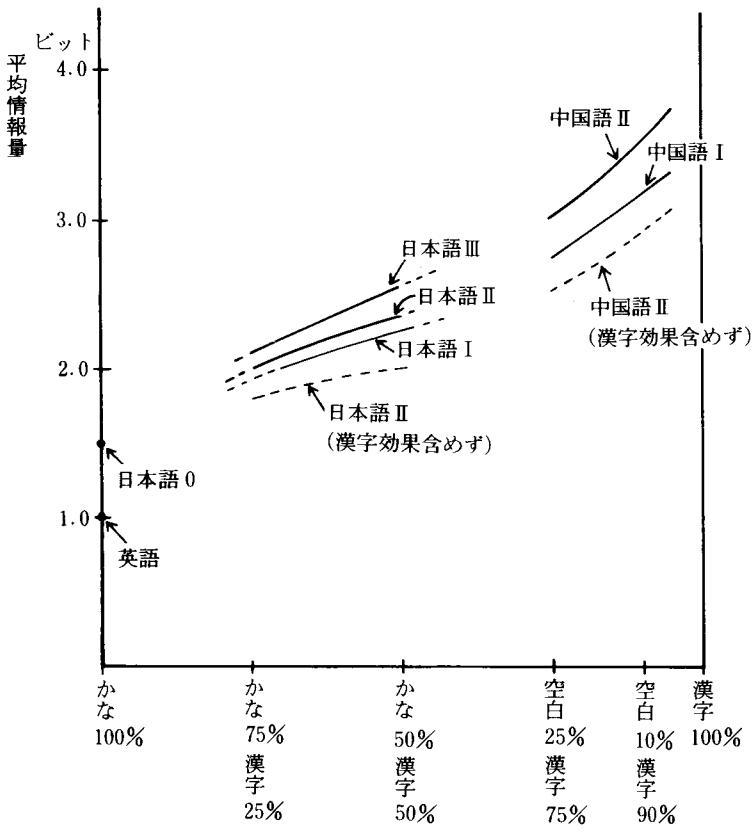

図1　モデル言語の「漢字効果」を含めた平均情報量の値

2．1　平均情報量

同じ内容の情報をいろいろな言語で表す場合、言語によって全情報量に差がないとすると、1.3節の結果から日本語中の漢字には1文字当り0.48個の仮名・空白に相当する符号としての言語の伝達効率を増加させる効果を余分に持っていることになる。

図1よりわかることは、日本語において平均情報量を増やそうとすると漢字の混在比率を上げればよいが、その時の寄与率は「漢字効果」の方が漢字の混在比率を上げる効果よ

りも大きいということである。日本語Ⅱでは漢字の混在比を25％から50％に上げると平均情報量は10％上がるが、漢字の混在比が50％の時の「漢字効果」は18％にものぼる。

中国語でも日本語と同じ程度の「漢字効果」があるとすると、空白の比率が10％の時「漢字効果」により22～25％も平均情報量を上げるのと同じことになる。

また、日本語では漢字の基本字数を8倍にする（日本語Ⅰから日本語Ⅲ）以上の効果を「漢字効果」は持っている。中国語でも基本字数を4倍にする（中国語Ⅰから中国語Ⅱ）のに倍する効果を「漢字効果」は持っていることになる。

