論文 80-17-2-4

空白素를 包含한 한글 字素發生 確率과 엔트로피

(The Probabilities and Entropies of HANGEUL Elements Including the Space)

安 秀 桔* , 安 之 煥**

(ANN, Souguil and AHN, Jeehwan)

要 約

C. E. Shannon에 의하여 情報理論의 基礎가 確立된 以來 各國言語에 對한 많은 硏究가 發表되고 있다.

마찬가지로 한글 element(要素)를 情報工學에 適用하기 위해서는 한글 情報源을 구성하는 모든 element(要素)에 對한 確率이 적용되어야 한다.

본 論文에서는 그간 누락된 space(空白素)가 包含되었을 때의 엔트로피(entropy)와 리던던시(redundancy)를 고려하였다. 따라서 본 결과는 旣存 各種 한글 情報處理裝置의 再評價 및 새로운 情報裝置의 코-드(code)化에 있어서 主要한 資料로서 사용될 것이다.

Abstract

The foundation of information theory mainly established by C. E. Shannon, opened the way to profuse studies on the linguistics of various languages.

The statistics of Korean letters composed of elements and space are analyzed. The Korean alphabetic elements including the space have their probabilities. And information techniques can be applied to them.

This paper extended the previous statistics of Korean letters to include the space. We have also calculated the entropies and redundancy for the case.

The result can be used to modify the evaluation of Korean letter information processing devices and used as the basis for letter coding method for eventual future information processors.

I. 序 論

人間의 意思疏通의 基本的인 媒介體인 言語는 情報交換의 一次的인 情報源으로서 이 情報源에 對한 定量的인 硏究는 科學文明의 利器를 使用하여 言語情報源을 다루게 되면서 始作되었으며 C. E. Shannon에 의하여 오늘날의 情報理論의 基礎가 確立된 以來 各國言語에 對하여 많은 硏究가 發表된 바 있다.

한글에 있어서 字素[1],[2]와 單音節[3] 낱말[2]에 對한 頻度分布가 發表된 바 있으나 한글의 code化 및 字盤配列 등에 있어서 한글 字素만으로는 그 特性을 充分히 把握할 수 없으며 單語와 單語, 文章과 文章을 分離시켜주는 space(空白素)를 한글의 字素와 對等하게 取扱할 必要가 있다.

이에 對한 이번 硏究는 韓國語에 space (空白素) 를 고려하였을 때의 엔트로피(Entropy)[4]와 리던던시(Redundancy)[4]를 計算하여 情報理論을 韓國語에 適用시키기 위한 기초로서 情報源의 code化[4] 및 한

*正會員, **準會員, 서울大學校 工科大學 電子科 (Dept. of Electronics Engineering, Seoul National Univ.), 接受日字 : 1979年 12月 17日

글타자기를 包含한 既存 各種 한글 情報處理裝置의 再評價 새로운 情報裝置의 開發에 主要한 資料로서 字素頻度, 文字頻度 및 單語頻度에 space(空白素)를 包含시키는 것에 主眼點을 두고 있다.

Ⅱ. 情報理論의 概念導入

言語는 人間의 생각을 他人에게 傳達하기 위한 媒介體이며 어떤 符號들의 連續的인 羅列로서 形成된 코-드(code) 群으로 볼 수 있다.

言語는 構成의 特性上 random process[5]가 되며 이에 對한 해석은 自然히 統計的인 確率分布에 따르게 되는데 이를 해석하는 方法에는 全體的 特性이 母集團의 一部와 a priori(事前的)[5],[6]한 것으로 부터 完全히 나타낼 수 있는 structure된 processes(組織된 過程)[6]와 母集團의 한 斷面만으로써 그 特性을 把握하는 partial characterization(部分的 特性化)[6] nonstructured processes(非組織過程)[6]이 있다.

이러한 言語의 定量的인 硏究는 情報理論에 바탕을 두고 있으므로 情報理論의 槪念을 우선 說明하겠다.

