論 文

論文 74-11-3-3

韓國語 音節의 Entropy에 관한 硏究

(Statistical Measurement of Monosyllable Entropy for Korean Language)

李 柱 根* · 崔 興 文**

(Lee, Joo Keun *and* Choi, Hoong Moon)

要 約

이 論文은 韓國語의 3聲音의 組織을 方程式化하고 그로부터 組織的인 상태 graph를 유도하여 情報源의 性質을 구명하고 나아가서 기본 要素의 統計量에 대한 Entropy와 Redundancy를 測定하여 西歐語와 비교검토하였다. 特히 한국어에서 情報源의 性質을 究明하기 爲한 syllable의 상태 graph는 單一化된다는 것을 明示함으로서 他言語에서 볼 수 없는 特異한 현상이 나타난다는 것을 지적하였다.

Abstract

The information amount of monosyllables(characters) in Korean language is measured, in order of the following 3 steps.

1) The basic consonants and vowels are partitioned into two sets,
2) These set symbols, C and V, are sequentially combined to obtain the equation which represent the flow state of monosyllables.
3) From the equation, the state graphs can be constructed to examine the properties of a stochastic process of monosyllables in Korean language.

Furthermore, the entropy of Korean language by statistics is measured and compared with that of the western languages. The proposed methods are more definite, systematic, and simpler than the usual methods in examining the nature of information sources.

I. 序 論

人間의 言語行爲에 의하여 형성되는 自然語는 情報의 傳達 및 處理過程에서 가장 기본이 되는 情報源을 이룬다. 自然語에 대한 統計學的인 연구는 일찌기 Pushikin의 「Eugene Onegin」 중에서 2萬字에 해당하는 文章에 대하여 길이 3의 文字系列의 頻度를 조사한 것이 最初이다(1913).

오늘날의 情報理論은 C.E.Shannon에 의하여 그의 基礎가 確立되었으며, 그는 情報의 傳達에 관한 기본 性質과 確率過程에서의 情報源의 구조를 해명하고 情報量의 測度로서 Entropy를 도입하였다[1](1948).

그로부터 言語情報源에 대한 定量的인 연구는 활발히 전개되었다. 그러나, 言語란 나라마다 다르고, 또 그의 表記要素인 文字, 文章에따라 그의 기본 性質을 달리한다. C.E.Shannon은 Nonrooted data를 가지고 英語의 Word Entropy를 발표하였고[3](1951), G. A. Barnard는 Rooted data를 사용하여 Shannon 方法에 따라 西歐 4個國語 (영, 불, 독, 스페인어)의 Word Entropy를 비교하였다[4](1955). 그러나 韓國語에 대해서는 情報理論에 기초를 둔 情報量에 대한 연구는 아직 發表가 없다. 그런데, Shannon이나 기타 연구에서는 英語에 대한 情報源의 性質을 구명하기 위한 상태 Graph의 유도에 있어서 文字 하나 하나를 對象으로

*,** 正會員, 仁荷大學校 工科大學 電子工學科
Dept. of Electronic Engineering, Inha University
接受日字 : 1974年 5月 9日

하였기 때문에 英語에서 單語의 性質上 상태 Graph가 一定치 않다. 즉 선택되는 Sample에 따라 상태 Graph가 다르게 되어 組織的이 아니다. 이 硏究에서 提案된 方法은 個個의 기본 文字를 대상으로 하는 것이 아니고, 기본 子母를 두개의 集合成分으로한 30種의 單音節의 集合을 구성하고, 그의 論理變換에 의한 方程式을 유도 하였다. 그것으로 부터 組織的인 상태 Graph를 유도하고, 다시 音節 Space를 도입하여 單音節의 確率過程에서의 性質과 制限條件을 검토하여 한국어의 情報源에 대한 性質을 구명하였다. 나아가서 기본 子母音, 音節要素의 統計量의 分析(別紙 68개 Table)[10]에 의한 Entropy와 Redundancy를 測定하였다.

이 硏究의 결과는 한국어를 현대 情報工學에 적응시키기 위한 기초가 마련된 것으로서, 從來의 각종 端末장치의 再評價, Code割當의 理論的 根據, 字盤배열등 새로운 情報장치의 개발에 主要한 資料가 될 것으로 기대된다. 單語에 대한 分析은 韓國語에서는 天文學的인 계산을 필요로 한다. 이 硏究에서는 單音節에 제한하였으나, 그것은 英語에서 單語에 필등하므로 이 연구의 결과만으로도 充分한 資料가 된다고 생각된다.

