한글문헌에 있어서 Zipfian현상에 관한연구

A Study of Zipfian Phenomena in Hangul Literature

신강현* 이두영**

초 록

본 연구는 Zipf가 최초에 유도한 공식이 한글문헌에 있어서도 그 타당성이 성립하는지의 여부를 조사연구하였다.

그 결과 한글문헌에 있어서도 단어의 수록빈도와 등급사이에 일정한 통계적인 법칙성이 존재하며 이 현상은 Zipf가 유도한 공식과 일치하는 것으로 나타났다. 한편 Zipf의 제2법칙은 한글문헌에 적용되지 않았기 때문에 본 연구에서는 이에 적합한 공식을 유도하였다.

ABSTRACT

The purpose of this Study is to investigate the Zipfian distribution in Hangul literature. The result shows that the formulas derived from the Hangul literature are in accordance with the generalized Zipf's first law. The result also shows that the formulas derived from the Hangul literature are not in accordance with the Zipf's second law and the generalized Zipf's second law.

*숙명여자대학교도서관
**중앙대학교 도서관학과

第1章 緒 論

第1節 硏究의 目的

情報 그 自體는 반드시 어떤 物理的인 現象을 통해서만 表現되며, 어떤 意圖된 目的下에 이를 受容할 相對方(受容者)에게 傳達된다. 이와같이 情報가 傳達되기 위해서는 情報發生者와 受容者사이에 이미 約定된 어떤 記號體系로 表現되어야 한다.

그러나 人間社會에 있어서 이미 約定된 記號體系를 통해서 情報를 傳達한다고 하더라도 情報發生者는 항상 그가 意圖한 目的을 보다 效果的이고 正確하게 受容者에게 傳達할 수 있는 經濟的인 方法을 模索하려고 노력한다. 따라서 情報發生者는 可能한 最少의 記號를 利用하여 많은 量의 情報를 傳達할 수 있는 方法을 選擇하려고 한다. 그러나 이와같은 情報發生者의 行爲는 正確한 情報를 迅速하게 認知하려는 受容者의 行爲와는 相反되게 나타나고 있음을 알수 있다.

만약 情報發生者가 노력을 最少로 하기 위해서 記號를 단 하나 혹은 몇개의 記號로 많은 量의 情報를 傳達하려고 試圖한다면 受容者側에서는 情報發生者가 意圖한 目的을 正確하게 認知하지 못할 可能性이 크며 正確하게 認知한다고 하더라도 많은 時間과 노력이 要求된다. 逆으로 情報發生者가 必要以上으로 많은 記號를 動員하여 情報를 傳達하려고 한다면 受容者側에서는 그 情報를 正確하게 理解할 수는 있지만 情報發生者側에서 볼 때는 매우 非經濟的이라 할 수 있다.

이와같이 情報를 傳達하는 過程에 있어서 記號의 選擇은 情報發生者와 受容者의 相反된 要求에 따라 影響을 받게된다. 이러한 相反된 두 행위를 G.K. Zipf는 "情報傳達記號數의 單一化 傾向(Force of Unification) 또는 情報發生側의 經濟性(Speaker's economy)과 情報傳達記號數의 多數化 傾向(Force of Diversification) 또는 受容者側의 經濟性(Auditor's economy)"으로 表現하고 있다.[1] 이와 관련하여 G.K. Zipf는 "言語學이란 언어산출을 精密科學의 客觀精神으로 探究되어야 하는 自然的인 心理學 내지 生物學的 現象"이라 보고 이에 관한 硏究의 한 方法으로 "관찰가능한 言語(말)現象에 統計的 原理를 適用시키는 것"이라고 제시한 바 있다. 이 原理에 따라 G.K. Zipf는 單語의 利用頻度를 調査하여 頻度가 높은 單語에서 낮은 單語의 順으로 等級을 할당해 얻은 單語의 利用頻度와 그 等級사이에 存在하는 어떤 一定한 統計的인 法則性을 찾으려 試圖하였던 것이다.

일찌기 單語의 利用頻度와 그 等級사이의 어떤 一定한 統計的인 法則性 즉 利用頻度와 等級의 곱이 一定하다는 것은 1915年 L.P. Ayres에 의해 最初로 계산되었으며, 1916年 J.B. Estoup에 의해 單語의 利用頻度와 等級사이에 存在하는 逆關係를 糾明하려는 試圖가 있었다.

이러한 統計的인 法則性은 1949年 G.K. Zipf에 의해 體系化되었다. 그는 이러한 現象이 "最少努力의 原理(The Principle of least effort)"에 起因한 것이라고 主張했으며 이를 根據로 해서 가장 많이 利用되는 單語는 費用이나 傳達의 側面에서 費用이 가장 적게 들거나

1) G.K. Zipf, Human behavior and the Principle of least effortian introduction to human ecology. Cambridge: Addison-Wesley, 1949. p.21.

單語의 길이가 짧은 것들이라고 主張했다. 그리고 더 나아가서 이 原理가 人間行動全般을 支配하는 法則이라고 믿었다.

그러나 이러한 法則性을 根據로 하여 G.K. Zipf가 誘導한 Zipf 法則에 對한 硏究는 아래의 세가지 側面[2]에서 現在 硏究가 계속 진행되고 있다.

(A) G.K. Zipf가 最初에 誘導한 公式의 妥當性與否의 調査

(B) G.K. Zipf가 誘導한 公式보다 더 適合한 公式의 誘導

(C) 利用頻度와 그 等級사이에 存在하는 어떤 一定한 統計的인 法則性에 對한 理論的 根據의 定立

本 硏究의 目的은 위의 세영역중 G.K. Zipf가 最初에 誘導한 公式의 한글문헌에서의 妥當性與否를 調査하고 이러한 統計的인 法則性의 存在를 糾明하며 그리고 만약 G.K. Zipf가 誘導한 公式이 適用되지 않을 경우에는 한글문헌에 適合한 公式을 誘導하는데 있다.

