語学力における情報理論の適用に関する研究

志　方　　泰

緒　　言

C. E. SHANNON による情報理論は現在では情報工学なる学問分野にまで発展をなしている。我々は範囲を更に拡大すべく語学力評価に適用を試みた。即ち語学力の一つの普遍的測度として対象国語の平均情報量もしくはエントロピーを考え，各人の有するその量を測定してこの結果より能力を客観的に判断する。斯の方法によると一般家庭において簡単に測定を為し得るので従来比較的把握困難であった家族の語学力が判断されるのでことに教育に関心を有する父母にとって有用であろう。この考えのもとに本学学生を対象に実験を行なった所予期のとおりとなり実用の見通しがつけられたのでとり敢ずここに御報告する次第である。

§1. 実 験 方 法

N字についてのエントロピーの定義式は

$$H_N=-\sum_{a} P_a(N)\log P_a(N) \qquad \cdots\cdots\cdots\cdots(1)$$

であることは衆知のとおりである。

ここに $P_a(N)$ は N字中に現われる文字の頻度確率である。対数の底は2であり Hは全てビットで表わす。

従って，MONOGRAM ENTROPY を H_1
DUOGRAM ENTROPY を H_2
TRIGRAM ENTROPY を H_3

と表わすと

$$H_1=-\sum_{i=a}^{Z} P_i\log P_i$$

$$H_2=-\sum_{ij=aa}^{ZZ} P_{ij}\log P_{ij}$$

$$H_3=-\sum_{ijk=aaa}^{ZZZ} P_{ijk}\log P_{ijk}$$

となる。N—GRAM ENTROPY H_N の1字あたりの値 $1/N\ H_N$ は　Nが大になると急速に一定値に近づく，この極限値をHとすると

$$H=\lim_{N\to\infty}\frac{1}{N}H_N\fallingdotseq\frac{1}{S}H_S \qquad \cdots\cdots\cdots\cdots(2)$$

となる。S は各国語についてほぼ100程度以下の値である。

更に

$$F_N=H_N-H_{N-1} \qquad \cdots\cdots(3)$$

とすると

$$H=\lim_{N\to\infty} F_N \qquad \cdots\cdots(4)$$

であり，一方(3)式は

$$F_N=-\sum_{ij}P_{bij}\log P_{bi(j)}$$

$$=-\sum_{ij}P_{bij}\log P_{bij}+\sum_{i}P_{bi}\log P_{bi} \qquad \cdots\cdots(5)$$

ここに　$b_i=(N-1)$ GRAM

$j=b_i$ につづく任意の文字

P_{bij} $=N-$GRAM b_{ij} の確率

$P_{bi(j)}=b_i$ のつぎに j のくる条件付き確率

$=P_{bij}/P_{bi}$

である。特に

$$F_0=\log n \qquad \cdots\cdots(6)$$

とする。

(5)式および(6)式により N 字連続の度数統計から F_0, F_1, ……F_N が計算によって求められるが，言語統計によっても F_8 以上は，莫大なデータを必要とし計算時間も甚大となり実用上求めることは困難であり実験によらなくてはならない。また仮に手数と時間に関係なく計算を行なっても得られるものは理論的な値である。我々の対象としてはこの値に対してではなく，同様の経験を有する者の総合の平均（例えば実験対象者が大学1年生ならば大学1年生の集合の平均）でなければならない。その為にも積極的に実験を行なう意義が存在する次第である。

実験としては被実験者が未経験である文章を選び，文章の1部 $N-1$ 文字を選び次の N 字目を当てさせる。この予測が正しくない場合は実験者は"否"と答え別の文字をあてさせ，何回目に正しい答が出たかを記録する。この様な実験を無数の $N-1$ 文字について行なうとき $N-1$ 文字を与えて次の1字が i 番目であたる確率 $q_i(N)$ が求められる。

この時 F_N の上限および下限は

$$\sum_{i=1}^{i}(q_i(N)-q_{i+1}(N))\log i\leq F_N\leq -\sum_{i=1}q_i(N)\log q_i(N) \qquad \cdots\cdots(7)$$

