言語の「空白」と情報量の研究

——書き言葉の「空白」と話言葉の情報量——

A Study on "Blank" and "Quantity of Information" of Language

横原恭士

1．はじめに

言語による情報伝達は話言葉による場合と、書き言葉による場合がある。話言葉による場合は情報を言葉で、書き言葉による場合は情報を文字で単語や文にして伝える。話言葉は有声部分と無声部分からなるが、有声部分のアクセント、さらに顔の表情やジェスチャーなども話言葉による情報伝達には欠かせない。書き言葉による情報伝達では文字以外にも文の構成要素として、句読点・空白・各種記号が使われる。

言語による情報伝達において、情報の内容を直接表現していない話言葉の単語と単語の間および文と文の間の無声部分や、書き言葉の句読点・空白・各種記号は単語や文を明確にするものとして、また言葉や文字で表現できない情報を伝えるものとして必要である。これら広い意味の記号は、情報伝達の信頼性の点で文字に劣らない重要な役割を果たしている。

話言葉では無声部分はどの言語においても単語と単語の間、文と文の間に必ずあり、書き言葉では句読点は全ての言語にあるが、空白は英語などの表音文字系の言語では単語間にあっても、中国語や日本語のような表意文字を使う言語では単語間にはない。

本論文では、情報伝達の信頼性向上の点で重要な役割を担っている話言葉の無声部分と、書き言葉の句読点と空白の役割に情報量やエントロピーによる定量的な検討を加える。検討の対象とする言語は表音文字を使う英語、表意文字を使う中国語、表意文字と表音文字の混成文字を使う日本語の三言語である。

2．書き言葉の「空白」、話言葉の無声部分

本論文では、書き言葉の「空白」と話言葉の無声部分を次のように考える。

書き言葉である文字で書かれた文章では、文と文を句切る句読点と各種記号と表音文字における単語と単語の空白を「空白」とする。書き言葉では、「空白」は文字と同じ大きさのスペースを占めており、明確にその部分が単語と単語の間、文と文の間の句切りであることがわかる。

話言葉では、単語と単語の間と文と文の間の句切りに無声部分がある。さらに一つの単語の中にも無声部分が見られる。無声部分のうち単語の一部とはみなされない有意の長さの無声部分が書き言葉の「空白」に対応している。

3．書き言葉の「空白」のエントロピー

書き言葉では文字の組合せで単語を作り、単語の組合せで文を作り、さらにいくつかの文の集まりを文章として情報を伝えている。文章に「空白」は必要不可欠であり、「空白」が情報の伝達をより信頼性の高いものとしている。英語、中国語、日本語の三言語とも「空白」には文字一字分のスペースが割り当てられている。本節では、書き言葉における「空白」のエントロピーについて検討する。

3．1　文字のエントロピーと対訳文の総情報量

文字のエントロピーと言う時、それは「空白」のエントロピーを含んでいる。三言語の「空白」を含めた文字のエントロピーは［横原1994］で得られている。［横原1994］で示された三言語の文字のエントロピーの値の妥当性を三言語間の関係において確認するため、日英、日中対訳文の総情報量を比較検討した。

その手順は次の通りである。

(1) 英語と日本語、中国語と日本語の対訳文の文字数と空白数を調べる。日本語については漢字比も調べる。

(2) 文字全体のエントロピーとして、［横原1994］の値を使って三言語の文の総情報量を計算する。

(3) 日英対訳文の日本語と英語、日中対訳文の日本語と中国語の総情報量が等しいかどうか比較検討する。

対象とした対訳文は次の文献の一部の文章である。

日英対訳文……［天声人語 '93春］

日中対訳文……［現代中国経済］［小さな出来事］［人間の始まり］

結果を次表に示す。

表１．対訳文の文字数と総情報量

文章		文字数			総情報量（ビット）
		空白数	表音文字	漢字（漢字比）	
天声人語	英 語	6279	29351	― ― ―	1.0×35630≒36000
'93春	日本語	776	6074	3921（0.36）	3.3×10771≒36000
現代	中国語	129	― ― ―	1338	5.5×1467≒8100
中国経済	日本語	138	1147	1074（0.45）	3.4×2359≒8000
小さな	中国語	27	― ― ―	229	5.0×256≒1300
出来事	日本語	34	333	110（0.23）	2.8×477≒1300
人間の	中国語	95	― ― ―	713	5.0×808≒4000
始まり	日本語	89	1021	269（0.20）	2.7×1379≒3700