第2節 硏究의 方法 및 範圍

本 硏究의 目的을 達成하기 위해서 다음과 같은 假說을 設定하였다.

(A) 한글문헌에 있어서도 單語의 利用頻度와 그 等級사이에 一定한 統計的인 法則性이 存在한다.

(B) G.K. Zipf가 最初에 誘導한 公式이 한글문헌에도 適用된다.

이와같은 假說下에서 實驗을 위한 分析對象資料로써 圖書館學分野에 관한 著作物中 單行本 一卷과 碩士學位論文 二卷을 任意로 選定하여 本文에 收錄되어 있는 單語中 外國語로 表記된 單語(英語, 獨語, 日語等)와 숫자를 除外한 한글(漢學包含)로 表記된 單語만을 分析對象으로 하였다.

分析對象資料中 單行本으로는 鄭獄謨의 "文獻情報學原論"(서울 : 아세아문화사, 1977)을, 碩士學位論文으로는 李振榮의 "大學圖書館의 利用率에 影響을 미치는 要因硏究"(서울 : 成均館大學校 大學院 圖書館學科, 1976)와 朴英熙의 "書庫移動時機 및 書庫區分모델에 관한 硏究"(서울 : 成均館大學校 大學院 圖書館學科, 1985)를 각각 選定하였다. 分析對象資料로 單行本과 碩士學位論文을 選定한 理由는 分析對象單位인 單語의 量의 많고 적음에 따른 Zipf 公式의 變化與否를 區別하기 위해서이다.

資料는 SPSS(Statistical Package for the Social Sciences)를 利用해 전산처리하였다.

第2章 Zipf 法則

第1節 Zipf法則의 理論的 根據

1920년대에 G.K. Zipf는 하바드대학의 대학원생으로써 言語에 있어서 音聲學上의 變化를 硏究하는 過程에서 音素의 利用頻度에 흥미를 갖게 되었다. 그러나 그의 關心은 音素의 利用頻度에서 單語의 相對的인 頻度로 바뀌어 갔으며 1932년에 出版된 그의 著書 "Studies of the principle of relative frequency in language"에 자신의 硏究結果를 發表하였다.[3] 그리고 1935년에 出版된 그의 두번째

2) R.E. Wyllys, "Empirical and theoretical bases of Zipf's law" Library Trends, Vol.30 No.1 (Summer 1981), p.56.

3) loc.cit.

著書 "The Psycho-biology of language"에서 利用頻度와 等級사이의 關係를 공식화하여 이를 Zipf 法則이라 命名하였다.[4] 1949년에 出版된 그의 세번째 著書 "Human behavior and the principle of least effort : an introduction to human ecology"에서 等級과 利用頻度 사이에 存在하는 一定한 統計的인 法則性에 對한 理論的 根據로서 "最少努力의 原理"라는 槪念을 說明했으며 이 原理를 확장하여 人間行動全般에까지 적용시켰다.[5]

G.K. Zipf는 그의 세번째 著書에서 "最少努力의 原理"를 다음과 같이 說明했다.[6]

> 作業의 最少化는 오늘의 問題를 解決하기 위해 作業을 행함에 있어서 만약 오늘의 作業이 最少化되지 않았다면 要求되어지는 것 이상으로 내일의 作業을 增加시킬 것이라는 結論이 誘導된다. 逆으로 오늘의 作業에 必要한 것보다도 더 많은 量의 作業을 행했더라면 내일에는 오늘 더 많이 행한 作業量만큼 作業을 節約할 수 있다.
>
> 그리고 여기서 언급된 오늘과 내일의 關係는 個個人 全體가 한순간에 作業에 對한 그의 投資比率이 다음 순간에 그의 作業의 最少化에 영향을 미칠 것이라는 완전한 연속의 관계를 의미한다.
>
> 한 개인이 그의 時間當 平均作業投資比率(Average rate of work expenditure over time)을 最少化할 수 있기 이전에 그는 우선 미래에 발생할 수 있는 돌발사고를 고려한 후에 作業의 最少平均比率을 계산해야 한다.
>
> 그렇게 함으로써 個人은 作業의 平均比率을 완전히 最少化하지는 못하지만 可能한 한도내에서 最少化하려고 노력할 것이다. 편의상 이것을 "最少努力(least effort)"이라고 부른다.

이와같이 G.K. Zipf는 사람들이 어떤 作業을 遂行함에 있어서 그의 時間當 平均作業投資比率을 可能한 한 最少化할 수 있는 方法으로 行動을 한다고 主張했으며 이러한 原理가 모든 人間行動에 適用될 수 있는 法則이라고 믿었다.

따라서 그는 人間의 情報傳達行爲에 利用되는 言語에 있어서도 이 原理가 適用된다고 主張했다. 즉 말을 할 수 있다는 行爲는 말을 할 수 없다는 行爲보다도 어떤 目的을 쉽게 達成할 수 있으므로 一般的으로 잠재적인 經濟性을 가진다고 말을 할 수 있다. 그리고 情報를 傳達하기 위해서 言語을 利用하는 方法에 있어서도 相對的으로 經濟的인 方法과 非經濟的인 方法이 存在한다고 믿고 있다. 즉 單語들은 一般的으로 어떤 意味를 지니고 하나 혹은 둘이상이 모여서 情報를 傳達하므로 情報發生者의 側面과 受容者의 側面에서 意味를 지닌 單語들을 選擇·組合하는 經濟的인 方法과 非經濟的인 方法이 存在하게 된다.