により求められる。実際問題として無限回の試行は不可能であるので各N字目につき数十回以上実験を行なって代用し，代表値として上限の値を採用した。

§ 3. 語学能力と情報量—エントロピー—の関係

前述のようにして $N-1$ 字を与え次の N 字目を当てさせる場合，文章構造および単語等

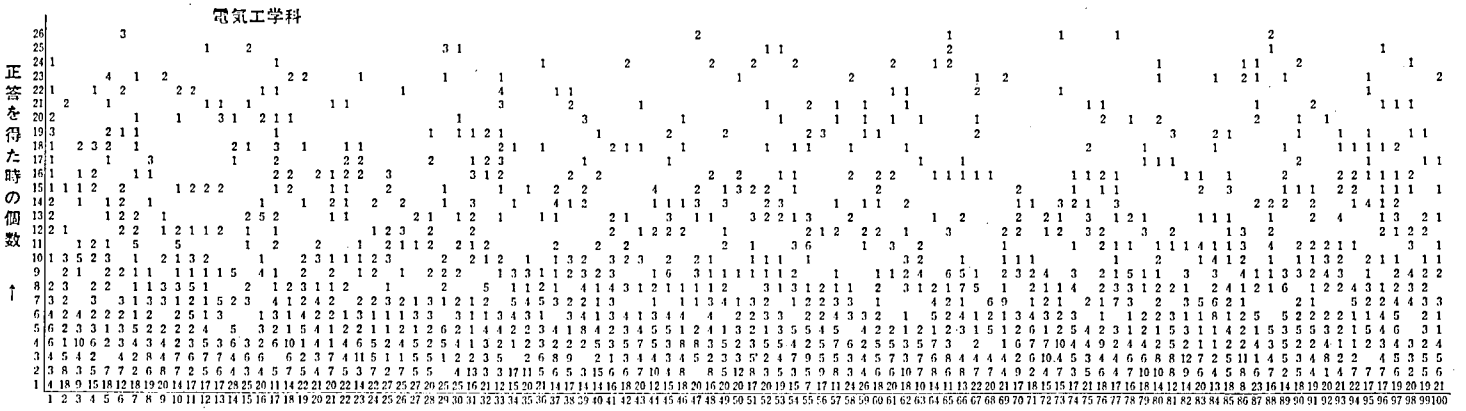

Fig 1A 電気工学科 度数表

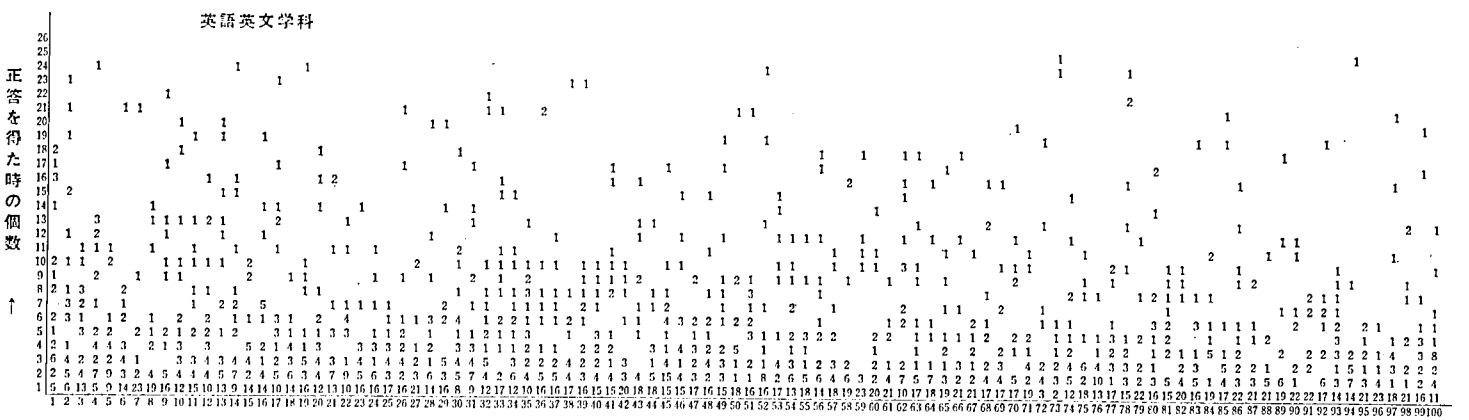

Fig 1B 英語英文学科 度数表

を熟知していれば当然次の字は容易に見出し得るであろう，又それらおよびその国語の遷移確率等も全く知らなければ乱字に等しくエントロピーは $\log n$（n は種類数）となる。文章構造，単語，成句，などの知識ということが語学能力の一つの尺度となっているのは当然であり，従ってこの観点より得た情報量が小さければ小さいほどその語学に関する学力は大きいことが言えよう。

