言語による情報伝達についての一考察

——文字のエントロピーと次文字の予測から——

A Study on Verbal Communication : from the Viewpoint of Information Theory

横　原　恭　士

はじめに

記号、特に言語による情報伝達においては、情報の伝達効率と信頼性の確保が重要である。情報の伝達効率と信頼性の目安として、情報源の事象の確率から計算される情報量とエントロピーをもとに算出される伝達効率と冗長度が用いられる。

言語には英語やドイツ語のように数十個の基本文字を持つものは多い。それらの言語は文字の一つ一つは意味を持たないが、その文字の配列である単語が意味を持ち、単語の組み合わせの文で情報を表す。また漢字や仮名のように、一文字でも意味を持ち得る数千の文字を基本とする日本語や中国語のような言語もある。このように違った特徴を持つ言語による情報伝率を考える時、文字(事象)の出現確率を拠りどころにしている指標の議論を展開するだけで十分なのであろうか。言語の情報伝達を議論する場合、ほかに適切な指標はないのであろうか。本稿では、英語と日本語、両言語における情報伝達についての比較検討を、情報伝達の指標の軸である文字のエントロピーと、次文字予測率を手掛かりに行ない、言語の情報伝達についての考察を行う。

1. 情報伝達の指標

情報伝達の様子を数的に示す指標としては、情報量、エントロピー、エントロピーから導かれる伝達効率と冗長度がある。これらの指標は、情報内容の意味、価値など質的な面を考慮せず、事象(記号)の出現確率だけを根拠として情報量をもとに計算できるものである。エントロピーは次に出現する事象の不確かさの程度を表すものでもある。

情報量とエントロピー

情報量は情報源の事象の出現確率の逆数の2を底とする対数の値であり、全事象の情報量の平均値がその情報源のエントロピーである。言語の文字単位の情報量は各文字の出現確率から求められる。この情報量とエントロピーの値は、言語が使われる全ての状況を考慮したもので統計的な平均値である。

英語のアルファベットの一文字当りのエントロピーは1ビット、英語のエントロピーを基準にして算出すると、日本語の一文字当りのエントロピーは仮名漢字混じり文で3ビットである［横原 1994］。

伝達効率と冗長度

情報伝達の効率は情報源の全事象が等確率で使用されるとき時最大で1ある。人間が情報を伝えるのに用いる情報源(＝記号群)には、無記憶情報源とマルコフ情報源がある。○か×、yes か no、サイコロの目、数字などは無記憶情報源である。これらの例は全事象が等確率で使用されることが多く、その時の伝達効率は1であり冗長度は0である。数字は無記憶だが文字として扱われる場合など状況によってはマルコフ連鎖性を持つこともある。これらの無記憶情報源は伝達効率重視の情報源といえる。

これに対し、言語と音楽記号は代表的なマルコフ情報源である。音楽が作曲者や演奏者の感情を聴き手に伝える、あるいは聴き手の感情を喚起する〔Sloboda 2001〕のに対し、言語は聞き手に効率よくしかも正確な情

報を伝えること重視する。

言語の文字による情報伝達の効率ηは、文字単位のアルファベットが全て等確率で使用される時にとり得る最大のエントロピーH_{MAX}と文字の実際の使用確率から計算されるエントロピー H とから$\eta = H/H_{MAX}$で与えられる。この値は全ての状況を考慮した平均値である。$(1-\eta)$である冗長度は信頼性の目安とされている。効率が事象の出現確率だけから算出されるので、冗長度も出現確率だけを根拠とする。

無記憶情報源は効率が1で冗長度が0であるが、言語では効率は二十数%で冗長度は七十数%である。この時の効率と冗長度の値も言語が使われる全ての状況を考慮した平均値である。

不確かさ

エントロピーは情報源の不確かさ、ここでは次に出現する文字の不確かさを表しているとされる。英語のエントロピーは1ビット、英語のエントロピーを基準にした日本語のエントロピーは3ビットなので、次文字の不確かさは英語で1ビット、日本語で3ビットと言える。次文字の不確かさを次文字が何であるかを予測できる率(予測率とする)だとすると、平均的な次文字の予測率は英語では0.5、日本語では0.125ということになる。この不確かさの値も言語が使われる全ての状況を考慮した平均値である。