따라서 情報를 傳達하는 單語를 選擇하는 情報發生者의 觀點에서 볼 때 오직 하나의 單一單語로 모든 情報를 傳達할 수 있다면 이 方法이 가장 經濟的인 方法임에 틀림없다. 왜냐하면 情報發生者는 많은 單語를 계속해서 알고 있어야 하는데 드는 노력과 알고 있는 많은 單語中에서 必要한 單語를 選擇하는데 드는 노력을 아낄 수 있기 때문이다. 그리고 情報를 傳達받

---

4) Ibid., p.57.
5) Ibid.
6) G.K. Zipf, Op. Cit. p.6.

는 受容者의 側面에서 볼 때 그 情報가 單一單語로 構成되어 있다면 그 單語가 內包하고 있는 情報(情報發生者의 意圖된 情報)를 正確히 把握한다는 것은 거의 不可能하다. 따라서 受容者는 情報發生者가 傳達하려는 情報를 各各 다른 意味를 지닌 單語들로 構成할 때 그 情報를 認知하는데 드는 노력을 최소화할 수 있다.

그러므로 單語의 選擇過程은 이런 두개의 相反된 힘에 의해서 支配된다.이 두개의 相反된 힘을 G.K. Zipf는 "情報傳達記號數의 單一化 傾向(情報發生者側의 經濟性)"과 "情報傳達記號數의 多數化 傾向(受容者側의 經濟性)"이라고 命名했다.

情報傳達記號數의 單一化 傾向은 오직 하나의 單一單語로 만든 意味를 統合시킴으로써 單一單語로 利用單語의 數를 減少시키려는 傾向을 말하며 그리고 情報傳達記號數의 多數化 傾向은 各各 다른 意味를 지니는 다른 單語가 반드시 存在할 것이라는 觀點에서 利用單語의 數를 最大限으로 增加시키려는 傾向을 말한다.

따라서 情報를 傳達하기 위해서 單語를 選擇할 경우 情報發生者와 受容者는 이러한 理論的인 두힘사이의 均衡點(Vocabulary Valance)을 찾아서 가장 經濟的으로 單語를 選擇해야 한다.

그러나 人間은 어떤 行動을 하더라도 항상 균등하게 노력을 투자할 수는 없다. 그러므로 두힘사이의 正確한 均衡點을 說明할 수는 없다. 따라서 均衡點의 存在를 正確히 說明하기 위해서 G.K. Zipf는 다음의 假定을 提示하였다.

(A) 人間은 어떤 作業을 遂行함에 있어서 항상 노력을 經濟的으로 投資한다.

(B) 前述한 두힘사이의 均衡點에 對한 論理가 타당하다.

따라서 均衡點은 다른 單語들의 數를 增加시키려는 힘과 1로 減少시키려는 힘을 각각 매개변수로 하는 函數關係에 있다.

이와같이 G.K. Zipf는 單語의 利用頻度와 그 等級사이에 存在하는 어떤 一定한 統計的인 法則性을 "最少努力의 原理"라는 前提下에서 說明하려고 노력했다. 그러나 이 原理만으로는 그 自身이 發見한 公式의 誘導過程을 論理的으로 說明하지 못했다.

따라서, G.K. Zipf가 그의 法則을 公式化하고 理論的 根據를 提示한 이래 여러 學者들에 의해서 公式이 修正되고 새로운 理論的 根據들이 提案되었다. 그 代表的인 學者들로서 B. Mandelbrot, D. J. D. Price, H. Simon, B.M. Hill, S.D. Haitun등을 들 수 있다.

B. Mandelbrot[7)8)]는 Zipf 法則의 理論的 根據를 提示하기 위해서 Shannon과 Weaver에 의해 發展된 情報理論의 側面에서 接近을 試圖하였다. 그는 情報傳達過程에 있어서 주어진 時間內에 情報의 最大量을 傳達하는 方法으로 單語를 選擇할 것이라는 假定下에 읽는 時間과 한 單語를 처리하는데 소모되는 시간을 대상으로 實驗을 試圖하였다. 이 實驗의 結果는 그가 假定했던 사실과 일치하였다.

따라서 B. Mandelbrot는 情報를 傳達함에 있어서 그 情報를 構成하는 文字와 空間으로

7) M.K. Buckland and A.Hindle, "Library Zipf," Journal of Documentation, Vol.25 No.1, 1969. p.55.

8) Jane Fedorowicz, "The theoretical foundation of zipf's law and application to the bibliographic data base environment", Journal of the American Society for Information science, Vol.33, No.5, 1982. p.287.

인하여 소모되는 傳達費用을 最少化하려는 노력 때문에 Zipf 法則과 같은 현상이 나타난다고 주장했다.

D.J.D. Price[9) 10)]는 많이 인용되는 資料가 거의 인용되지 않았던 資料보다 그후 더 자주 인용되며 자주 이용되는 단어가 이후 더 자주 이용된다는 상황을 통계학적인 측면에서 만든 모델인 누가편익분포(Cumulative Advantage Distribution)를 利用해서 Zipf 法則을 說明하려고 했다. 그는 누가편익분포가 Zipf法則뿐 아니라 Bradford法則, Lotka法則, Pareto 法則(부의 분산), 인용빈도의 분산등 모든 實驗에 의한 結果를 說明할 수 있는 기초가 되는 확률 이론이라고 주장했다.

이러한 성공의 법칙[11)](Law of success)은 1774년 P.S. Laplace에 의해서 最初로 誘導되었으며 매우 多樣한 社會現象에서 그 실례를 찾아볼 수 있다.