1. 言語 情報源의 性質

한글에서는 Markov source[5]의 性質을 뚜렷하게 나타내고 있다. 그 例로서 文字의 구성이 子音(1혹은 2字素) + 母音이거나 여기에 받침이 1, 2 子音 字素가 첨가되는 形態이며 이 規則을 벗어나지는 않는다. 重母音이 쓰일 경우에도 'ㅡ' + 'ㅏ', 혹은 'ㅣ', 'ㅜ' + 'ㅓ' 혹은 'ㅣ'가 반드시 오게 된다.

즉 뒤에 올 文字는 앞 文字의 影響을 받고 있으며 이것은 한글의 主要한 特性으로 Markov source[5]의 性質을 완벽하게 나타내 주고 있다.

어떤 source(情報源)로 부터 sample을 取하여 統計資料를 얻었을 때 이 統計資料가 情報源의 性質을 그대로 나타내기 위해서는 어떤 임의의 情報源으로 부터 取해진 sequence(數列)의 特性이 sample로 부터 取해진 ensemble[7]과 같은 ergodic process[7]이어야 한다. 그러나 이 점에 있어서는 統計資料가 커질수록 情報源의 性質을 나타내게 됨으로 markov processes[4]이면서 ergodic processes[7]가 된다. 따라서 統計에 의해서 나타난 特性은 情報源의 特性을 그대로 나타내 주게 된다.

2. Mandelbrot의 낱말 分布모델[8]

낱말 分布에 對하여 G.K.Zipf가 統計的으로 제시한 $P_n = \frac{k}{n}$의 形態가 있으나 Mandelbrot는 낱말을 sequencial coding으로 생각하여 수식으로 전개하여 낱말 分布가 $P_n = P(N+B)^{-r}$의 形態를 가질 것이라고 예측하였다. 실제 西歐語에 있어서 統計的인 結果와 一致하고 있음을 나타냈다.

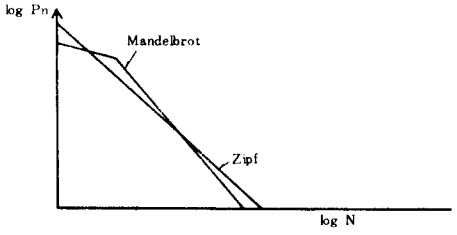

그림 1. 單語의 發生確率순위 N對 發生確率 P_N 그래프

Fig. 1. The log-log plot of word probability against word freq. order.

Zipf의 모델은 Mandelbrot 모델(그림 1)에서 B = r = 0인 특별한 경우에 해당된다.

Ⅲ. Space의 機能과 調査必要性

人間의 文字生活에 있어서 情報源을 구성하는 element(要素)로 字素 뿐 아니라 space도 包含되어야 한다. (space는 單語와 單語, 句와 句, 文章과 文章을區分시키는 element(要素)이다) 지금까지의 發表[1],[9]에는 space element(空白素)가 누락되었다.

한글 情報源의 element(要素)에 對한 定量的인 考察은 한글 情報源의 code化 및 타자기에 있어서 最大能率을 주는 字盤의 制定, 電算裝置에서의 人力問題[10],[11] 解決等에 使用하게 되며 既存 Morse code 等의 再檢討에도 必要되는 重要한 기초가 되는데 space의 發生確率이 統計에 빠져 있음은 큰 하자가 되겠다.

Ⅳ. 統計節次 및 資料

文字 情報源으로 부터 다음 절차에 의하여 統計資料를 얻어 確率모델로 잡았다.

1. 小說과 新聞 等에서 random하게 임의로 구명을 뚫어 조사할 경우 特定의 單語가 sampling되는 것은 그 單語가 차지하는 表面積 즉 文字數에 比例하게 되는데 space의 길이는 例가 없이 고르지 않았기 때문에 space 表面積의 計算 나아가서는 發生確率의 計算에 지장을 주었다.