＊［天声人語 '93春］は20の文章の合計である。

＊文字のエントロピーは［横原1994］から次の値とした。

a．英語は文字、空白とも1.0ビット。

b．中国語は［現代中国経済］の基本文字数は9000、［小さな出来事］［人間の始まり］の基本文字数は6000とし［横原1994－図10］から、各々5.5と5.0ビット。

c．日本語は［横原1994－図9b］から、漢字比の差により3.3、3.4、2.8、2.7ビットとした。

対訳文の総情報量の比較から、日英対訳文［天声人語 '93春］の英語と日本語の総情報量は各々36000ビットと36000ビットで等しい。また日中対訳文の中国語と日本語の総情報量は、［現代中国経済］で8100ビットと8000ビット、［小さな出来事］で1300ビットと1300ビット、［人間の始まり］で4000ビットと3700ビットと三つの文章でほぼ同じ値である。このことから、［横原1994］の三言語の文字のエントロピーの値が妥当なものであると言える。ある同じ内容の情報を書き言葉によって伝える場合、どのような言語を比べても伝える総情報量は等しいと言う［横原1993］での主張が日中対訳文で裏付けられた。従って、次節の「空白」のエントロピーの検討では、言語の文字のエントロピーとして［横原1994］の値を使う。

［横原1994］と本論文の今までの議論では、「空白」のエントロピーの値は英語のアルファベットや仮名のエントロピーと同じ値であるとしてきた。この英語のアルファベットと仮名のエントロピーの値は、情報源を多重マルコフ情報源とした実際の使用時（極限エントロピー）を考えると、アルファベットでは1.0ビットであり、仮名では［横原1994－図9a］より漢字比による幅があり、2.0～30.ビットである。しかし、「空白」のエントロピーはアルファベットや仮名のエントロピーと同じではない。「空白」のエントロピー、さらに「空白」を句点、読点、単語間空白と分けた時の各々のエントロピーを検討するため、いくつかの文章の空白数と文字数の割合を調べた。

3．2　「空白」数と文字数

「空白」数と文字数は次の文献の一部の文章を調べた。

英　語……［天声人語 '93春］［Robinson Crusoe］

中国語……［現代中国経済］［中国文化基礎知識］［小さな出来事］［人間の始まり］

日本語……［天声人語 '93春］［現代中国経済］［小さな出来事］［人間の始まり］
［杜子春］

「空白」は単語間、読点、句点、その他（各種記号）に分けその割合を調べた。

表2．文章の「空白」数と文字数の割合

文　章		「空白」数の割合（%）					文字数の割合（%）		
		単語間	読点	句点	その他	合計	表音	漢字	合計
英語	天声人語 '93春	16.0	0.6	1.0		17.6	82.4	……	82.4
	Robinson Crusoe	20.9	1.5	1.6		23.2	79.0	……	79.0
中国語	現代中国経済	……	5.0	2.4	1.4	8.8	……	91.2	91.2
	中国文化基礎知識	……	5.2	3.8	1.1	10.1	……	89.9	89.9
	小さな出来事	……	8.1	2.7	0.4	11.2	……	88.8	88.8
	人間の始まり	……	7.7	3.2	0.9	11.8	……	88.2	88.2
日本語	天声人語 '93春	……	3.9	3.3		7.2	56.4	36.4	92.8
	現代中国経済	……	3.9	1.7	0.2	5.9	48.6	45.5	94.2
	小さな出来事	……	4.2	2.9		7.1	69.8	23.1	92.9
	人間の始まり	……	4.1	2.3		6.5	74.0	19.5	93.5
	杜子春	……	5.6	1.9		7.4	63.0	29.6	92.6