§4. 実 験 結 果

実測は電気工学科4年生50名および3年生120，名英語英文学科30名，全て本学学生を対象に行なった。電気工学科学生を選んだのは筆者らの所属の関係で最も実験が行ない易かった故である。英語英文学科の学生を対象にした理由は英文に関してはこの学科の学生の方が語学力を有すると判断して，前項に述べた仮定を検定するためにほかならない。試料としては英字新聞の社説を主として用い固有名詞を含まない箇所を選んである。予測は第1字目より行わせ，100字まで実験を行なった。資料としてはコンサイス程度の辞書を持参させたが殆んど活用せずこの影響は無視して差支えない。

電気工学科50名および英語英文学科30名の実測の結果はそれぞれ Fig. 1 A，B として示したとおりである。

Fig. 1 より各々のエントロピーを求めたものを Fig. 2 として掲げた。図には参考のため米人の例[1]も併記してある。

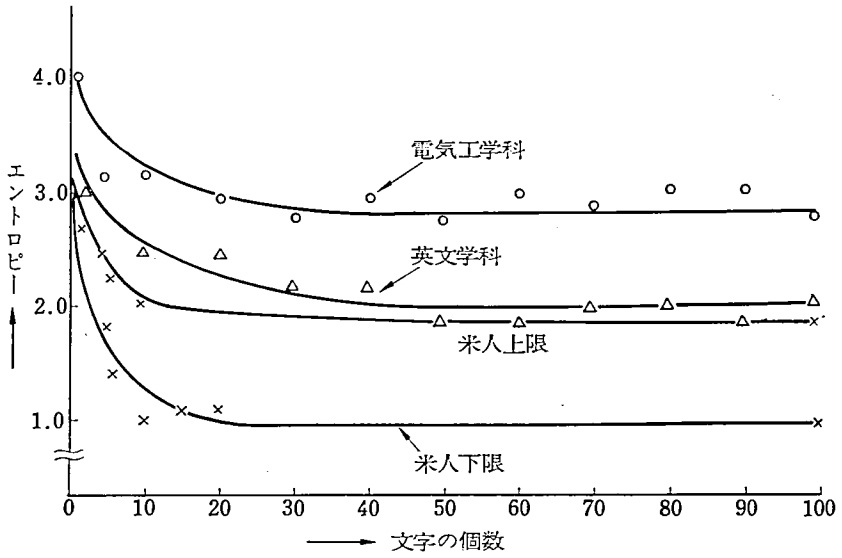

Fig. 2 エントロピー—文字の個数

Fig. 2 より英語の語学力が高いと思われるグループ程エントロピーは低下しており我々の仮定は満足させられたかと思える。

一方エントロピーの計算は家庭で行なうにはやや複雑過ぎるので，平均情報量 (I) を用いることを考えた。これはN字目に平均何ビットの情報を有するかと言うことで

$$I=\frac{1}{N}\sum_{i=1}^{26}m_i \log i=\frac{1}{N}\sum_{i=1}^{N}\log i \qquad \cdots\cdots\cdots(8)$$

