# 日本語の次文字予測率からの検討

The Predictability of Sequential Characters in the Japanese Language

横原恭士

## はじめに

情報伝達、特に言語による情報伝達における2つの重要な点は、情報の伝達効率と伝達された情報の正確さである。その重要性は、情報伝達の手段が言葉でも文字でも変わらない。

ある言語による情報伝達の効率と正確さを検討する時、その言語のエントロピーの値が示されれば、効率と冗長度が計算でき情報伝達の効率と正確さを定量的に検討できる。

エントロピーの値は、情報伝達の効率の面と正確さの確保とのバランスによって決まっており、記号のマルコフ連鎖性を考慮した出現確率から導き出すことができる。実際のエントロピーの値とエントロピーの最大値との比によって情報伝達の効率がわかり、1と効率の値の差から情報伝達の正確さの目安となる冗長度が求まる。

しかしまた、情報伝達の正確さは次に出現する記号をいかに推定できるかと言うことによっても確保できる。出現する記号の推定がマルコフ連鎖性を考慮した出現確率だけで行なわれているのかどうか、また言語によってその差はあるのかどうかなどを比較することにより、日本語の特徴を検討する。

言語による情報伝達において伝達される情報を量的に表わす時、エントロピーを使用することが有効であり一般的である。文字による情報伝達で、

エントロピーの値を考える時の単位として単語と文字が考えられるが、本稿では一文字当りのエントロピーを対象とする。

一文字あたりのエントロピーは、言語の文字のマルコフ連鎖性を考えた統計的な出現確率から計算できる値であるが、英語のように基本文字数の少ない言語では容易に計算でき、日本語のように基本文字数の多い言語では計算することは困難である。

筆者は［横原1994］で、日本語、英語、中国語各言語の一文字当りのエントロピーの値を求めた。英語のエントロピーを1.0ビットとした時、日本語のエントロピーはおよそ3.0ビット、中国語のエントロピーはおよそ5.0ビットである。

情報には量と質の両面がある。量的にはある言語で書かれた文章の総情報量は、エントロピー掛ける字数で計算できる。しかし、情報の受け手は文章から情報を量的に受け取っているだけでなく、情報量では表わせない「意味」や「価値」と言う質的なものも獲得している。さらに、情報の受け手は文脈理解など言語を総合的な観点から捉えて、情報の獲得を行なっている。

言語による情報の獲得が質量両面で総合的な観点から行なわれていると考えると、次文字の統計的な出現確率から求まるエントロピーだけを、情報伝達の効率と正確さの保証の根拠とすることは必ずしも適切とは言い切れない。特に情報の正確さの確保について、情報の総合的な獲得の観点からの検討も必要ではなかろうか。

情報の受け手は、情報伝達の正確さの保証に関する次文字の予測を、文字の出現確率だけで行なっているとは限らないであろう。もし、次文字の予測を次文字の出現確率からだけで行なっているのであれば、その予測値はエントロピーの値から計算できることになる。しかし、情報の受け手が情報の獲得を文脈の理解など言語の総合的観点からも行っているなら、次文字の的中率はエントロピーから計算できる値よりも高い値となるはずである。

本稿では、日本語と英語について、次文字の的中率がエントロピーから計算できる予測率と同じかどうかを検討することによって、情報伝達を担

う言語の特徴を考察する。

## 1　次文字の確率的な予測率

文字による情報伝達において、情報伝達の正確さを読み手は読んでいる文字の次の文字を予測することで確保していると言ってもいい。一方、次文字に何が出現するかの不確かさあるいは確かさは、文字間のマルコフ連鎖性を考慮した文字の出現確率だけから得られるエントロピーの値として定量的に与えられる。

次文字の予測が文字の出現確率のみによっているのであれば、次文字の平均的な予測率はエントロピーの値から算出できる。

### 1．1　英語

英語の読み手は、次文字の予測を文字の単独な出現確率だけを考えて行なっているのではなく、前の文字との関連、マルコフ連鎖性を考えて行なっている。英語をマルコフ情報源とした時のエントロピーはおよそ1.0ビットなので、文字の条件付き出現確率は0.5である。このため次文字の出現確率による予測率は、英語の場合およそ0.5となる。