この結果から、言語によって「空白」の「空白」を含めた全文字数に対する割合が異なっていることが分かる。その割合は英語では約20%、中国語では約10%、日本語では約7%である。三言語とも全文字数に対する「空白」の割合は個々の文字の割合に比べて大きく、したがって「空白」の文字のエントロピーの値に与える影響もかなり大きいと言える。さらに、同じ言語の文章の間でも「空白」数と文字数の割合に若干の差が見られる。ここでは次節で検討する「空白」のエントロピー算出のための三言語の「空白」数と文字数の割合の基準値を、表2と［佐藤］を参考に表3のように設定する。またこの基準とする言語の極限エントロピーは［横原1994］を参考に、英語1.0ビット、中国語5.0ビット、日本語3.0ビットとする。

表3．基準三言語の「空白」数と文字数の割合

言語	「空白」数の割合（%）				文字数の割合（%）			極限
	単語間	読点	句点	合計	表音	漢字	合計	エントロピー
英　語	16.6	0.8	1.2	18.6	81.4	……	81.4	1.0ビット
中国語	……	7.0	3.0	10.0	……	90.0	90.0	5.0ビット
日本語	……	4.1	2.6	6.7	62.0	31.3	93.3	3.0ビット

＊表2の「空白」数のその他は読点に含めた。

3．3　「空白」のエントロピー

基準三言語の「空白」数と文字数の割合から、「空白」全体と「空白」を単語間空白、読点、句点に分けた時の各々のエントロピーの値を次の手順により求める。

(1) 言語を多重マルコフ情報源とした実際使用時のエントロピー（極限エントロピー）と、文字および「空白」を完全事象系とした時のエントロピーとの比をaとすると、次式が成り立つ。

$$-a\{\sum_{i=1}^{n} P_{ci}\log_2(P_{ci})+P_b\log_2(P_b)\}=\text{極限エントロピー}$$

＊各文字の出現確率…P_{ci}

＊「空白」の出現確率…P_b

(2) P_{ci}を各文字均等としてaを求める。

(3) $a\times\log_2(P_b)$から「空白」全体のエントロピーを計算する。

(4) 「空白」全体のエントロピーから、単語間空白、読点、句点、のエントロピーを求めるための係数bを次式より求める。

$$P_b\log_2(P_b)=b\{P_{bw}\log_2(P_{bw})+P_{bt}\log_2(P_{bt})+P_{bk}\log_2(P_{bk})\}$$

＊単語間空白の出現確率…P_{bw}

＊読点の出現確率…P_{bt}

＊句点の出現確率…P_{bk}

(5) 単語間空白、読点、句点のエントロピーを$b\times\log_2(P_{bw})$、$b\times\log_2(P_{bt})$、$b\times\log_2(P_{bk})$から求める。

基準三言語について求めた「空白」と文字のエントロピーを表4に示す。

表4．「空白」と文字のエントロピー

言語	「空白」				文字		
	単語間	読点	句点	合計	表音文字	漢字	合計
英語	0.50	1.36	1.24	0.54	1.11	……	1.11
中国語	……	1.29	1.70	1.41	……	5.40	5.40
日本語	……	1.35	1.54	1.35	2.44	4.49	3.12

3．4　結論

(1) 対訳文の総情報量は、日本語と英語の間、日本語と中国語の間で等しい。このことにより、同じ内容の情報を伝える文字で書かれた文章全体の総情報量は、言語が異なっても等しいと言う［横原1994］の仮説が裏付けられる。