Ⅱ. 韓國語의 情報源에 대한 考察

(1) 單音節의 形成過程

한국어를 표기하는 한글은 單音節文字이고, 그것이 또한 音에 따라 組合될 뿐만 아니라, 그들 表記要素는 初, 中, 終聲이란 3聲音의 明確한 조직규칙이 作用하며, 構造上으로는 集團變化를 한다는 點[7]에서 表記要素의 橫的나열만의 西歐語에 비하여 특별한 次元을 이룬다.

이것은 情報源의 性質의 구명에 있어서 새로운 결과를 초래할 것이란 點에 着目하여 그 着目點으로 부터 單音節의 組織過程을 분석하였다. 즉

기본要素의 集合을

$$X=\{x_i\}, \qquad i=1,2,\cdots\cdots 24 \tag{1}$$

라 했을 때 子音集合은

$$\begin{aligned}\mathcal{C}&=\{x_i|x\in X,\ x:\text{consonant}\}\\&=\{ㄱ,\ ㄴ,\ ㄷ,\ \cdots\cdots,\ ㅎ\}\end{aligned} \tag{2}$$

$$i=1,2,3\cdots\cdots,m,\quad m=14$$

또 母音集合은

$$\begin{aligned}V&=\{x_i|x\in X,\ x:\text{vowel}\}\\&=\{ㅏ,\ ㅑ,\ ㅓ,\ ㅕ,\ \cdots\cdots,\ ㅣ\}\end{aligned} \tag{3}$$

$$i=1,2,3,\cdots\cdots n,\quad n=10$$

이되며, 이들 集合要素 $\mathcal{C},\overline{V}$로서 3聲音을 구성하는 形式은

$$\mathcal{C}_x=[C_i, C_iC_j]$$

$$V_y=[V_i,\ V_j,\ V_iV_j,\ V_jV_k,\ V_iV_jV_k]$$

$$\mathcal{C}_z=[\phi, C_k,\ C_k\mathcal{C}_k',] \tag{4}$$

$$C_i,C_j\subset\mathcal{C},\ (V_i\cup V_j)\subset V, V_k\subset V_j$$

이들 式에서 $\mathcal{C}_x$,는 初聲, V_y는 中聲, $\mathcal{C}_z$는 終聲音의 集合要素을 表示하고, 添字 i,j,k는 文字에서 位置別을 표시 한다. 또 이들 3聲音의 形式은

$$f(\mathcal{C},V)=\mathcal{C}_x\cdot V_y\cdot\mathcal{C}_z \tag{5}$$

으로서 표시된다.

즉 이들 $\mathcal{C}_x$, V_y, $\mathcal{C}_z$가 單音節을 구성하는 形態는 30種類가 된다.

$$N(\mathcal{C}_x)\times N(V_y)\times N(\mathcal{C}_z)=2\times5\times3=30 \tag{6}$$

이들 30種의 form化는 著者의 一人이 이미 指摘한바 있으며[7] 이와 關聯된 연구에서 그들 form으로 부터 세가지의 着想點을 유도한바 있다.[6][7]

첫째는 Matrix 組織에 의한 文字의 組織的 配列에 대한 硏究였다.[6]

둘째는 기본 要素의 入力으로서, 合理的인 文字를 發生할 수 있는 Character Generator의 구성을 可能케한 바 있다.[7]

세째는 30種의 form으로 부터 또 다시 6개의 form으로 規格化하므로써, pattern 空間의 部分分離의 可能性을 주어 效果的인 文字 Pattern의 認識을 可能케 한바 있다.[7]

네째로 이 硏究에서는 그들 form (또는 單音節)으로부터 또 다시 確率過程에서의 性質을 규명하기 위한 單音節의 組織的인 상태 Graph를 유도하였다. 즉 (4)(5)式 또는 30種의 文字 form의 論理變換에 의하여 다음 (7)式이 유도된다.

$$\begin{aligned}f(\mathcal{C}\overline{V})&=(\mathcal{C}+\mathcal{C}\mathcal{C})(V+VV+VVV)(1+\mathcal{C}+\mathcal{C}\mathcal{C})\\&=(\mathcal{C}_i+\mathcal{C}_i\mathcal{C}_j)(V_i+V_j+V_iV_j+V_jV_k\\&\quad+V_iV_jV_k)(1+C_k+C_kC_k')\end{aligned} \tag{7}$$