누가편익분포의 간단한 원리는 G. Polya 의 항아리모델(urn model)을 변형시킴으로써, 즉 成功이 이전에 發生한 成功에 의해서는 影響을 받지만 失敗는 이전에 發生한 失敗에 의해 影響을 받지 않는 수정된 항아리모델로 부터 誘導할 수 있으며 보다 一般的인 境遇는 Feller에 의해 誘導된 確率論的 純粹誕生過程(Stto-chastic pure birth process)으로 부터 誘導할 수 있다.[12)](부록1 참조)

A. Rapoport[13)]는 쌍곡선에 근거를 둔 누가편익분포이론을 반박했다. 그는 만약에 分析對象物이 頻度順에 따라 가장 높은 것에서 부터 낮은 것의 順으로 배열된다면 약간 단조롭게 하향하는 곡선으로 표현되지만, 그러나 이들곡선이 상당히 쌍곡선에 유사하다는 사실만으로 쌍곡선이라는 결론을 유도할 수 없다고 主張했다. 그러므로 理論的 結論은 만약 곡선이 어떤 계층에 속해야 할 根本的인 理由가 提案될 수 있어야만 誘導할 수 있다. 따라서 根本原理의 內容은 Content-bound theory라고 그는 主張했다.

H. Simon은 Zipf 法則의 特徵을 說明하기 위해서 "비대칭분포(Skew distribution)"가 지니는 세가지 特徵을 利用했다.[14)]

(1) 비대칭분포는 매우 긴 상부(Upper tail)를 가진 역J형태로 나타나거나 매우 비대칭적이다. 따라서 끝부분은 일반적으로 다음의 함수와 매우 일치할 것이다.

$$f(r) = (a/r^k)^{b^r} \qquad (A)$$

여기서 f(r)은 r번 발생하는 단어의 수이며 a, b, k는 상수이다. 그리고 상수 b는 거의 1에 가까운 가치를 지니므로 f(r)에 거의 영향을 미치지 못한다.

9) S.M. Lawani, "Bibliometrics:its theoretical foundations methods and applications", Libri, Vol.31 No.4, 1981, p.296.

10) Derek de solla price, "A General theory of bibliometric and other cumulative advantage processes", Journal of the American Society for information science, Vol.27 No.5-6, 1976, p.292.

11) B.C. Brookes, "Toward informetrics:Haitun, Laplace, Zipf, Bradford and the alvey programme" Journal of Documention, Vol.40, No.2 1984, p.122.

12) Derek de Solla Price, Op, Cit, pp.293-295.

13) D.O. O'Conner and Henry Voos, "Empirical laws, theory construction and bibliometries", Library Trends, Vol.30, No.1 (summer 1981), p.11.

14) G.Herdan, "A Critical examination of Simon's model of certain distribution function in linguistics", Applied statistics, Vol.10 No.2 1961, p.65.

(2) 指數 K는 1보다 크며 單語頻度分散에 있어서 2에 매우 가깝다(이 경우 b의 값은 1이다).

$$f(r) = a/r^2$$
$$r^2 \cdot f(r) = a \qquad (B)$$

(3) 單語頻度分散에 있어서 公式(A)은 단지 끝부분에 있어서 分散뿐 아니라 r의 매우 적은 가치를 위한 分散도 나타낸다. 이 경우 f(1)/V은 일반적으로 약 1/2이다. (여기서 V는 전체어휘수를 의미한다). 그리고 f(2)/f(1)은 일반적으로 약 1/3이다.

H.A. Simon은 그의 假定을 根據로 하여 다음의 公式을 誘導했다.[15] (부록2 참조)

$$f(r) = c \cdot r^{-m}$$

따라서 H.A. Simon이 誘導한 公式이 Zipf 法則과 같은 형태를 취함을 알 수 있다.

G. Herdan은 單語頻度分散이 복합포아송분산(Compound Poisson Distribution)과 유사하다고 主張했으며 公式을 誘導하기 위해서 Waring의 公式인 $\frac{1}{p-q}$ 을 利用했다. G. Herdan이 誘導한 頻度의 平均과 分散은 다음과 같다.[16] (부록3 참조)

$$\mu = \frac{q}{P-q-1}$$

$$\sigma^2 = \frac{q(p-1)(p-q)}{(p-q-1)^2(p-q-2)}$$

B.M. Hill은 生物學의 種(species)과 屬(genus)의 關係에 있어서 전통적인 점유모델(Classical occupancy model)의 보오스-아인쉬타인분산(Bose-Einstein distribution)을 利用해 Zipf 法則을 誘導했다. 그는 "屬-種關係(generic-specific form)"와 이 형태를 위한 모델의 변형으로 부터 유도될 수 있는 "等級-頻度關係(Rank-Frequency from)"라고 칭한 Zipf 法則의 두가지 형태를 아래와 같이 상술했다.

① 屬-種關係

屬-種關係는 n개의 種을 지닌 屬의 比率이 $n^{-(1+a)}$에 비례하는 關係를 意味하며 이것은 Zipf 第二法則과 거의 일치한다(여기서 $a>0$이다). 이러한 屬-種關係는 1922년 Willis에 의해 最初로 發見되었다.[17] Willis는 n개의 種을 지닌 屬의 數 g(n)가 다음과 같음을 관찰했다.

$$g(n) = kn^{-(1+a)} \qquad a>0,\ k: \text{상수}$$

② 等級-頻度關係

等級-頻度關係는 r번째로 큰 集團의 量이 $r^{-(1+a)}$에 비례하는 關係를 意味하며 이것은 Zipf 第一法則의 정의와 비슷하다.($0<a\leq 1$)

B.M. Hill과 M. Woodroofe가 유도한 屬-種關係는 다음과 같다(부록4 참조).

$$E\{\theta(1-\theta)^{n-1}\} = [n(n+1)]^{-1}$$

15) G. Herdan, Op. Cit., p.67.