(2) 英語の極限エントロピーを1.0ビットとすると、日本語と中国語の極限エントロピーの値は［横原1994－図9、図10］から得ることができる。

(3) 「空白」を単語間空白、読点、句点と分けた時、各々のエントロピーおよび文字のエントロピーは表4の通りとなる。

① 「空白」全体のエントロピー、句点のエントロピーの大小関係は文字の極限エントロピーの大小関係と同じ傾向にある。　中国語＞日本語＞英語

② 読点のエントロピーは三言語ともほぼ同じ値約1.3ビットである。

③ 「空白」を含む文字のエントロピー（英語1.0ビット、中国語5.0ビット、日本語3.0ビット）より、「空白」を除いた文字のエントロピーは少し増加する。「空白」のエントロピーは、中国語と日本語と英語の単語間空白では減少するが、英語の句読点では増加する。

4．話言葉の情報量

話言葉では書き言葉の文字の部分に対応する有声部分と、書き言葉の「空白」に相当する無声部分との組合せで情報を伝達する。書き言葉との違いはアクセント、表情、ジェスチャーなども情報の伝達に関っていることである。本論文では、アクセント、表情、ジェスチャーは検討の対象から除外し、話言葉の有声部分と無声部分の情報量について考察する。

書き言葉では「空白」の占めるスペースは文字一字分と一定であるが、話言葉では一つの無声部分の占める時間の長さは一定ではない。単語と単語の間の無声部分と句読点の無声部分では時間の長さが違うし、単語と単語の間の無声部分や文と文の間の無声部分の長さも一定とは限らない。同じ言語でも話す内容や話し方によっても無声部分の長さの違い

がでてくる。また言語による差もあるであろう。三言語の話言葉の有声部分と無声部分の時間の長さの測定を、時間軸に沿った音声波形により行った。

4．1　音声時間

英語、中国語、日本語の話言葉の音声時間の測定は、次の市販のカセットテープの音声を記録することにより行った。

英　語……"Robinson Crusoe" Longman Group UK Limited 1987

中国語……「中国文化基礎知識」東方書店　1986

日本語……芥川龍之介　「杜子春」　新潮カセットブック　新潮社　1988

音声波形の記録を有声時間と無声時間に分け測定すると、話言葉の無声部分は次のように分類できる。一、文中の句点に対応する文と文の間の句切りの無声部分。二、文中の読点に対応する無声部分。三、単語と単語の間の無声部分。四、単語の中の無声部分である。

音声波形の分析より得た無声部分のうち、句点と読点に対応する部分一つ当りの時間は次表のとおりである。単語と単語の間、単語の中の無声部分は句読点の無声部分と比べると非常に短い。

表5．句読点に対応する無声部分の時間（秒）

	読点	句点
Robinson Crusoe　（英語）	0.51	1.22
中国文化基礎知識（中国語）	0.52	1.20
杜子春　　（日本語）	0.38	1.18

句点での無声部分の時間は、三言語とも約1.2秒とほぼ同じである。読点での無声部分の時間は英語と中国語が約0.5秒、日本語が0.38秒である。読点での無声部分の時間に言語による差があるものの、句点の箇所の無声時間が三言語とも同じことから、三言語のカセットテープとも話す速さは同じであると言ってよいのであろう。また、単語と単語の間の無声部分と単語の中の無声部分は非常に短いので、文の一部と見なすことにする。

4．2　時間比率と1秒当りの情報量

「文」時間を導入し、話言葉の時間が「文」時間と句読点の無声部分の時間から成るとする。単語の中の無声部分の時間と単語と単語の間の無声部分の時間と有声部分の時間と合わせた時間を「文」時間とする。「文」時間の導入により、話言葉の時間は、句点、読点、「文」時間の三つに分けられる。音声波形の分析から算出した句点、読点、「文」時間の比率を表6に示す。話言葉の総情報量として文字のエントロピーから算出した総情報量を用い、総情報量を話言葉の総時間で割った1秒当りの情報量、総情報量を「文」時間で