但 "1"은 identity를 定義하며, $a\cdot1=a$로 부터 $C_iV_j\cdot1=$"가"형을 定義한다. (7)式은 한국어의 모든 單音節의 상태를 표시하는 동시에 한글文字의 모든 형태를 표현한다. 添字 i,j,k는 順序와 位置別로 표시한다. (7)式으로서 Flow graph를 작성하고 또 거기에 다시 單音節의 Space S를 도입하고[3], 音節의 상태를 J_i로서 표시하면 그림 1과 같은 Syllable의 상태를 표현하는 Garph가 유도 된다. 그림 1에서 상태 $J_1\sim J_6$은 音節구성이 完成되지 않은 過渡 junction이며, J_7은 終聲이 없는 音節의 상태, J_8은 終聲音이 한개가 存在하는 상태이고, J_9은 終聲音이 두개로서 구성되는 音節의 상태를 표시한다. 이 Graph는 情報源에서 방출되는 기본 子母音이 音節을 형성하는 確率過程에서의 Flow 상태를 보여준다. 즉 J_0 點에서 시작하여 상태 J_7에서 J_0로 돌아오면, 한 音節이 完成되고, 그 J_0點에서 다시 다른 音節이 시작된다. 이것을 계속 반복

허 가면 한국어의 單音節뿐만 아니라 모든 文章을 표현 할 수 있다는 것을 보이고 있다. 이것은 대단히 흥미 있는 것으로서, 從來의 ABC……한字 한字를 對象으로 한 상태 Graph[3]로서의 검토와는 本質的으로 次元을 다르게 한 것이다.

이 論文에서는, 우리 言語에는 3聲音이란 特殊조직이 存在한다는 點에 着目하여 그것을 2段集合過程을 거쳐 30종의 상태로 규격화 함으로써, 音節을 集團決定짓는 單一 상태 Graph로서, 音節群을 處理한 點이 이 硏究의 主要 Point의 하나이며 이點은 西歐의 어느 言語에서도 單一 Graph로서 言語의 모든 情報를 나타낼수 없는데 反해서 特異한 點이란 것을 지적한다.

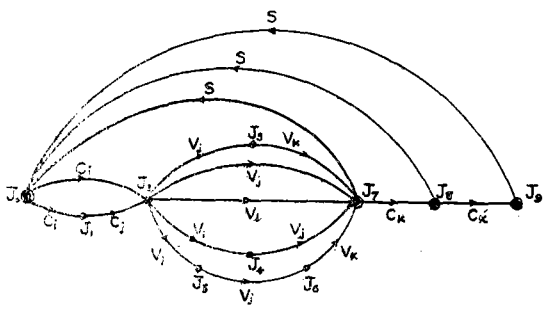

Fig. 1. A graph corresponding to the mono-syllable forming process.

또 이것은 다음 節에서 情報源의 性質을 判定짓는데 중요한 Factor를 제공한다.

(2) 音節組織의 制限條件

그림 1의 Graph에서 音節을 구성하는 制限條件은 다음 5개 項으로 요약된다.

i) 音節은 初聲과 中聲의 最小限 各 한개 要素를 포함한 2~7개 要素로서 구성된다.

ii) 初聲은 1~2개의 기본子音으로 구성되나, 2개로서 구성될때 (C_iC_j)는 반드시 $i=j$인 2重子音형태를 취한다.

iii) 中聲은 1~3개의 기본 母音으로 구성되나, 2個以上의 要素를 취할때는 $i\neq j\neq k$의 복합 母音의 형태만을 취할 수 있다.

iv) 終聲은 0~2개의 기본 子音으로 구성되나, $K=K'$의 2重子音, $K\neq K'$의 복합 子音의 형태를 취한다. (여기서 C_iC_j와 C_kC_k'의 구분은 그들의 順序位置를 표시하기 위함이다).

v) 2개 以上의 母音이 연속될때, 뒤에 오는 母音은 반드시 앞 母音에 의하여 결정된다. 즉 母音 "ㅗ"는 "ㅏ" ("ㅛ"는 "ㅑ"와)만이 결합되고, "ㅜ"는 "ㅓ"와 ("ㅠ"는 "ㅕ"와)만이 결합되어, 이것은 극히 중요한 문제로서 Markov 情報源으로 이끌어 진다. (母音"ㅣ"는 아무 제한도 받지 않는다).

(3) 情報源의 性質

確率的인 情報源에서는 일반적으로 정보원의 각 要素의 出現確率이 앞서 출현한 요소의 영향을 받는 Markov Source로서 記述한다. 즉 임의의 정보원에서 발생한 文字의 時間系列을

$$x_{i-m-n},\cdots\cdots\ x_{i-m},\ x_{i-m+1},\ \cdots\cdots,\ x_{i-2},\ x_{i-1}$$

라 할때 다음 순간 x_i라는 文字가 발생할 條件付確率이 다음 관계를 가질때, 일반적으로로 M^{th} order Markov Source로 알려져 있다.

$$P(x_i/x_{i-1},\ x_{i-2},\cdots\cdots\ x_{i-m},\cdots\cdots,x_{i-m-n})$$
$$=P(x_i/x_{i-1},\ x_{i-2},\cdots\cdots,x_{i-m}) \qquad (8)$$

$m=0$일때 각 文字의 出現確率이 相互獨立인 Zero-Memory Source가 된다.