16) Jane Fedorowicz, "The theoretical foundation of Zipf's law and its application to the bibliographic database environment", Journal of the American Society for information Science, Vol.33, No.5, 1982, p.290.

17) R.A. Fairthone, "Empirical hyperbolic distributions) (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction", Journal of Documentation, Vol.25, No.4, 1969, pp.324-325.

이 公式이 매우 여러 資料에 實際로 일치되는 Zipf 法則의 간단한 형태이다.

그러나 이 公式에서 θ의 가치의 적절한 선택에 의해서 公式을 誘導했으므로 이 公式을 屬一種關係의 약한형태 (weak form) 라고 한다.

屬一種關係의 강한형태 (stronger form) 는 In/nT 그 자체가 $n^{-(1+a)}$에 거의 인접하게 변화하는 경향이 있음을 의미하는 것으로 다음과 같다.[18] (부록 4 참조)

$$p(n) \sim cn^{-(1+a)} \qquad c : 상수$$

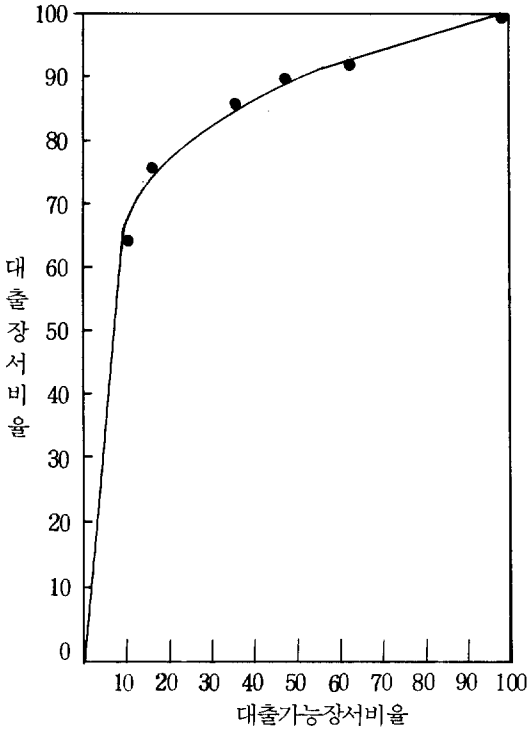
여기서 $p(n) \geqq 0$이고 $\sum_{n=1}^{\infty} p(n) = 1$ 이다.

等級一頻度關係는 屬一種關係를 수정함으로써 유도할 수 있다. 만약 $L^{(r)}$이 r 번째 큰 種에 있는 屬의 수라면 혹은 $L^{(r)}$이 한 국가에 있는 r 번째 큰 도시의 인구라면 $L^{(r)}$은 $r^{-(1+a)}$에 거의 비례한다.[19] ($a>0$)

R.A. Fairthorne은 Lotka, Bradford 그리고 Zipf 분산이 各各 그 출발점이 무엇이던지간에 그것들 모두는 대상물이 분류되고 등급지워지는 等級一頻度順 (rank-order frequencies) 에 기초를 두기 때문에 쌍곡선분포(hyperbolic distribution)의 일종이라고 주장했다.[20]

R.W. Trueswell 은 Zipf 법칙이 도서관에 있어서 책과 정간물의 분산에 적용된다고 주장했다. 그는 공공도서관과 전문도서관, 학원도서관에 있어서 총대출의 백분율과 대출가능장서의 백분율을 各各 軸으로 하여 그래프로 나타낸 결과 이 分布를 80/20 法則이라고 主張했다.[21] 아래의 그림 1 은 Air Force Cambridge Research Laboratory에서 수집된 대출장서비율대 대출가능 장서비율의 그래프이다.

L.S. Kozachkov 와 L.A. Khursion[22] 은 情報의 흐름에 있어서 알려진 많은 統計規則들의 根本的인 類似性을 主張했다. 그들은 소위 "쌍곡선사닥다리 (hyperbolic ladder)"라 불리우는 모델을 構成했으며 이 모델을 根據로 Zipf 法則을 說明했다.



출전 : Air Force Cambridge Research Laboratory에서 수집된 데이터

그림 1. 대출가능장서비율대 대출장서비율

18) B.M. Hill and M. Woodroofe, "stronger forms of Zipf's law", Journal of the American Statistical Association, Vol.70, No.349, 1975, p.212.

19) Ibid.

20) R.A. Fairthorne, Op. Cit., pp.319-343.

21) F.W. Lancaster, The Measurement and evaluation of library service, Washington, D.C.:Information resources press, 1977, p.347.

22) M.K. Buckland and A. Hindle, op.cit, p.52.

소련의 統計學者 S.D. Haitun[23] 은 實驗에 의한 分散들이 구지안유형(Gaussian type)과 지피안유형(Zipfian type)중 하나라고 지적했으며 이 두유형을 모멘트(moment)의 存在與否에 따라 구별했다. 그리고 지피안유형의 分散이 社會全般에 發生하며 구지안유형의 통계기술을 이 분야에 적용시킬 수 없다고 주장했다.

그러나 G.K. Zipf가 그의 法則을 公式化한 이래로 많은 學者들에 의해 理論과 公式이 수정되고 새로 제안되었지만 통일된 공식과 이론은 아직 정립되지 않은 상태이다.

第2節 公式의 發展

어떤 책자의 본문에 수록되어 있는 異種單語의 收錄頻度를 조사하여 收錄頻度가 가장 높은 단어에서 낮은 단어의 順으로 배열하여 1, 2, 3, …N과 같이 등급을 매겼을 때 이들 개개 단어들의 等級과 收錄頻度의 곱은 일정하다는 統計的인 法則性을 Zipf 법칙이라 하며 이것을 공식화하면 다음과 같다.

$$r \times f = c \qquad (C)$$

위의 공식(C)를 Zipf 第一法則이라고 하며 여기서 변수 r 은 등급을, 변수 f 는 收錄頻度를 상수 C는 어떤 특정책자에 해당되는 상수를 의미한다.