割った1秒当りの情報量も表6に示す。ここで、英語の文字のエントロピーは1.0ビットとし、中国語の文字のエントロピーは「中国文化基礎知識」の基本文字数を5000と考え［横原1994－図10］から4.8ビットとし、日本語の文字のエントロピーは、［横原1994－図9b］で「杜子春」の漢字比29.6%から3.1ビットとした。

表6．時間比率と情報量（ビット）

文　章	時間比率			1秒当りの情報量	「文」時間1秒当りの情報量
	「文」	読点	句点		
Robinson Crusoe（英語）	0.66	0.10	0.25	12.9	19.7
中国文化基礎知識（中国語）	0.73	0.09	0.18	15.2	20.8
杜子春　（日本語）	0.79	0.10	0.11	15.4	19.5

「文」時間1秒当りの情報量は中国語でやや高いものの、三言語でほぼ同じ値である。情報が句読点の無声部分以外の「文」時間中に伝達されると考えると、情報量の伝達速度は言語によって差はないと言える。

「文」の時間比率と1秒当りの情報量には三言語に差が見られるが、これは言語間の差というより、文章の難易の差と言うべきであろう。"Robison Crusoe"は平易な文章であるので、句点の割合も多く1秒当りの情報量も少ない。

さらに、三言語の種々の文章による比較を行うため、表2の文章と表3の基準三言語について表6をもとに係数を掛けて句点と読点の時間を求め、「文」・読点・句点の時間比率と1秒当りの情報量の値を求めた（表7）。

表7．種々の文章の時間比率と1秒当りの情報量

文　章		時間比率			1秒当りの情報量（ビット）
		「文」	読点	句点	
英語	Robinson Crusoe	0.65	0.10	0.25	12.9
	天声人語 '93春	0.77	0.05	0.18	15.2
	基準英語	0.73	0.06	0.21	14.4
中国語	中国文化基礎知識	0.73	0.09	0.18	15.2
	現代中国経済	0.78	0.10	0.12	16.3
	小さな出来事	0.75	0.13	0.13	15.5
	人間の始まり	0.73	0.12	0.15	15.1
	基準中国語	0.75	0.10	0.14	15.6
日本語	杜子春	0.79	0.10	0.11	15.4
	天声人語 '93春	0.75	0.07	0.18	14.6
	現代中国経済	0.82	0.08	0.10	16.0
	小さな出来事	0.76	0.08	0.16	14.9
	人間の始まり	0.79	0.08	0.13	15.4
	基準日本語	0.78	0.08	0.15	15.2

1秒当りの情報量は、三言語とも約15ビットとほぼ同じとなった。

4．3　効率・冗長度からの考察

話言葉の1秒当りに伝達できる情報量は三言語ともほぼ同じであった。しかし、書き言葉の効率・冗長度は三言語間にかなりの差がある。ここでは書き言葉の効率・冗長度からの検討を行う。極限エントロピーの値が英語1.0ビット、中国語5.0ビット、日本語3.0ビットの時の書き言葉の効率及び冗長度は表8に示すようになる。この書き言葉に対応する話言葉の情報量と、書き言葉の冗長度が0（効率が1）の時に対応する話言葉の情報量も表8に示す。

表8．書き言葉の効率・冗長度および話言葉の情報量

言 語	書き言葉		話言葉の情報量（ビット）			
			冗長度有り		冗長度0（仮定）	
	効率	冗長度	1秒当り	「文」1秒当り	1秒当り	「文」1秒当り
英　語	0.21	0.79	14.4	19.7	68.5	93.8
中国語	0.41	0.59	15.6	20.8	38.4	51.2
日本語	0.27	0.73	15.2	19.5	55.8	71.6

今までのことから次のことが言える。

(1)書き言葉では、基本文字数が多いほど情報の伝達効率は高く冗長度は小さい。いいかえると、同じ情報を伝達する言語を比べると、基本文字数の多い言語ほど少ない文字数で効率よく情報を送ることができる。基本文字数の多い言語は書き言葉に適した言語と言える。