이러한 定義를 한국어의 단절음에 비추어보면 Markov Source에 속함이 명백하다. 또 앞 節에서는 한국어에서 單音節의 集合에 대한 性質을 구명하였고 여기서는 그들 內部 상태를 관측한다. 즉 前節에서 지적된 制限條件에서 例를 들면, "ㅗ"의 뒤에 다른 母音이 연속될때 "ㅏ"아니면 "ㅣ"가 반드시 온다. 또 "ㅜ"일때는 "ㅓ""ㅣ"이다. 즉 뒤에올 母音은 반드시 앞 母音의 영향을 받는다. 이것은 모든 한국어의 單音節에서 일어나는 현상으로서 한 자모가 다른 자모와 결합하여 音이 發生 하기 때문에 Markov Source의 性質을 完全無缺하게 나타낸다. 이것은 西歐語에서 大數의 法則이 成立할때에 限해서만 Markov Process中의 Ergrodic Process가 되는 것과는 본질적으로 다른 현상이다.

한편, 그림 1에 제시한 音節의 상태 Graph를 Shannon의 判定法에 비추어 보면

(i) Graph 內에서 서로 隔離된 두개의 junction이 없고

(ii) 화살표를 따라 原點 J_0를 포함한 各 loop의 길이의 G.C.D가 1이다. (단위 길이는 한개의 화살표를 가진 각 선분) 즉, 分解不能이고, 非周期的인 Graph라는 點에서 한국어의 情報源은 Markov Process는 물론 Ergodic Process임이 明白하다. 이것은 音節集合으로서 표시되거나, 從來의 方法과 같이 基本要素하나 하나로서 表現될 때도 마찬가지이다.

이상에서 한국어의 情報量을 결정지을 수 있는 根據를 제시하였으며, 計算에 필요한 統計方法 및 統計量을 결정할 수 있다. 그런데, Ergodic Process의 性質에 따라 임의의 文字系列의 평균과 系列集合의 평균과는 같으므로 어느 특정한 無限系列에서 文字의 相對頻度를 구하는 것이나, 系列의 集合에서 相對頻度를 구하는 것이나 같은 결과를 얻을 수 있다.

Ⅱ. 統計資料

(1) 統計의 節次

文字情報源은 不確定系에 속하므로 統計學的인 確率 Model로서 그의 性質을 취급할 수 있으며, 각 要素의 確率的인 性質을 알아야 情報量의 계산이 可能하다. 따라서 한국어의 自然語에서 기본 子母音, 單音節, 單語 등의 出現確率을 測定해야 하며, 나아가서 N-gram의 條件付 및 同時確率을 구하여야 한다.

한국어의 어휘頻度조사는 文敎部에서 발표한바 있으나[6], 현대語에서 사용하지 않는 표기(ㅱ, ㅳ, ㆁ, ㅶ) 등이 포함돼 있는 반면 音節區分, 單語 Space가 統計에 고려돼 있지 않으므로 Markov Source에 따른 多目的分析이 곤란하다. 따라서, 이 연구에서는 다음과 같은 條件과 節次에 따라 統計資料를 새로 마련하고, 計算機에 의하여 多角的인 分析을 하였다.

(i) 현대語로 表記되고, 맞춤法과 띄어쓰기가 正確한 31種의 書籍에서 28萬字에 해당되는 文章에서 資料를 채취하였다.

(ii) 특정 分野나 階層에 속하는 文章에 치우침으로 因하여 생기는 統計의 歪曲을 막기 위하여 가능한 여러 종류의 書籍 (교과서, 신문, 잡지, 소설 등)을 자료로 삼았다.

(iii) 外國語표현이라도 표준 한글로 表記되어 있을 때는 統計資料로 삼았다.

(iv) 數字등은 除外하고 音節區分과 單語 Space는 統計에 넣었다.