で表わされる。ここに　N：試行回数$=\Sigma m_i$，m_i：i番目に当った回数である。

エントロピーは情報量の測度であるから直接情報量を用いて理論的に差支えないが，斯る有限個の資料につき現実問題としてどの程度相関関係を有するか，各学年毎に求めた。

その結果をFig. 3 A，Bとしてそれぞれ掲げた。

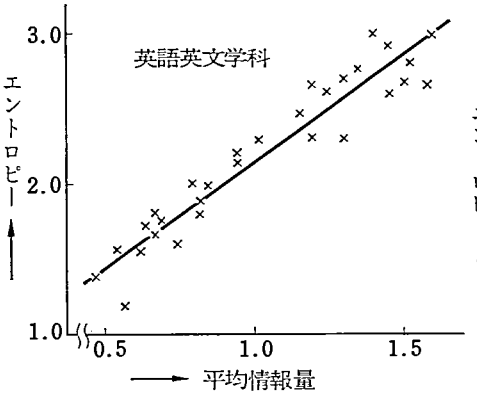

Fig. 3A　電気工学科の相関関係

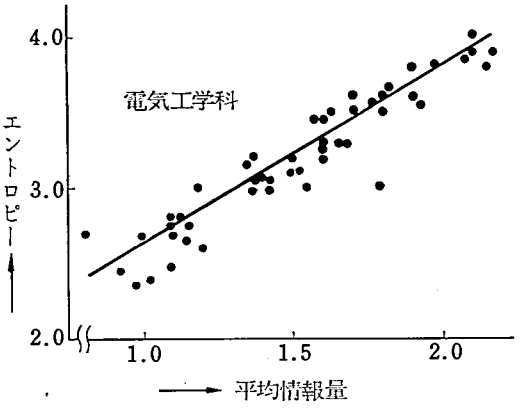

Fig. 3B　英語英文学科の相関関係

この場合のような少数例においても電気工学科の相関関係数は0.97，英語英文学科のそれは0.99となり非常に強い正の相関関係を有し，従って平均情報量を学力評価の測度として差支えない。そこで平均情報量をそれぞれ求め，Fig. 4 として掲げた。

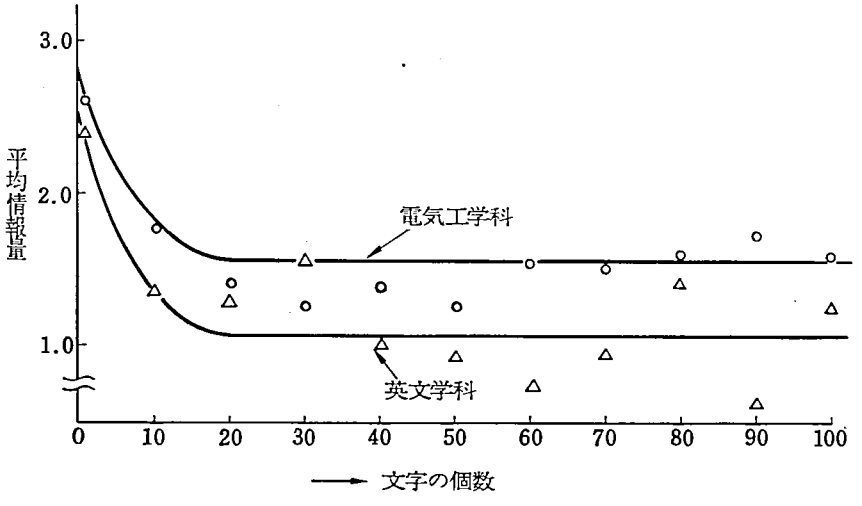

Fig. 4　平均情報量－文字の個数

実際に各家庭で実験を行うときは，Fig. 4 に見られるように10数字以後は最小自乗法を用いて計算した結果，電気工学科においては方程式　$y=0.081x+1.57$　英語英文学科においては　$y=-0.023x+107$　と勾配が殆んど零になるのでこの程度の字数を先行語として与えて，この後100文字程度の実験を行い平均情報量を求めるのが妥当であろう。更に電気工学科と英語英文学科の差が偶然によるものか否かを判定するために，偶然によると仮定を立てこれに対して独立性の検定を行った。計算に際して資料数が増大する程精度が向

上するので前述の定常状態になった50字以後のデータをまとめて使用した。

今電気工学科の実測結果の度数分布をf_{1i}，英語英文学科のそれをf_{2i}とし，それぞれの

理論値を f_{1i}^* f_{2i}^* とするとき $\sum_i f_{1i}=2500$ $\sum_i f_{2i}=1500$

総数 $n=\sum_i f_{1i}+\sum_i f_{2i}=4000$, $f_{1i}+f_{2i}=f_i$ ……(9)