情報伝達の指標は、情報源が使われる全ての状況を考慮した事象の確率に基づく情報量の平均値であるエントロピーが軸である。一方、次章で議論する次文字の的中率(予測率)はコミュニケーションの実際の状況下で次の文字を正しく予測する値を表す。エントロピーの表す不確かさと実験による次文字の的中率が一致すれば、あらゆる状況下での確率的要素だけで計算された情報伝達指標である情報量、エントロピー、効率、冗長度および不確かさが、どの状況下でも指標として適用できることになる。

2. 次文字の予測

ここで言う予測とは次に出現する文字の予測のことである。エントロピーの表す不確かさと個々の状況の下での次文字の的中率(予測率)を比較し、検討する。

次文字の的中率(予測率)

[横原 1997] で、文字による情報伝達で次に出現する文字の的中率について、実験により結果を得ている。英語と日本語に習熟した者が次の文字を言い当てる率、すなわち次文字の的中率を実験で求めると、英語でも日本語でも約 50% の割合で次の文字を言い当てた。このことから英語も日本語も習熟者では次文字を正確に当てる的中率は 0.5 であると言える。この的中率は次文字の予測率とも言える。

次文字の不確かさと予測率の比較

英語のアルファベットを A～Z と空白(句読点などを含む)とすると、英語の一文字当りのエントロピーすなわち文字の不確かさは 1 ビットであり、次文字の予測率 0.5 と一致する。日本語のアルファベットは仮名、漢字と空白(句読点などを含む)で、日本語の一文字当りのエントロピーすなわち不確かさは 3 ビットであり、次文字の予測率 0.5 とは整合しない。

3. 情報伝達指標と予測率からの検討

実際の次文字の予測率 0.5 が、エントロピーの表す次文字の不確かさと、英語では一致し、日本語では数倍の開きがある。このことをもとに言語の情報伝達指標について、状況の違い、事象の形態、知識の活用性、などの点から検討を行う。

3.1 状況の違いと情報伝達指標

実際のコミュニケーションの場における情報伝達の指標は、どの状況下でも文字(事象)の出現確率に基づくエントロピーと同じとできるのであろうか。それとも固定的なものではなく、その場のあらゆる状況に対応して変動していると考えるべきなのであろうか。

変動しない情報伝達指標

情報の送り手と受け手の間でのコミュニケーションが進行している間、特定の事象の情報量は一定なのであろうか、変動しているのであろうか。情報の送り手と受け手はその場に応じて常に情報伝達の効率と信頼性に最適化を図ってコミュニケーションしており、実際の情報量は変化していると考えられる。しかし言語による情報伝達では多くの文字を使用するため、統計と確率のうえからその変化は小さく、情報量が統計値としての確率を根拠とする限り、情報量もエントロピーもどの状況の下でも値は同じとみなせる。

ある事象の情報量の根拠を確率以外の要因にも求めるなら、情報交換の進行とともにその事象が含まれる情報源の実質的な情報伝達の信頼性は向上すると考えられる。その事象に関する知識の活用度や知識そのもの内容が、コミュニケーションの進行や変化にともない変化・向上していることによると考えられる。

しかし、情報量を確率だけに基づくとする限りコミュニケーションの進行とともに刻々と変化していると考えるべきではないであろう。

エントロピーは事象の確率を根拠にしているということ、また定義から全ての状況を含んだ全状況の平均値であることから、状況の違いによってその値は変わらないと考えるべきである。伝達効率も冗長度もエントロピーから導かれるものであり、その値は事象の出現確率で決定される。情報伝達の指標である情報量、エントロピー、エントロピーから導かれる伝達効率と冗長度は、全て事象の確率だけをもとに計算されている。

不確かさと予測率

エントロピーは情報源の全状況を考慮した平均の不確かさを表すとされる。しかし、実際の異なった状況下での不確かさは、全状況下での確率的要素だけを根拠にしているエントロピーをもとにした不確かさと同じではないであろう。人は実際の状況に応じて時々刻々最適な情報交換を行っていると考えられるので、全状況での不確かさと、異なる個々の状況下での不確かさは同じではないと考えるべきである。

情報の受け手は、次文字の予測をその場の状況に応じて行っている。会話の進行にともない、その瞬間々々の情報を効率よく信頼性の高い最適な方法で受け取ろうとする。次に出現する文字の予測は、できるだけ正確に送り手からの情報を受け取ろうとする方策なのである。次文字予測の根拠を確率以外の要因にも求めるなら、事象に関する知識の活用度や知識そのもの内容が、コミュニケーションの進行や変化にともない変化・向上していることによると考えられる。次文字の予測は不確かさは事象の確率にもとづくだけではない。