(v) 以上의 資料는 풀어 쓰기로 하고, alphabet로 代置 Coding하여, Data card를 作成하였다. 이렇게 준비된 資料로서, Ergodic Process 및 定常情報源으로서의 性質을 관측하기 위하여 처음 2萬字까지는 千字 단위로 區分하여, 順次로 상태 變化를 관측하고, 다음은 2萬字를 單位로 하여 統計量의 變化에 따른 기본 要素의 確率分布를 조사하였다. 그 結果, 그림 2와 같이 10萬字 以內에서 거의 一定한 安定상태에 도달하였다. 이것은 한국어에서 統計量의 범위를 얼마로 할 것인가 하는 문제에 중요한 資料가 된다.

그러나 單語의 確率分布를 결정짓는 데는 더욱 많은 資料를 필요로 하므로, 約 3倍에 해당하는 28萬字를 統計量으로 設定 (이것은 결과적으로 다음 (11)式에 보인 한文字의 平均要素의 數의 근거와 일치한다)하고 基本子母音, 重複合子母音, 音節의 出現確率을 구하였다.[10]

(2) 기본 子母音의 頻度와 確率分布

24개의 기본 子母音의 確率分布를 頻度順에 따라 나열한 것을 그림 3에 표시 하였다(附錄 table—1). 그림 3의 確率分布에 대한 近似式은 多項式으로 展開되며, 點線上의 i番째要素의 確率 P_i는

$$P_i=0.1403-0.1933\times10^{-1}i+0.164\times10^{-2}i^2 -0.797\times10^{-4}i^3+0.145\times10^{-5}i^4 \quad (9)$$

로 접근 된다.

한편, 기본子母音의 集合 C와 V의 確率은 다음과 같이 계산된다.

$$P(C)=\sum_{i=1}^{14} P(C_i)=0.5653$$

$$P(V)=\sum_{i=1}^{10} P(V_i)=0.433 \quad (10)$$

또 기본 子母音의 總頻度數는 741,630字이고, 全體統計量이 28萬字이므로 한 文字 (또는 음절)를 구성하

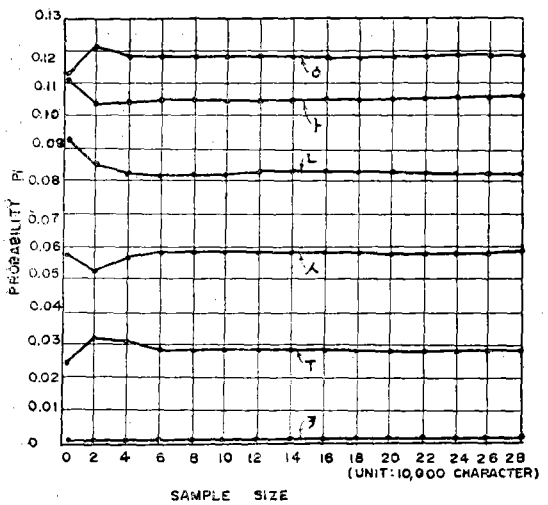

Fig. 2. The probability distribution according to the sample amount.

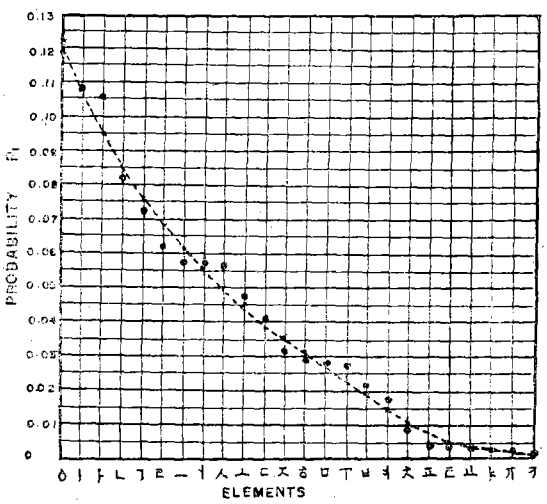

Fig. 3. Probability distribution of basic elements

는 平均 기본 要素의 數 β는

$$\beta=\frac{741,630}{280,000}=2.64 \text{ (Letters)} \qquad (11)$$

이된다. 따라서 한글 文字에서 基本 要素가 平均 3개 골이란 것을 明白히 하였다. 또 기본 要素의 頻度와 確率分布의 조사는 비단 情報量을 計算하는데 필요할 뿐만 아니라, 情處報理장치에서 速度向上을 위한 字盤配列에도 직접적인 資料가 된다.

(3) 音節의 頻度와 確率分布

音節의 頻度는 初聲 19, 中聲 21, 終聲 28개의 要素로 구분하고 각 要素를 기준으로 하여 合計 68개의 Table을 別紙附錄[10]에 수록했으며, 音節의 確率은 初聲 19개의 각 要素를 기준으로 統計를 취하였다. 이들의 分析結果, 全體統計量 28萬에서 사용頻度가 零이 아닌 文字의 總數는

1,603字로 밝혀졌다.