とするとき

$$f_{1i}^*=\left(f_i\sum_i f_{1i}\right)\Big/n,\quad f_2^*=\left(f_i\sum_i f_{2i}\right)\Big/n \qquad \cdots\cdots(10)$$

である。

これらより，前述の仮説のもとでは

$$X^2=\sum_{j=1}^{2}\sum_{i=1}^{26}\left[\left(f_{ji}-f_{ji}^*\right)^2\Big/f_{ji}\right] \qquad \cdots\cdots(11)$$

は自由度 $\phi=\max i-1$ なる X^2 分布をなすと考えられる。従って危険率5％とした値よりも大であれば仮説は棄てられ即ち2学科の差は偶然ではないという結論が得られる。これらの関係および計算結果はまとめて Fig. 5 として示したとおりである。

i	実測値		実測値の和	理論値		$\frac{(f_{1i}-f_{1i}^*)}{f_{1i}^*}$	$\frac{(f_2-f_{2i}^*)^2}{f_{2i}^*}$
	電気 f_{1i}	英文 f_{2i}		電気 f_{1i}^*	英文 f_{2i}^*		
1	867	888	1755	1096.8	958.1	48.1	80.3
2	292	187	479	299.3	179.6	0.1	0.3
3	250	102	359	220.3	132.0	4.0	6.8
4	177	66	243	151.8	91.1	4.1	6.9
5	134	57	200	125.0	75.0	0.6	4.3
6	106	36	142	88.7	53.2	3.3	5.5
7	98	22	120	75.0	45.0	7.0	11.7
8	90	23	113	70.6	42.3	5.3	8.8
9	87	21	108	97.5	40.5	5.6	9.3
10	43	20	63	39.3	23.6	0.3	0.5
11	54	12	66	41.2	24.7	3.9	6.5
12	43	13	56	35.0	21.0	1.8	3.0
13	43	10	53	33.1	19.8	2.9	4.8
14	35	3	38	23.7	14.2	5.3	8.8
15	29	3	32	20.0	12.0	4.0	5.3
16	31	9	40	25.0	15.0	1.4	2.4
17	10	5	15	9.3	5.6	0.1	0.1
18	15	6	21	13.1	7.8	0.2	0.4
19	20	5	25	15.6	9.3	1.2	1.9
20	15	2	17	10.6	6.3	1.8	2.9
21	13	3	16	10.0	6.0	0.9	1.5
22	6	2	8	5.9	3.0	0.2	0.3
23	15	0	15	9.3	5.6	3.4	5.6
24	15	3	18	11.2	6.7	1.2	2.0
25	6	2	8	5.3	3.0	0.2	0.3
26	6	0	6	3.7	2.2	1.4	2.2
和	2500	1500	4000			108.3	182.4
	$X^2=X_1^2+X_2^2=108.3+132.4=290.7$					$=X_1^2$	$=X_2^2$

Fig. 5　独　立　性　の　検　定

図中にも示したように

$$X^2=290.7>X^2_{\substack{0.01\\ \phi=25}}=44.3 \qquad \cdots\cdots\cdots\cdots(12)$$

となって有意性が認められる。従ってこの2学科の差は偶然であるという仮説は1％の危険率をもって棄却される。換言すれば2学科の差は偶然ではなく，これより語学力の差と平均情報重—エントロピー—の間に関連を有することが立証されたといえよう。

§ 5. 結　　論

本実験において語学力の評価の一方法として平均情報量を用いることが可能でかつ実用的に有用であることが認められた。今後は英語学習初年度よりこの種の実験を進め，学力評価方法として実用化を計り順次他国語におよぼし，教育の一助とする所存である。

文　　献

(1) J. R. PIERCE: SYMBOLS, SIGNALS, AND NOISE HARPER & BROTHERS N. Y. 1961.

付　　記

本研究は本学電気工学科助手植村辰久君が中心として行なったものである。ここに記して同君に深謝すると共に，実験に協力した本学電気工学科1回生石川清助君，データ処理に尽力した本学第3学年田総元雄君および快く被実験者となった学生諸君に謝意を表わす次第である。