2. Random하게 sampling된 page 內의 space, 字素, 文字의 頻度數를 조사하였다. 이것은 既調査 字素發生確率[1]을 space를 包含시켜 보완하기 위한 資

料로서 利用하였다.

3. 여러개의 小說을 對象으로 삼아 sampling에 따라 確率의 變動을 보인 것이 그림 2 로서 표본수의 증가에 따라 6,000 element(要素) 內至 8,000 elment (要素)에서 이미 分布상태가 一定함을 보여 주었고, 따라서 50,000 sample數는 充分함을 알 수 있다. 그러나 특정표본에 의한 편중을 피하기 위하여 한 冊에 對하여 16,000 element(要素) 정도로 한정한 결과가 그림 2 로서 冊의 著者의 特性等에 따른 편차가 最大 0.5 以下로서 安定되어 나타났다.

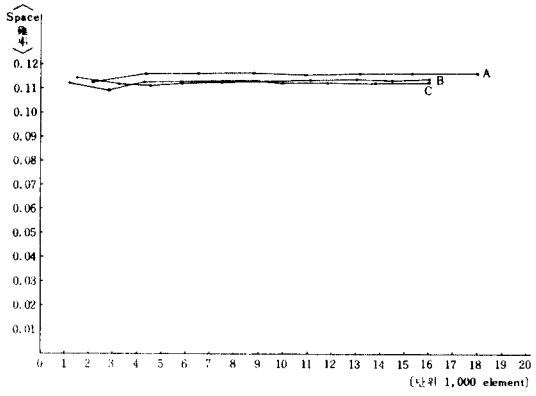

그림 2. 한글 element 發生頻度 對 space 確率의 그래프

Fig. 2. The plot of space probability against element frequency for HANGEUL.

4. 전체 element(要素) 50,000개 중에서 space 가 5,652개 字素가 44,348개, 文字가 16,712개로써 나타났다. 따라서 space 確率은 0.1131(space/element), 字素의 確率은 0.8869 (字素/element)이다.

5. 한글 각각의 字素에 對한 確率은 총 字素確率에 각 字素確率의 곱으로 구하였다. 결과는 표 1에 나타내었다.

표 1. 한글 element(要素) 確率과 entropies

Table 1. The probabilities and entropies of HANGEUL elements.

element	probability	$P_i \log_2 \frac{1}{P_i}$
space	0.1131	0.3556
ㅇ	0.1056	0.3425
ㅣ	0.0953	0.3232
ㅏ	0.0946	0.3218
ㄴ	0.0726	0.2747
ㄱ	0.0651	0.2566
ㄹ	0.0545	0.2288
ㅡ, ㅓ	0.0517	0.2209
ㅅ	0.0511	0.2192
ㅗ	0.0423	0.1930
ㄷ	0.0359	0.1723
ㅈ	0.0281	0.1448
ㅎ	0.0265	0.1388
ㅁ	0.0258	0.1361
ㅜ	0.0255	0.1350
ㅂ	0.0200	0.1129
ㅕ	0.0166	0.0982
ㅊ	0.0067	0.0486
ㅍ	0.0042	0.0332
ㅌ	0.0040	0.0319
ㅛ	0.0027	0.0230
ㅑ	0.0022	0.0194
ㅠ	0.0016	0.0149
ㅋ	0.0011	0.0108
total	0.9985	4.0769

※ 母音確率 0.3842

V. 한글의 情報量의 測定

統計資料에 의한 한글의 情報量은 다음과 같이 나타났다.

1. 最大 entropy(24 字素 + space = 25 element)

$$H\max = \log_2 25$$
$$= 4.644 \text{ (bits/element)}$$

2. Element entropy

$$H = \sum_{i=1}^{25} P_i \log_2 \frac{1}{P_i}$$
$$= 4.0769 \text{ (bits/element)} \quad \text{(第一表)}$$

3. Redundancy

$$R = 1 - \frac{H}{H_{max}}$$
$$= 0.12$$

4. 한글의 각 文字당 평균字素

$$\frac{\text{字素}}{\text{文字}} = 2.65 \text{ (字素/文字)}$$

5. 한글의 文字와 space의 比

$$\frac{\text{space}}{\text{文字}} = 0.34 \text{ (space/文字)}$$

6. space 와 element(要素)의 比
   0.1131 (space/element)
7. space 와 (space +字素)의 比

$$\frac{\text{space}}{\text{space + 字素}} = 0.25$$

Ⅳ. 한글 情報源의 特性 檢討

한글 element 의 情報量은 統計量의 분석 결과 다음과 같은 特性이 나타났다.