한국어에서 日常 生活에 使用되는 文字의 수는 얼마나 하는 問題는 매우 회의적인 것이었으나, 이로써 그 數가 明白해졌다. 물론 統計的인 결과이므로 엄밀한 숫자라 할수는 없겠으나, 가장 實際에 접근되는 方法에 의한 결과라는 點에서 큰 差가 없을 것으로 본다. 특수 表記를 포함할 때는 약 2000字內外가 될 것으로 推定된다.

또 音節要素 즉 3聲音의 確率은 Table—I에 주었다.

Table—I The probability of each syllable components.

Syllable Component (C_x)	Syllable Component (V_y)	Syllable Component (C_z)
$P(C_i)=0.9770$	$P(V_i)=0.3098$	$P(\phi)=0.5544$
$P(C_iC_j)=0.0229$	$P(V_j)=0.5414$	$P(C_k)=0.4141$
	$P(V_iV_j)=0.0602$	$P(C_kC_k')=0.0312$
	$P(V_jV_k)=0.0878$	
	$P(V_iV_jV_k)=0.0006$	
0.9999	0.9998	0.9997

單音節(文字)의 確率分布를 頻度順位에 따라 全對數 Graph에 표시한 것이 그림 4이다. 그런데 英語에서는 그림 4의 對角點線上에서 單調減少函數로 나타나므로 頻度順位 n番째 單語의 確率 P_n는 Zipf의 法則에 따라 $P_n=\frac{k}{n}=\frac{0.1}{n}$이라는 간단한 近似式으로 표시된다. 그러나 한국어의 音節의 確率分布는 그림 1과 같이 $\frac{1}{4}$圓周上에 分布되어 Zipf의 法則이 成立하지 않음이 밝혀졌다. 이것은 무엇인가, 그로 因한 特異點이 무엇인가 하는 문제가 提起된다. 이 問題는

(i) 한 音節의 구성要素가 英語의 單語에 비하여 그 數가 적고, 子音과 母音의 出現確率은 그림 3에서와 같이 單調減少函數로 나타나지만 音節구성에 있어서는 出現確率이 높은 子母音끼리 結合될 때와 確率이 가장 낮은 子母音끼리 結合될 때가 있을 것이다. 따라서 극히 높은 確率分布와 극히 낮은 確率分布로 偏重되어 나타날 것이다.

(ii) 또 그 中間部分에서는 確率이 높은 子音과 母音의 結合, 또는 그 反對로 因한 音節의 형성過程에서 確率分布가 均等해 진다는데 起因된다고 보겠다.

(iii) 또 기본 子母音이 한 文字에서 反復結合되기 때문에 文字의 出現確率이 偏重되며 均等分布 된다고 보겠다.

以上은 直觀的인 관찰이지만 더욱 면밀한 검토의 여지가 있다고 보아 따로 미룬다. 音節의 確率分布가 半圓의 등고선上에 나타난다는 것은 西歐語와는 判異한 現象이다.

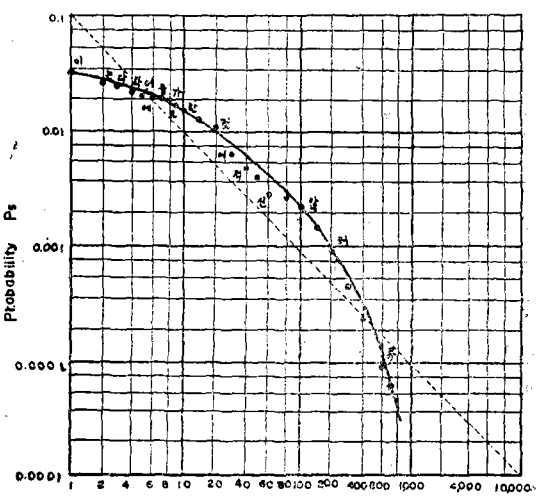

Fig. 4. Probability v.s. frequency rank of Korean monosyllables.

Ⅳ. 韓國語의 情報量의 測定

앞에서 記述한 諸硏究에서 한국어의 情報源의 성질은 Markov, Ergodic Process 임이 명백하고 또 各 統計値가 算出되었으므로[10] 情報量이 결정지어진다.