### 1．2　日本語

日本語の文字の出現確率も各文字によって異なる。日本語でも読み手は、次文字の予測を文字の単独な出現確率だけで行なっているのではなく、少なくとも前の文字との関連、マルコフ連鎖性などを考慮して行なっている。日本語のエントロピーはおよそ3.0ビットなので、文字の条件付き出現確率は0.125である。すなわち次文字の出現確率による予測率は、日本語の場合およそ0.125となる。

## 2　実験による次文字の的中率

文字で書かれた文章を読んでいく時、読み手は次の文字を予測すること

によって情報伝達の信頼性を確保している。予測した次文字の的中率を、英語と日本語について実験によって調べた。

### 2．1　英語

英語のエントロピーは、A～Zと空白の27文字の情報量を平均したものである。次文字の予測も、A～Zと空白の27文字について行った。

50個の文について、その文の後半における次文字の位置を乱数で出しその文字が何かの予測を行なった。対象は学生、日本人の英語教員、アメリカ人教員である。次文字の的中率は学生の平均値が0.33、最高が0.52であった。日本人の英語教員とアメリカ人教員の的中率は0.53と0.54であった。

この結果から、学生の英語に対する次文字的中率はおよそ0.33であるが、アメリカ人や日本人の英語習熟者では次文字的中率はおよそ0.5と考える。

### 2．2　日本語

日本語のエントロピーは、かな、カナ、漢字の2000以上の文字の情報量を平均したものである。次文字の予測も、およそ2000から4000の漢字を使用していると思われる通常の文章について行った。

25個の文について、その文の後半における次文字の位置を乱数で出して予測を行なった。対象が大学生の時では平均が0.43で、最高が0.52であった。さらに、高校生、社会人を対象に、長文として数冊の本を利用した次文字の予測を行った。40代の社会人の場合5冊の本の次文字的中率は0.52、0.54、0.47、0.53、0.46とほぼ0.5であった。特定の一冊の本で、高校生、40代、70代の人で次文字の予測を行なったところ、的中率は0.47、0.51、0.47とほぼ0.5であった。

これらの結果から、日本語の次文字の的中率はおよそ0.5と考える。

## 3　考察

英語と日本語の一文字当りのエントロピーは、各々1.0ビット、3.0ビットである。一文字当りのエントロピーは次にどういう文字が出現するかと

いう不確かさを表すので、英語の次文字の予測が確率0.5で、日本語の次文字の予測が確率0.125で可能といえる。すなわち、エントロピーから逆算できる出現確率による次文字予測率は、英語で0.5、日本語で0.125である。

一方、英語、日本語とも実験による次文字的中率はおよそ0.5であった。英語では次文字予測率と次文字的中率は同じであるが、日本語ではエントロピーから計算できる次文字予測率と実際の次文字的中率とはかなり差がある。

次文字的中率とエントロピーから得られた次文字的中率などを比較することにより、英語と日本語の特徴が考察できる。

### 3．1 英語

英語のアルファベット一文字は平均的に1.0ビットの情報量を持つ。伝達効率と信頼性に関わる冗長度についての値は、基本文字数を27文字とした時の最大エントロピー4.76ビットと実際のエントロピー1.0ビットから求まる。英語の情報伝達に関する値は次の通りである。

基本文字数はA～Z、空白の27文字
実際のエントロピー＝1.0ビット
伝達効率 1.0÷4.76＝0.21
冗長度　　1－0.21＝0.79
次文字予測率　　＝0.5　（エントロピーから計算）
次文字的中率　　＝0.5　（実験による）

英語では次文字予測率と次文字的中率とは同じ値である。

したがって、英語では次文字の予測はマルコフ連鎖を考えた条件付き出現確率だけから行われていると考えられる。文字による情報の伝達誤り防止は、マルコフ連鎖性を考えた文字の出現確率からのみ行なわれている。

### 3．2　日本語

日本語の１文字は平均的に3.0ビットの情報量を持つ。

日本語の文字の種類数を4000個としたときの最大エントロピーは12ビットであり、実際のエントロピー3.0ビットとから伝達効率と冗長度が求まる。日本語の情報伝達に関する値は次の通りである。