1. 한글의 element 당 redundancy(리던던시)는 12 %로 나타났다. 따라서 12 %의 정보손실은 해독이 가능하다. 英語의 경우 element(要素)의 發生確率[8]과 entropy를 표 2에 나타내고 있는데 英語의 경우 Redundancy 15 %(표 2의 결과를 이용하여 구함)에 비하여 낮다.

표 2. 英語 element 確率과 entropies[4]

Table 2. The probabilities and entropies of alphabet.

element	probability	$P_i \log_2 \frac{1}{P_i}$
word space or blank	0.2	0.4644
E	0.105	0.3414
T	0.072	0.2733
O	0.0654	0.2573
A	0.063	0.2513
N	0.059	0.2409
I	0.055	0.2301
R	0.054	0.2274
S	0.052	0.2218
H	0.047	0.2073
D	0.035	0.1693
L	0.029	0.1481
C	0.023	0.1252
F, U	0.0225	0.1232
M	0.021	0.1170
P	0.0175	0.1021
Y, W	0.012	0.0766
G	0.011	0.0716
B	0.0105	0.0690
V	0.008	0.0552
K	0.003	0.0251
X	0.002	0.0179
J, Q, Z	0.001	0.0100
total	1.0044	4.0453

∴ 母音確率 0.3109

2. 한글의 element(要素) 分布모델 그림 3 은 英語 그림 4 의 경우와 마찬가지로 Mandelbrot 모델[8]을 따르고 있다.

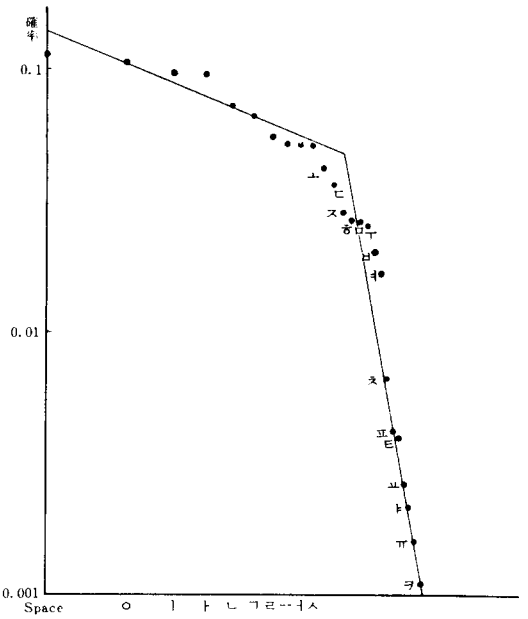

그림 3. 한글 element 發生할 순위 對 確率의 그래프

Fig. 3. The log-log plot of element probabilities against element freq. order for HANGEUL.

3. Element (要素)의 分布모델이 情報量에 비례하는 것으로 볼 때 이 모델을 情報量 分布모델이라하면 이 式은

$$N = \log \frac{1}{P_n} + K$$

〈여기서 N 은 element(要素)의 確率이 높은 순서를, R은 일정한 상수를 나타낸다〉.