(1) Entropy와 Redundancy[13][14]

一般的으로 Entropy는 情報量의 測度로서 統計熱力學에서의 Entropy에 대응되는 것이다. 言語情報源의 경우 Entropy는 個個의 文字가 갖는 情報量의 平均값을 나타내는 統計學的인 parameter로서 각 文字의

平均 digit數를 뜻하므로 能率的인 code化의 기초가 되는 것이다. 한편 Redundancy는 相對 Entropy가 1보다 얼마나 적은 가를 나타내는 量으로서 情報源의 不要部分을 말하며, 또 이것은 情報源의 統計學的 構造로 인하여 言語에 부여되는 制限量을 표시하는 값이기도 하다.

한국어에서 各種 Entropy는 다음과 같이 구하여 진다

(a) 最大 Entropy는 ($P_1=P_2=\cdots=P_n$)

$$F_0=H_m=\log_2 n=4.58\text{〔Bits/Letter〕} \quad (12)$$

(b) Zero Memory Source Entropy는 부록 Table—Ⅰ로부터

$$F_1=H_0=-\sum_{i=1}^{24} P_i\log_2 P_i=4,029 \quad (13)$$

(c) Syllable Entropy F_s는

$$F_s=\frac{1}{\beta}\left[-\sum_{i=1}^{N} P_s\log_2 P_s\right]=2.92\text{〔Bits/Letter〕} \quad (14)$$

β: 음절을 구성하는 평균 자모의 수, $\beta=2.64$

P_s: 각 음절의 확률 〔별지 부록참조[10]〕

N: 통계상의 문자의 총수, $N=1,603$자

(d) Redundancy

$$\Gamma=1-\frac{F_s}{F_0}=0.36 \quad (15)$$

그런데 한국어에서 음절의 구성요소를 24개의 基本 子母로 구분하는 것이 정당한 것이지만 typewriter 등에서 器具의 구조상 필연적으로 기본 要素를 증가 시킨다. 이러한 경우의 Entropy는 (51개 요소를 기준)

$$F_1^*=-\sum_{i=1}^{51} P_i\log_2 P_i=4,405\text{〔Bits/Letter〕} \quad (16)$$

가 되어 F_1과 F_1^*를 비교하면 24개 기본 要素만의 경우는 (13)式으로 부터 Entropy가 적어서 Code 割當등에 유리하다.

以上의 結果를 Talbe-Ⅱ에 西歐 4個國語의 Entropy와 비교하였다.

Table—Ⅱ. Entropies of Korean and four Western languages.

	Korean	English	French	German	Spanish
A	24	26	26	26	26
α, β	2.64	4.5	4.84	5.92	4.96
F_0	4.58	4.70	4.70	4.70	4.70
F_1	4.029	4.124	3.984	4.095	4.015
F_s	2.92	—	—	—	—
F_w		2.62* 1.648**	3.02	1.08	1.97

* : Shannon의 Nonrooted data에 의한 결과

** : Barnard의 rooted data에 의한 결과

α : 西歐語의 單語를 구성하는 평균 문자 數

β : 한국어 單音節을 구성하는 평균 문자 數

西歐 4個國語의 Word Entropy와 單語의 平均長과의 관계를 보면, 대체로 單語長이 긴 言語일수록 Entropy가 적은 것을 알 수 있다. 대표적인 例로 獨語의 경우는 平均長 α가 가장 큰 반면 word Entropy는 가장 적다. 그러나 佛語의 경우는 가장 큰 Entropy를 나타냈다.

한국어에서 Syllable Entropy(文字 entropy)가 2.92bit로서 佛語에서의 Word Entropy 3.02bit에 대등한다. 이것은 Entropy의 정리에 비추어 볼때 佛語에서는 한 單語가 다른 西歐語에 비해 표시內容이 풍부하다는 것을 뜻한다. 다시 말하면 한국어에서는 文字가 그들 單語보다도 充分한 內容표시를 할 수 있다는 科學的인 立證이기도 하다. 또 한국어의 文字가 佛語의 한 單語에 대등하고 또 單語는 한 個의 單音節이 모여 이루어 짐으로 平均長 $\beta=2.64$란 點에서 볼때 한국어의 word Entropy는 더욱 더 적어질 것이 명백한 사실이며, 큰 Redundancy를 나타낼 것이다.

Ⅴ. 總括 및 結論

(1) 文字情報源에 대해서는 Shannon등 많은 研究가 있으나 西歐語의 單語의 性質上 文字 하나 하나를 對象으로 하기 때문에 情報源의 性質을 研究하기 위한 상태 Graph가 組織的이 아니다. 이 研究에서는 單音節組織을 方程式化함으로써 體系的이고 간결한 상태 Graph를 유도 할 수 있었다.

世界에는 250種의 文字가 있다고하나, 文字 및 音節組織이 單一 Graph化 되는 예는 없는것 같다.