言語による情報伝達では信頼性の高い情報伝達を常に保つため、次事象の予測値を最高の値である 50% に維持し続ける。エントロピーは次事象の不確かさを表さない。

3.2 事象の形態と情報伝達指標

事象の形態の違いによって情報伝達の指標はどう違うのであろうか。

事象の形態と意味

言語という情報源は有限個の事象の集合であり、事象の捉え方は色々考えられる。事象を文字単位で捉えると、英語のアルファベットは言語の情報源の最小要素で、I や a を除いてその一つ一つは意味を持たずその繋がりは確率とマルコフ性で議論できる。また、日本語のアルファベットの仮名や漢字は一つでも意味を持ち得る。

事象の形態は、英語の場合はアルファベット、音素、単語、単文、文章であり、日本語の場合は音素、単語、単文、文章である。英語のアルファ

ベットと音素は意味をもたず、単語、単文、文章となって意味を持つ。日本語の漢字には一文字で単語になるものもあり、複数個で単語になるものもあって一文字でも意味を持ちうる。また、仮名も助詞や名詞で一文字が意味を持つ場合もある。単語は両言語で意味を持つ単位である。

英語のアルファベット：それがひとつだけでは意味を持たない。
（I と a は例外）
かな（音素）：ひとつで単語になり意味を持つものもある。（助詞、かなで書かれた名詞）
漢字　　：ひとつで意味を持つものは多い。（葉、木、鳥……）
単語　　：意味を持つが品詞によりその重みは違う。（名詞、動詞）と（冠詞、助詞）
文・文章　：意味を持ち情報となりうる。

事象の形態の違いによる知識の活用性の違いということから考えてみると、英語のアルファベットのように一つでは意味を持たない事象は、出現確率とマルコフ連鎖だけで文字単位の情報伝達を議論できるし、意味や知識ネットワークを考慮する必要はない。一つの事象で意味を持ちうる仮名や漢字、単語、単文、文章などは出現確率とマルコフ連鎖だけでは情報伝達を議論できず、知識のネットワークあるいは意味のネットワークなどとの関係も考慮すべきである。

事象の形態と予測率

事象の形態の違いによる情報伝達を考えるとき、確率だけでなく意味を持ち知識のネットワークの利用を含めば、情報量やエントロピーだけにもとづく議論だけでは十分ではない。予測率を情報伝達の指標として導入すればよいのではないか。

4. 事象の形態の違いと情報伝達

言語による実際の状況下での情報伝達で、事象の出現確率から計算されるエントロピーにもとづく指標は変化しない。文字のエントロピーは以下のようであり、異なる言語間の情報伝達指標の比較では、このエントロピーを基に議論すればよい。

表1　文字のエントロピー

英語1文字のエントロピー	1ビット〔横原 1992〕
仮名1文字のエントロピー	1.5～2ビット〔横原 1992〕
日本語の漢字1文字のエントロピー	4～5ビット〔横原 1994〕
日本語の1文字のエントロピー	3ビット〔横原 1995〕
英単語1つのエントロピー	5.6ビット〔横原 1993〕
中国語1文字のエントロピー	5ビット〔横原 1995〕

しかし、不確かさと次文字予測率は状況や事象の形態の違いによって同じではないことが分かった。言語の情報伝達を、事象の形態—英語のアルファベット、日本語、単語、単文、文章—の違いごとに整理する。

(1) 英語のアルファベット

一つ一つの事象は単なる記号であり確率的要素だけで連結しており、情報伝達指標は全状況でも個々の状況下でも殆ど変わらない。エントロピーは1ビットで、不確かさと予測率は50% で一致する。一方、事象の確率にもとづく計算上の伝達効率と冗長度は H_{max}=4.75 ビットなので、0.21 と 0.79 となる。エントロピーは1文字当たり1ビットであるので、不確かさと予測率の値は矛盾しない。また、予測率が50% であることから、英語は26個の文字を使用しているが、信頼性という点では確率50% の2事象の情報源と等価である。