실제 韓國語나 英語에 있어서 element(要素) 分布모델이 이러한 情報量分布모델에 따르고 있음이 그림 5, 그림 6에서 각각 보여 주고 있다. 그러나 이러한 特性에는 threshold(急減點)가 存在하며 여기에 위치하는 音은 "ㅏ", "ㅓ", "ㅛ", "ㅠ" 等의 젖은 母音(voyalles mouillées)과 "ㅊ", "ㅋ", "ㅌ", "ㅍ" 等의 강한 마찰음과 파열음의 情報量은 급격히 감소함을 알 수 있다. 이는 英語의 경우 "K", "X", "J", "Q", "Z" 等의 알파베트(Alphabet)에 對하여도 관찰된다. 〈그림 5와 그림 6 참고〉

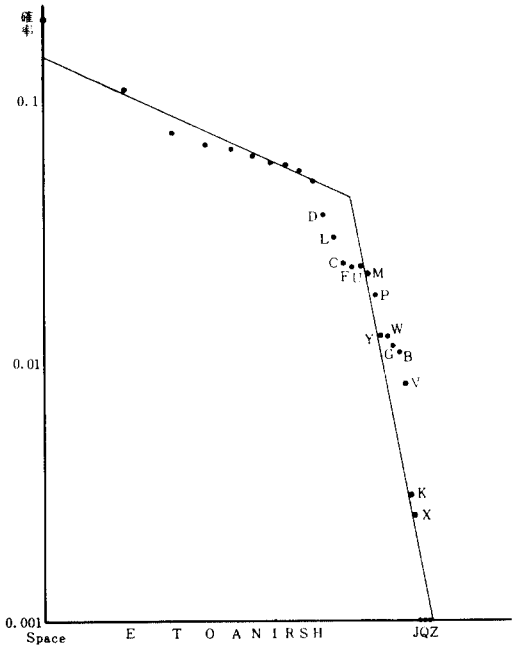

그림 4. 英語 element 發生頻度 순위 對 確率의 그래프

Fig. 4. The log-log plot of element probabilities against element freq. order for English.

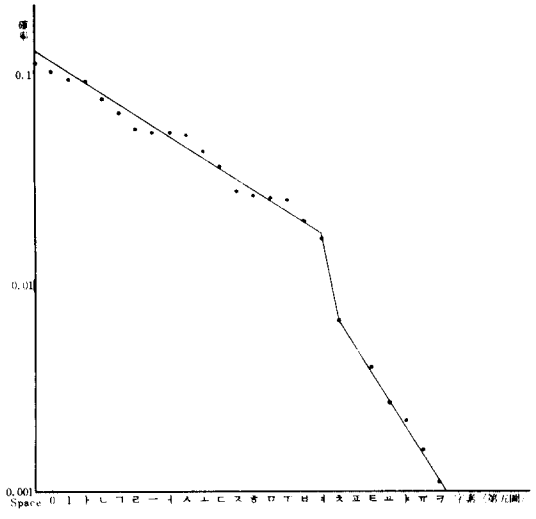

그림 5. 한글 element 발생순위 對 確率의 그래프

Fig. 5. The semi-log plot of element probabilities against element freq. order for HANGEUL.

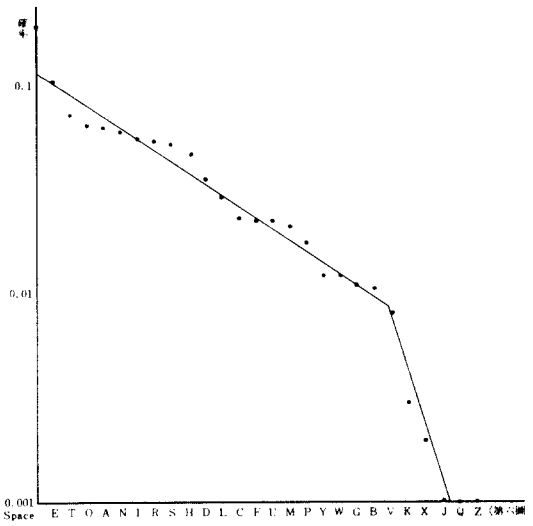

그림 6. 英語 element 發生순위 對 確率의 그래프

Fig. 6. Semi-log plot of element probabilities aganist element freq.order for English.