2．2　字数

1000ビットの情報を表現するのに必要な各モデル言語の字数を図2に示す。

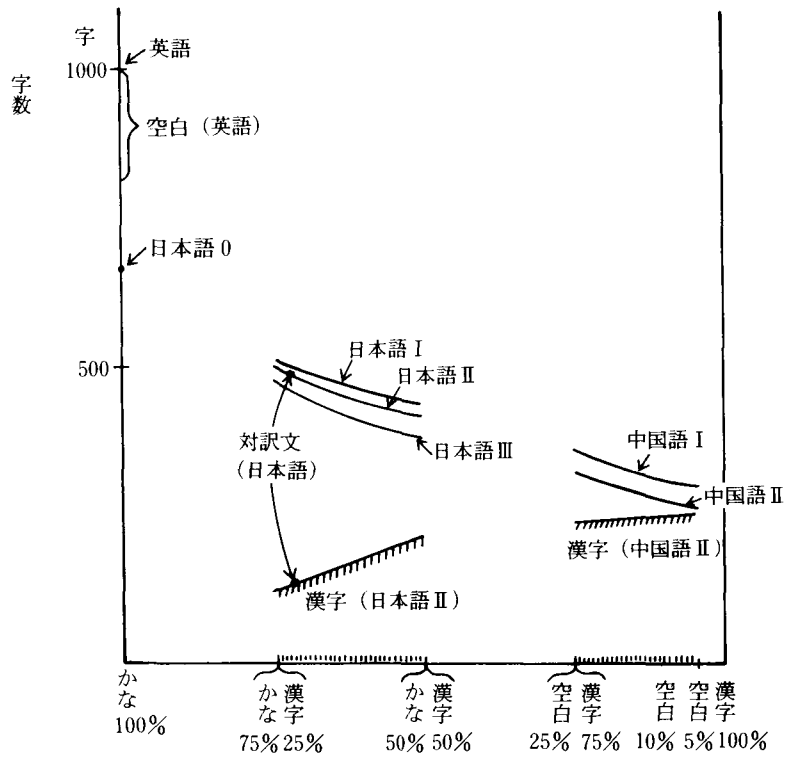

図2　1000ビットの情報に要する字数

（1）総字数

英語の総字数は日本語の総字数の約2倍である。中国語の総字数は日本語の6割程度である。

（2）空白

各言語により比率はほぼ一定と思われ、日本語では5％程度、英語では18％程度、中国語では10～15％程度であろう。中国語と日本語の空白比が英語に比較して少ないのは、単語と単語の間に空白を入れる必要がないからである。日本語では仮名を使っているために中国語より総字数が多くなり、空白比率は中国語より小さくなる。日本語では表音文字である仮名を使っているにもかかわらず空白の比率が低いのは、漢字が仮名の間に適当な比率で入って「漢字効果」を発揮しているからである。

（3）字数

中国語では基本字数を8192字から32768字へ4倍に増しても、10％しか字数は減らない。日本語では2048字から8192字へ4倍に増しても5％しか字数は減らないが、使用する漢字の比率を25％から50％に増やすことによって16％字数を減らすことができる。

英語は表音文字を一次元に並べているため字数を減らすことができない。中国語は基本漢字数を増やすことによって字数を減らすことができるが、基本漢字数の大幅な増加を招くことになる。一方日本語は、漢字の混在比率を増すことによって基本漢字数の大幅な増加なく文字数を効果的に減らすことができるのである。

（4）漢字

中国語では情報を漢字と空白で記述しているため平均情報量は大きいが、学習し覚えなければならない基本漢字数が非常に多い。使用する字数を少なくするためには、基本漢字数が膨大となり学習に大変な努力が必要である。

日本語では清・濁71字の仮名と空白と通常数千の漢字で情報を表現するので、漢字学習は中国語に比べ容易である。また仮名単語間の空白は漢字の存在により不要なので、空白比率は英語の3割程度で情報表示面積上効率がよい。

3．結論

日本語と英語の対訳文の比較と日本語・英語・中国語のモデル言語の比較を定量的に行うことによって、日本語における漢字の役割を中心に日本語について考察した。結論をまとめると次のようになる。

（1）漢字は表音文字だけからなる言語における空白の効果を含んでいる。その効果は漢字1個につき空白約0.5個分に相当する。（本文の「漢字効果」）

同じ内容の情報を文字で表した場合には、日本語は英語に比べて文字数も少なく情報の

言語への符号化の効率が良い。すなわち情報伝達効率が良い。（本稿の対訳文では約10％）これは日本語における漢字が英語における単語間の空白にあたる効果を含んでいるからである。この効果は漢字１個につき約0.5個の空白に相当する。この効果は平仮名だけからなる日本語と比較する場合にも言えることである。

（２）２つの表音文字間では、同じ内容の情報について“文字数と平均情報量の積”はほぼ等しくなる。

基本文字数の差によって、文字で書き表した時の言語間には平均情報量と使用文字数の差が生じる。英語と仮名日本語のように表音文字間では、平均情報量と文字数の積はほぼ等しくなる。しかし、漢字を含んだ日本語ではこの積は小さな値となる。すなわち漢字が入ることによって情報の言語への符号化の効率が良くなるのである。

（３）日本語は仮名によって基本漢字数の増加を抑えている。

日本語は仮名と漢字が混在している。仮名は情報伝達の効率を大きく落すことなく、漢字の基本文字数の大幅な増加を抑制する役割を果たしている。また漢字は、表音文字である仮名の単語間に必要な空白の役割を内に含んでいるため空白が少なくてすむのである。日本語では仮名と漢字がお互いの情報伝達上の短所を補いあっているのである。

4．まとめ

日本語の漢字の情報伝達における役割を中心にして日本語について考察した。

情報処理機械を使う場合には、情報を言語として符号化する効率すなわち言語による情報表現の効率が、情報入力時間の短縮や情報表示面積の低減につながることになる。情報化社会が進展すると、二次元で情報を表現しているため情報表現の効率の高い漢字の価値が高まるのではなかろうか。

漢字を使うと情報伝達の効率が上がる反面、覚えなければならない基本字数が多くなるという欠点がある。しかし、日本語では仮名と漢字を併用することによって、漢字の基本字数を適正な量に抑え、かつ情報伝達の効率も維持できることが分かった。日本語は表音文字と表意文字の双方の長所を活用している言語と言えるのである。

参考文献

（１）神山正治　1961　*Bertrand Russell's Best,* 金星堂
（２）佐藤憲市　1982　『情報科学』八千代出版
（３）大村　平　1970　『情報のはなし』日科技連