일반적인 Zipf 법칙은 대수학적으로 나타낼 때는 공식(C)의 형태로 표현되지만 그래프상에서 두 변수 r 과 f 사이의 일정한 統計的인 法則性의 존재여부를 보다 쉽게 알아 볼 수 있도록 하기 위하여 공식(C)를 로그함수로 변화시킬 필요가 있다. 공식(C)의 양변에 log를 취하면

$$\log(r \times f) = \log c$$

$$(\log r) + (\log f) = \log c \qquad (D\text{-}1)$$

가 된다. 여기서 log의 밑은 10인 상용대수이다.

앞의 공식(D-1)에서 B의 기울기를 지니는 임의의 방정식을 유도함으로써 공식(C)를 일반화 시킬 수 있다.

$$B \log r + \log f = \log C \qquad (D\text{-}2)$$

공식(D-2)를 다시 공식(C)와 같이 代數學的으로 나타내면

$$r^B \times f = C \qquad (D\text{-}3)$$

가 된다.

이때 B의 값이 만약 1이라면 공식 (D-3)은 공식 (C)와 동일해진다. 따라서 공식 (D-3)을 일반화된 Zipf 법칙이라고 한다.

이상에서 언급된 Zipf 第一法則과 일반화된 Zipf 法則은 收錄頻度가 낮은 단어에는 정확하게 적용되지 않는다는 것이 실험을 통해 밝혀졌다. 아래의 표1과 표2의 Western Reserve University의 Center for Documentation and Communication Research에서 행한 3가지 단행본에 대한 分析結果와 Eldridge의 分析結果를 나타낸 것이다.[24] 이 표에 의하면 단 한번 수록된 異種單語의 數가 異種單語總數에서 차지하는 비율은 WRU1의 경우 541/1001,

23) B.C. Brooks, op. cit., p.120-121.

24) Andrew D. Booth, "A law of occurrence for words of low frequency", In: Introduction to information science, edited by Tefko Saracevic. New York: R.R. Bowker Co., 1970, p.220.

表1) 단행본3권과 신문에 대한 收錄單語調査

	WRU1	WRU2	WRU3	Eldridge
收錄單語總數 T	4325	4409	8734	43989
異種單語總數 D	1001	1211	1698	6002
收錄頻度1인 單語數 I_1	541	710	887	2976
收錄頻度2인 單語數 I_2	152	227	273	1079
收錄頻度3인 單語數 I_3	94	91	151	516
收錄頻度4인 單語數 I_4	56	41	90	294
收錄頻度5인 單語數 I_5	36	32	62	212

表2) 異種單語總數에 대한 단 한번 收錄된 단어의 비율

	異種單語總收(D)	收錄頻度1인 單語數	I_1/D
WRU1	1001	541	·54
WRU2	1211	710	·58
WRU3	1698	887	·52
Eldridge	6002	2976	·50

WRU2의 경우 710/1211, WRU3의 경우 887/1698이고 Eldridge의 경우 2976/6002이므로 한번 수록된 단어의 數가 異種單語總數의 반이상임을 알 수 있다. 따라서 이와같이 낮은 빈도의 단어들에게도 적용될 수 있는 새로운 Zipf법칙을 유도해 내야 할 필요성이 대두된다.

表1은 단행본 3권에 대한 分析結果와 Eldridge의 分析結果를 나타낸 것으로 分析對象資料에 收錄되어 있는 單語의 總數와 異種單語의 總數, 그리고 數錄頻度가 낮은 單語들을 頻度에 따라 기록한 表이다.

表2는 각 자료의 分析結果 收錄頻度가 1인 單語들이 異種單語總數中 차지하는 비율을 나타낸 表로 1번 收錄된 單語가 異種單語總數의 50%이상임을 알 수 있다.

만약 모든 異種單語를 서로 중복되지 않고 완벽하게 等級을 매길 수 있다면, 그리고 등급r에 해당하는 한 단어가 收錄될 확률을 P(r)이라고 한다면 收錄單語의 總數가 T인 어떤 책자의 경우 단어의 收錄頻度를 다음과 같이 나타낼 수 있다.[25]

T·P(1) : 등급1인 단어의 수록빈도
T·P(2) : 등급2인 단어의 수록빈도
⋮ ⋮
T·P(r) : 등급r인 단어의 수록빈도

따라서, Zipf第二法則은 다음과 같다.

(부록5 참조)

$$In/I_1 = \frac{3}{4n^2 - 1} \qquad (E)$$

A.D. Booth는 공식(D-3)을 이용해 Zipf第二法則을 보다 일반화시켰다. 일반화된 Zipf第二法則은 다음과 같다.(부록6을 참조)

$$In/I_1 = \frac{2}{n(n+1)} \qquad (F)$$

그러나 G.K. Zipf가 유도한 공식과 실제로 실험에 의해 얻어진 수치사이에 차이가 있음이 자주 발견되었다.[26] 실험에 의해 얻어진 收錄頻度의 가장 전형적인 일탈은 收錄頻度가 높은

25) A.D. Booth, op.cit., pp.220-221.
26) A.D. Booth, op.cit., p.221.

단어들의 收錄頻度가 Zipf 곡선에 의해 예견되어지는 수록빈도보다 아래에 위치되어진다는 점이다. 이와 같은 일탈은 공식의 유도과정중 회귀방정식을 이용해 원식을 유도하였기 때문에 소수의 수록빈도가 높은 단어에 있어서는 오차가 생기기 때문이다.