4. 韓國語의 母音確率과 英語의 母音確率은 각각 0.3842와 0.3109로서 韓國語에서 母音愛用 特性이 강한 것으로 나타났다. 〈표 1, 표 2 참고〉

이는 英語의 경우 모음충돌을 不協和音으로 보고, 子音의 삽입으로 이를 피하려는 경향(Ex. an apple등의 a → an)이 강한데 반하여 韓國語에 있어서는 前文字에 받침의 有無에 따라 "은→는", "을→를" "이→가", 등과 같은 모음충돌회피 현상이 있기는 하지만 母音의 어울림이 받아 들여지고 있음을 나타낸다. (Ex. 무수→무우, 무스다→모으다, 기슴마다→기음매다. 무이자(無+利子), 비이슬(비+이슬), 사오(四+五), 새아씨(새+아씨), 아우→아우, 等)

5. 韓國語의 element(要素)당 space(空白素) 確率과 英語의 경우 각각 0.1131, 0.2로써 〈第一表, 第二表 참고〉 韓國語의 space 確率이 英語에 比하여 매우 낮게 나타났다. 그러나 한글의 文字는 音節 單位로 한 묶음되는 것이니 文字와 space를 對等하게 취급하면 space 밀도가 현저히 증가한다.

6. 文字당 space 確率은 $0.25 \left(= \frac{\text{space}}{\text{문자} + \text{space}}\right)$로써 우리글의 모아쓰기 현상에서 오는 space 표면적 증가로 文字對 space의 입장에서 한글의 space가 英語에 比하여 높은데, 여기에 띄어쓰기의 過用은 정보전달의 기능을 감소시키며, 또한 space entropy도 감소시킨다.

Ⅶ. 結 論

1. 한글의 文字에 對한 space 確率은 0.25로써 나타났다.

2. 한글 entropy는 4.0769로 나타났다. 즉 한글情報源을 redundancy zero로 code化할 경우 각 element(要素)당 평균 bit 數로는 4.0769以上이 되어야 한다.

3. 한글 각 文字는 평균 2.65 字素로 되어 있다.

4. 한글 element(要素)의 redundancy는 12%로 나타났다.

5. 한글의 字素와 英語의 Alphabet 두가지 모두情報量 分布모델이 관찰되나 한글에서는 "ㅏ" "ㅓ" "ㅗ" "ㅜ" 등의 짧은 母音과 "ㅅ" "ㅋ" "ㅌ" "ㅍ" 等의 격음에서, 英語에서는 "K" "X" "J" "Q" "Z"等의 音에서 threshold(急減點)가 관찰된다.

〈그림 5, 그림 6 참고〉

參 考 文 獻

1. 文敎部 : "우리말에 쓰인 글자의 頻度調査" 1956. 6. 12.
2. 南宮健 ; " 한글낱말의 發生頻度分布와 entropy 에 관한 硏究" 서울大學校 大學院 碩士學位論文 1979. 2.
3. 李柱根, 崔興文 ; " 韓國語 音節의 entropy에 관한 硏究" 論文 74-11-3-3.
4. Abramson; "Information theory and coding"
5. Papoulis; Probability, Random Variables, and Stochastic Processes".
6. Harry L. Van Trees; "Detection, Estimation, and Modulation Theory".
7. R. E. Zimer and W. H. Tranter ; "Principles of Communications".
8. Leon Brillouin; "Science and Information Theory".
9. 임종철 ; "타자연구물집" p. 30.
10. 安秀桔 ; "電子計算機의 한글 入出力에 관한 硏究現況과 한글 반 풀어쓰기 提案" 電子工學會誌 第十卷 第二號 1973.4.
11. 安秀桔 ; "電子計算機의 한글 入出力에 관한 硏究現況과 한글 반 풀어쓰기 提案(Ⅱ)" 電子工學會誌 第十卷 第三號.
12. 安秀桔 ; " 한글文字모아쓰기 Display의 한 방안" 電子工學會誌 第十二卷 第一號 1975.2.