基本文字数はかな、カナ、漢字、空白の4000字程度
実際のエントロピー＝3.0ビット
伝達効率　3.0÷12＝0.25
冗長度　　1－0.25＝0.75
次文字予測率　　＝0.125　（エントロピーから計算）
次文字的中率　　＝0.5　　（実験による）

日本語では、次文字的中率は0.5で次文字予測率0.125の４倍である。このことから、日本語では次文字の予測は条件付き出現確率だけから行われているのではないと考えられる。文字の伝達誤り防止は文字の出現確率とマルコフ連鎖性からのみだけではなく、別の観点からも行なわれていると考えられる。

日本語の次文字の実際の的中率と条件付き出現確率から得られる統計的なエントロピーとの差の２ビットの情報量は何によってもたらされるのであろうか。

### 3．3　英語と日本語の比較

英語と日本語について、情報伝達の２つの重要な点である伝達効率と情報伝達の正確さを、前節の値をもとに比較する。

伝達効率の値は、英語、日本語とも20数％でほぼ同じである。情報伝達の正確さに関わる冗長度の値が各々70数％、次文字的中率の値が双方ともおよそ0.5とほぼ同じである。情報伝達の２つの重要な点に関わる値が同じことから、英語と日本語の情報伝達能力は同程度と考えられる。

英語と日本語の違いは、基本文字数とエントロピーの値である。基本文

字数とエントロピーは比例しており、基本文字数を多くすると、一文字当たりのエントロピーは大きくなり総文字数は少なくて済むが、文字の学習のための時間が長くなる。

基本文字数とエントロピーが大きく違う英語と日本語の情報伝達能力の値が同じであるということは、どの言語でも定量的な情報伝達能力は同じであることを示唆している。

### 3．4　次文字的中率からの日本語の考察

英語と日本語では、実用時のエントロピーが1.0ビット、3.0ビットと3倍の開きがあるにも拘わらず、次文字的中率が0.5と同じである。

英語と日本語の違いの一つは基本文字数の違いで、この違いがエントロピーの値の差となっている。

日本語の文章は数千種の漢字の使用を念頭に入れたものであり、実用時のエントロピーはおよそ3.0ビットとなる。しかし、ある文章中では①使用する漢字が実質的に限定され使用文字数がそれほど多くならない、②漢字と「かな」の交互使用によって、次文字が漢字か「かな」かの区別がしやすいなどの理由で、予測対象となる次文字の数はかなり少なくなる。次文字の予測は、限定された数の漢字と「かな」が対象となる。これが、表意文字である漢字を多数使い基本文字数の多い、日本語の次文字予測率向上の工夫であろう。

日本語の次文字的中率向上の他の要因として他に、①漢字の表意性との関連、②文脈理解のしやすさ、③漢字と「価値」、「意味」との関連、も考えられる。また、これらの要因を総合した次文字の予測が行われいるかどうかの検討も必要であるが、今後の課題とする。

## まとめ

本稿での結果をまとめると、次の3点になる。

（1）情報伝達における2つの重要な点である伝達効率と情報伝達の正確さに関わる冗長度の値は、英語と日本語でほぼ同じ値である。かなり特徴

の異なる言語である英語と日本語での結果から、他の言語の伝達効率と冗長度もほぼ同じ値であることが想定できる。

（2）次文字的中率は、英語と日本語で0.5と同じ値である。他の言語でも同じ0.5という値になることが想定できる。

（3）日本語の次文字の予測率の向上は、ある文章に使用する漢字の種類数が絞られることと、漢字と「かな」の交互使用による予測のしやすさなどによる。これが、表意文字である漢字を多数使い基本文字数の多い日本語の情報伝達の信頼性確保の工夫であろう。

日本語は、情報の伝達効率と冗長度および次文字的中率（予測率）が英語とあまりかわらない。予測対象文字の数を絞り込むことにより、次文字的中率を0.5に上げている。

日本語の情報伝達の特徴をさらに探るには、①漢字の表意性との関連、②文脈理解からの検討、③漢字と「価値」、「意味」との関連、などの点とこれらを総合的した考察が必要である。

## 参考文献

［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巻、1992

［横原　1994］横原恭士：「日本語の漢字比率と平均情報量につて」
『相愛大学研究論集』第10巻、1994

［横原　1995］横原恭士：「言語の「空白」と情報量の研究」
『相愛大学研究論集』第11巻、1995