(2)話言葉では、1秒当りに送れる情報量は約20ビットで言語による差はない。

(3)英語の話言葉では冗長度を0にすると、1秒当り約100ビット近くの情報を送ることができる。しかし、中国語では冗長度を0にしても、1秒当り50ビットの情報しか送れない。英語は基本文字数が少ないので、冗長度を下げ情報伝達の信頼性を犠牲にすれば、情報伝達の速度を上げることが日本語や中国語に比べて容易なのである。実際には信頼性を確保するため、冗長度を下げ話す速度をそれほど上げることはできない。英語は基本文字数が少なく文字の学習が容易であり、データあるいはデータに近い冗長度をあまり必要としない情報を言葉で送るのに適した言語と言える。

4．4　結論

(1)　話言葉では、三言語とも句点一つ当りの時間は約1.2秒と同じである。

(2)　話言葉では英語、中国語、日本語の三言語とも1秒当りの情報量は約15ビットと同じであり、1秒当りの情報量に言語間の差はない。句読点時間を除いた「文」時間1

秒当りの情報量は三言語とも約20ビットと同じである。

5．まとめ

書き言葉の「空白」のエントロピーと話言葉の情報量からの検討結果をまとめると、次のようになる。

(1) 日英、日中対訳文の総情報量の比較から、同じ内容の情報を伝える文章全体の文字の総情報量は、言語が異なっても等しいと言う［横原1994］の仮説が裏付けられる。

(2) 英語の極限エントロピーを1.0ビットとすると、日本語と中国語の極限エントロピーの値は［横原1994－図9、図10］から得ることができる。

(3) 「空白」を単語間空白、読点、句点と分けた時、各々のエントロピーおよび文字のエントロピーは表4の通りとなる。

(4) 言語としては、基本文字数を少なくして効率を下げて冗長度を上げるか、基本文字数を多くして冗長度を下げて効率を上げるかの選択がある。

(5) 話言葉では英語・中国語・日本語の間で1秒当りの情報量は約15ビットと同じであり、1秒当りの情報量に言語間に差はない。句読点時間を除いた時間1秒当りの情報量は三言語とも約20ビットと同じである。

(6) 話言葉の単位時間当りの情報伝達効率はどの言語でも同じであるが、文字のエントロピーは言語によって異なる。

(7) 書き言葉と話言葉での三言語の比較を表9に示す。

表9．三言語の比較

言　語	基本文字数	書き言葉効率(冗長度)	文字学習	文字エントロピー	文章文字数	話言葉1秒当り情報量	話言葉の速度アップ
英　語	27	0.21 (0.79)	易	1.0 ビット	多	20 ビット	冗長度の不要な情報の時にはかなり可能
中国語	5000	0.41 (0.59)	難	5.0 ビット	少		難しい
日本語	2017	0.27 (0.73)	やや難	3.0 ビット	中		かなり難しい

［横原1992］［横原1993］［横原1994］と本論文の書き言葉の「空白」と話言葉の情報量を検討することによる英語、中国語、日本語の比較から、表意文字は書き言葉に適し表音文字は話言葉に適しているという結果が得られる。

参考文献

[横原1994] 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

[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

[現代中国経済] 佐々木信彰（編）：〔中日対訳〕「現代中国経済」 東方書店、39-45、1994

[小さな出来事] 魯迅、横山宏（訳）：対訳／魯迅画文選集「小さな出来事」 同時代社、8-15、1988

[人間の始まり] 孫宗光：原文対照 中国模範訳シリーズ「人間の始まり」 国書刊行会、2-5、1986

[Robinson Crusoe]Daniel Defoe:「Robinson Crusoe」(simplified edition)Longman Group UK Limited, 1987

[中国文化基礎知識] 中山時子（監）：「中国文化基礎知識」 東方書店、8-27、1986

[杜子春] 芥川龍之介：「鼻・芋粥・杜子春」 新学社文庫、新学社、101、1968

[佐藤] 佐藤憲市：「情報科学」 八千代出版、63、1982

[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