이상과 같은 일탈은 B. Mandelbrot가 유도한 공식에 의해 해결되어질 수 있다.[27] B. Mandelbrot는 지수B가 고정된 가치를 가지는 것이 아니라 역으로 변화한다는 개념을 도입해 Zipf 공식을 다음과 같이 수정하였다.

$$(r + m)^B \times f = C$$

여기서 r은 등급, f는 수록빈도를 의미하며 그리고 3개의 상수m, B, C는 어떤 특정 책자에 해당되는 상수를 의미한다. 이 공식에 있어서 주요한 점은 m이 r의 값이 낮을수록 최대효과를 가지며 그리고 이 공식이 공식(C)나 공식(D-3)보다 특히 등급이 높거나 낮은 단어들에 잘 일치된다는 점이다.

H.P. Edmundson은 3개의 매개변수를 이용하여 G.K. Zipf가 유도한 공식을 보다 일반화시켰다.

$$f(r : c, b, a) = c(r + a)^{-b}$$

여기서, 매개변수 a, b, c는 양의 상수이다.

第3章 實驗을 위한 資料의 蒐集 및 分析

第1節 資料의 蒐集

實驗을 위한 分析對象資料로는 한글로 쓰인 圖書館學分野文獻中 碩士學位論文 二卷과 單行本 一卷을 任意로 選定하였으며 이들 册子의 本文에 收錄되어 있는 單語들을 對象으로 分析하였다. 分析對象資料로서 碩士學位論文과 單行本을 選定한 理由는 分析對象單位인 單語의 量의 많고 적음에 따라 Zipf 公式의 變化與否를 구별하기 위해서이다.

分析對象資料인 碩士學位論文과 單行本은 아래와 같다.

資料Ⅰ : 李振榮, 大學圖書館의 利用率에 영향을 미치는 要因硏究, 서울 : 成均館大學校 大學院, 圖書館學科, 1976.

資料Ⅱ : 朴英熙, 書庫移動時機 및 書庫區分모델에 관한 硏究, 서울 : 成均館大學校, 大學院 圖書館學科, 1985.

資料Ⅲ : 鄭駁謨, 文獻情報學原論, 서울 : 아세아문화사, 1977.

分析對象單位인 單語의 分割은 1963年에 發表된 "學校文法統一案"을 基準으로 하였으며 同音異議語는 同一單語로 處理하였고 그리고 外國語(英語, 獨語, 日本等)와 숫자는 實驗對象에서 除外시켰다.

第2節 資料의 分析

以上의 方法으로 蒐集된 資料는 收錄頻度가 높은 單語에서 낮은 單語의 順으로, 그리고 收錄頻度가 같은 경우에는 가, 나, 다, 순으로 等級을 할당하여 各 資料의 等級一頻度表를 作成하였다.

다음의 表3은 資料Ⅰ의 等級一頻度를 나타낸 表이고 表4는 資料Ⅱ의 等級一頻度를 나타

27) Ju. K. Orlor, Why how and When does the Zipf-Mandelbrot law fail", J. Ling. Calc (1977) No.4, p.5.

낸 表이며 表5는 資料Ⅲ의 等級-頻度를 나타낸 表이다.(表의 나머지 부분은 부록7 참조)

表3, 表4, 表5의 첫번째 난은 等級을 나타내며 두번째 난은 頻度를, 그리고 세번째 난은 그 등급-빈도에 해당되는 單語를 기입한 것이다. 네번째 난은 頻度가 같은 單語들의 갯수를 기입하였으며 다섯번째 난은 等級이 1인 單語에서 부터의 단어수누계로 頻度가 같은 單語들이 있는 경우에는 합산하여 한번만 기입하였다. 여섯번째 난은 頻度에 單語數를 곱하여 한번만 기입하였다.

일곱번째 난은 여섯번째 난의 누계이며 여덟번째 난은 等級에 그 해당빈도를 곱하여 구한 수치를 기입한 난으로 頻度가 같은 單語들의 경우에는 그 해당등급들의 계급값을 곱하여 구한 수치를 기입하였다. 아홉번째 난은 여덟번째 난의 누계이며 열번째 난은 여덟번째 난의 로그값을 기입하였다. 그리고 열한번째 난은 單語總數를 100,000으로 환산하였을 때 다섯번째 난의 값을 기입한 난이며 열두번째 난도 일곱번째 난의 값을 위와같은 方法으로 구한 수치를 기입한 난이다. 따라서 이 두난을 통해서 異種單語總數中 몇 퍼센트가 單語總數中 몇 퍼센트를 차지하는데 알 수 있다.

가. Zipf 第一法則[28] 과의 比較分析

資料Ⅰ, Ⅱ, Ⅲ의 本文에 收錄된 單語들의 總數와 異種單語總數, 한단어의 평균수록수, 그리고 單語總數中 80%를 차지하는 異種單語들의 數와 比率은 다음의 表6과 같다.

表6에서 資料Ⅰ의 單語總數를 5,027개이고 異種單語總數는 1,341개이며, 따라서 한단어의 평균수록수는 3.75번이고 資料Ⅱ와 Ⅲ에 비해서 語彙의 選擇이 多樣함을 알 수 있다. 그리고 資料Ⅱ는 單語總數가 4,053개이고 異種單語總數는 1,011개이며 한 단어의 평균수록수가 4번으로 資料Ⅰ보다는 語彙의 選擇이 多樣하지 못하지만 單語總數중 80%를 차지하는 異種單語의 數가 264개로 異種單語總數의 約 26%를 차지해 80/20法則이 적용되는 것 같다. 資料Ⅲ은 單語總數가 50,023개이고 異種單語總數가 9,383개이며 한단어의 평균수록수가 5.33번으로 語彙選擇의 多樣性이 가장 낮지만 異種單語總數의 約 15%가 單語總數의 約 80%를 차지해 80/20法則보다 더 적은 數의 單語가 集中的으로 利用되었음을 알 수 있다.