(2) 또 한국어의 情報源에 대한 性質을 명확히 밝혔고 Redundancy에 영향을 주는 制限條件을 제시하였다.

(3) 28萬字의 統計量에 대한 頻度, 確率 및 情報量등 68개 Table을 別紙에 수록하여 관련된 研究에서 資料가 될 것으로 기대된다.

(4) 기본 子母音의 적절한 統計量은 10萬字 범위 內임을 밝혔고, 日常生活에서 사용되는 현대 文字의 數는 1603字임을 밝혔다.

(5) 音節의 確率分布는 西歐語와는 달리 $\frac{1}{4}$圓周上에 分布되어 Zipf法則은 成立하지 않는다는 사실을 제시하였다.

(6) 한국어의 Syllable Entropy는 2.92〔Bits/Letter〕로 나타나서, 佛語의 Word Entropy에 대응한다. 이것은 文字의 特異性에서 오는 것으로서 지극히 注目된

다. 한편 英語의 Redundancy는 50%로 알려졌는데 反해서 한국어에서는 單音節當 (文字當) 36%의 Redundancy를 가진다. 따라서 單語는 1個以上의 音節이 모여서 구성되므로 단음절(한文字) 平均長이 $\beta=2.64$이란 點에서 Word Entropy는 더욱 적어질 것이며 message에서 많은 部分이 탈락되어도 (영어의 수개 단어가 탈락 되는 것에 대응) 充分히 解讀할 수 있다는 理論的인 根據를 제시하였다. 이 硏究의 結果는 情報工學의 硏究에서 多方面에 利用될 것으로 기대된다.

끝으로 이 硏究는 仁荷大學校의 支援과 1973年度 蓮庵文化財團 硏究費에 의하여 이루어졌음을 밝힌다. 또 data 分析에 있어서 朴 집氏의 노고에 대해서 感謝한다.

參 考 文 獻

1. C. E. Shannon: Mathematical Theory of Communication. Bell System Tech. J. Vol. 27. pp. 379~423 July 1948
2. Fano, R.M: Transmission of Information. MIT. Res. Lab. Electron. Report, 65. 1949.
3. C. E. Shannon: Prediction and Entropy of Printed English. Bell System Tech J. Vol. 29 pp. 147—160 1951.
4. G. A. Barnard Ⅲ: Statistical Calculation of word Entropies for Four Western Languages. I.E.E.E. Trans. On Information Theory. Vol. IT—1. pp. 49—53. 1955.
5. 文敎部 : 우리말에 쓰인 글자의 頻度調査 1956. 6月, 12月.
6. 李柱根 : 한글 文字의 認識에 관한 硏究(Ⅳ) 電子工學會誌 Vol. 9 pp. 197~204. 1972. 9.
7. Joo Keun Lee; Recognition and Display of Korean Characters. Ph.D, dissertation in Keio-university 1972. 12.
8. L. Brilloin; Information and Entropy, Ⅰ, Ⅱ. J. Appl. Phys. Vol. 22 pp. 334~337, pp. 338~343. 1951.
9. 李柱根, 崔興文 : 韓國語의 情報量에 관한 硏究. 電子工學會 秋季學術會 論文集 1973. 10.
10. 崔興文 : 韓國語의 Entropy에 관한 硏究. 仁荷大學校 大學院碩士論文集 Vol. 12. 1973. 12.

Appendix. 1. The probability and the information amount of the basic elements.

Elements		Probability(P_i)	$-P_i\log P_i$
Consonants	ㅇ	0.1191	0.3656
	ㄴ	0.0819	0.2958
	ㄱ	0.0734	0.2767
	ㄹ	0.0615	0.2476
	ㅅ	0.0576	0.2373
	ㄷ	0.0405	0.1874
	ㅈ	0.0317	0.1580
	ㅎ	0.0299	0.1516
	ㅁ	0.0291	0.1487
	ㅂ	0.0226	0.1237
	ㅊ	0.0076	0.0539
	ㅍ	0.0047	0.0365
	ㅌ	0.0045	0.0354
	ㅋ	0.0012	0.0123
Vowels	ㅣ	0.1074	0.3457
	ㅏ	0.1067	0.3446
	ㅡ	0.0583	0.2392
	ㅓ	0.0583	0.2391
	ㅗ	0.0477	0.2096
	ㅜ	0.0288	0.1476
	ㅕ	0.0187	0.1077
	ㅛ	0.0031	0.0258
	ㅑ	0.0025	0.0222
	ㅠ	0.0018	0.0167
Total		0.9986	4.0287