(2) 日本語

日本語のエントロピーは文中の漢字の比率によって幅があるが、およそ3.0ビットである〔横原 1994〕。この値は英語のエントロピーを基準としたものであり、事象の確率だけをもとに情報量、効率、冗長度などを議論するときにはこの値を使用できる。この値を不確かさの目安とすると、次文字の予測値(的中率)とは一致しない。予測値(的中率)は個々の状況下での実測値であり、予測率が50%であることから、数千の文字を使用する日本語も情報の信頼性ということでは、英語と同様に確率50%の2事象の情報源と等価である。

(3) 単語

〔横原 1994、1995〕より英語の単語の平均長さは4.65文字であるので、英語のアルファベットのエントロピーを基準にすると、英語の単語のエントロピーは4.65ビットとなる。日本語の場合も品詞の数は英語の文章と変わらないと思われるので、単語のエントロピーは英語の空白を考慮すると5.5ビット程度であろう。単語単位の伝達効率、冗長度、不確かさ、予測率(的中率)についての詳細な検討は今後の課題とする。

(4) 単文、文章

英語のアルファベットのエントロピーを基準にすると、英語の単文と文章の場合、文字数に1ビットを掛けると英語の単文と文章の情報量となる。日本語の単文と文章の場合は、英語のアルファベットのエントロピーを基準にすると、文字数に〔横原 1994〕漢字比率に応じたエントロピーの値を掛けると日本語の単文と文章の情報量となる。単文、文章単位の伝達効率、冗長度、不確かさ、予測率(的中率)については今後の課題とする。

ま と め

(1) 言語の情報伝達での予測率という指標

情報伝達の指標として、事象の確率から計算される情報量、エントロピー、効率、冗長度がある。しかし、言語による情報伝達は事象の確率だけ

でなく、意味や知識を考慮した指標からの検討も必要である。

a. 確率にもとづくエントロピーを軸にする指標：情報量、エントロピー、効率、冗長度（このときの文字のエントロピーは表1の値）

b. 意味・知識を考慮した情報の信頼性に関わる指標：次文字の予測率（＝不確かさ）

(2)「言語の文字による情報伝達では、次文字予測率（＝不確かさ）は常に50% である。」

事象の出現確率だけによる次事象の予測では、いろいろな形態の事象による情報伝達の信頼性を確保できない。言語の次文字予測率は、文脈、状況、意味理解など総合的な情報の利用による結果である。事象の形態が異なっても、意味や知識を活用できる場合、次文字予測率によって情報伝達の信頼性を確保できる。

文字を事象の形態とするとき、英語と日本語では次文字の予測率が50%であるので、これ以上予測率を上げることはできない。文字による情報伝達は最適なのである。日本語の文字のエントロピーにもとづく不確かさより予測率が高い理由は、確率以外の事象の意味や知識などの活用による。

言語の他の事象の形態に拡張すると、言語におけるどの形態の事象の場合も不確かさ、次事象の予測率は50% と考えられ、どの形態の事象も確率だけでなく知識のネットワークの活用などなど高度な情報活動によって、次事象の予測を行なっていることが推測できる。

(3)「言語は2事象情報源と等価である。」

英語と日本語で任意の文字の次の文字は常に50% の確率で予測できるので、

「英語と日本語の文字の信頼性は50% の出現確率を持つ2事象の情報源と等価であり、情報の伝達効率は1である。」

このことは、事象が英語と日本語以外の言語の文字、事象が単語、単文、文章の場合でも成り立つであろう。

結論：「英語と日本語の文字の情報量の値は表 1 を基準とできる。英語と日本語の文字は 50% の出現確率を持つ 2 事象の情報源と等価で、効率は 1 で信頼性を示す次文字の予測率は 50% である。」

「言語の情報量は言語、事象の形態によって異なる。(4 章)」

仮説：「言語はどの言語どの形態の事象でもでも 50% の出現確率を持つ 2 事象の情報源と等価であり、効率は 1 で信頼性を示す次文字の予測率は 50% である。」

本稿では、言語の情報伝達を文字の場合で検討したが、事象が単語、文、文章に場合については仮説の域を出ない。事象が文字以外の場合での検証は今後の研究課題である。

参考文献

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

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

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

横原恭士：「日本語の字文字予測からの検討」『相愛大学研究論集』第 14(1) 巻、1997

Sloboda, J. A. & Juslin, P. N. ; *Music and Emotion* OXFORD Univ. Press, New York, 2001