表6) 各 資料의 收錄單語現況

單語總數	5,027	4,053	50,023
異種單語總數	1,341	1,011	9,383
單語當平均收錄數	3.75	4	5.33
單語總數中80%를	404	264	1,436
차지한異種單語數	(30%)	(26%)	(15%)

그림2는 資料Ⅰ의 等級을 X축, 頻度를 Y축으로 한 그래프로 쌍곡선과 유사한 형태를 가짐을 알 수 있다. 그리고 그림3은 그림2의 X축과 Y축에 해당하는 값의 로그값을 각각 X, Y축으로 한 散布度이다. 이 散布度는 相

28) 어떤 冊子의 本文에 收錄되어 있는 異種單語의 收錄頻度를 調査하여 收錄頻度가 높은 單語에서 낮은 單語의 順으로 등급할당했을 때 各 等級과 頻度사이에 다음과 같은 法則性이 存在한다.

$r \times f = c$

(r : 등급, f : 빈도, c : 특정책자에 해당되는 상수)

表7) 資料Ⅰ의 等級－頻度表

1	2	3	4	5	6	7	8	9	10	11	12
RANK	빈도	WORD	단어길이	단어수	(4)*(2)	(5) 빈도	(1)*(2)	(6) 빈도	LOG(3)	(5) 백분율	(7) 백분율
1	253	이	1	1	253	253	253.0	253.0	2.4116	7%	7.0%
2	242	의	1	2	242	500	484.0	742.0	2.6493	14%	9.9%
3	173	에	1	3	173	673	519.0	1,261.0	2.7331	22%	13.0%
4	169	은,는	1	4	169	842	676.0	1,937.0	2.6239	29%	16.7%
5	165	을	1	5	165	1,007	825.0	2,762.0	2.9164	37%	20.3%
6	113	하기만	1	6	113	1,120	678.0	3,440.0	2.8312	44%	24.4%

表8) 資料Ⅱ의 等級－頻度表

1	2	3	4	5	6	7	8	9	10	11	12
RANK	빈 도	W O R D	단어길이	단어수	(4)*(2)	(6) 빈도	(1)*(2)	(8) 빈도	LOG(3)	(5) 백분율	(7) 백분율
1	235	이	1	1	235	235	235.0	235.0	2.3710	9%	5.1%
2	155	은	1	2	155	390	310.0	545.0	2.6913	19%	9.5%
3	120	에	1	3	120	510	360.0	905.0	2.5563	29%	12.5%
4	106	을	1	4	106	616	424.0	1,329.0	2.6273	39%	15.1%
5	100	에	1	5	100	716	500.0	1,829.0	2.6989	49%	17.6%
6	82	는	1	6	82	798	492.0	2,321.0	2.6919	59%	19.6%
7	66	과	1	7	66	864	462.0	2,783.0	2.6646	69%	21.3%
8	52	기	2	9	104	968	442.0	3,225.0	2.6454	79%	23.8%
9	52	은					442.0	3,667.0	2.6454		

表9) 資料Ⅲ의 等級－頻度表

1	2	3	4	5	6	7	8	9	10	11	12
RANK	빈 도	W O R D	단어길이	단어수	(4)*(2)	(5) 빈도	(1)*(2)	(6) 빈도	LOG(3)	(5) 백분율	(7) 백분율
1	3,010	이	1	1	3,010	3,010	3,010.0	3,010.0	3.4785	11	6.17
2	1,865	의	1	2	1,865	4,875	3,730.0	6,740.0	3.5717	21	9.7%
3	1,516	은	1	3	1,516	6,391	4,549.0	11,298.0	3.6578	32	12.7%
4	1,420	에	1	4	1,420	7,811	5,680.0	16,978.0	3.7543	43	15.5%
5	1,143	을	1	5	1,143	8,954	5,715.0	22,693.0	3.7570	53	17.6%
6	1,079	가	1	6	1,079	10,033	6,474.0	29,157.0	3.8111	64	20.7%

關係數가 -0.96772로 相關이 아주 높으며 SPSS를 利用해 전산처리한 결과 다음의 회귀 방정식을 구했다.

$$Y = -0.85137 \times + 2.5105$$

위의 回歸方程式은 標準誤差가 0.0952이고 決定 係數가 0.93648이므로 回歸方程式에 의한 豫測이 매우 正確함을 알 수 있다.

이 回歸方程式에서 Y는 頻度의 로그값을 X는 等級의 로그값을 意味하므로 X, Y대신 대입하면 다음과 같은 公式을 誘導할 수 있다.

$$R^{0.85137} \cdot F = 323.971 \qquad (G)$$

그림4는 資料Ⅱ의 等級과 頻度를 各各 X軸과 Y軸으로 한 그래프이고 그림5는 그림4의 X와 Y의 값을 로그값으로 변화시켜 구한 散布度이다.

그림5의 散布度는 相關係數가 -0.9763으로 그림3의 相關係數보다 절대값이 1에 더 근사하므로 더 相關이 높다. 이 散布度의 回歸方程式은 다음과 같다.

$$Y = -0.92495 \times + 2.64016$$

이 回歸方程式은 標準誤差가 0.0877이고 決定係數가 0.95327이므로 回歸方程式에 의한 豫測이 資料Ⅰ보다 더 正確하다. 이 回歸方程式의 X대신에 logR을 Y대신에 logF를 대입하면 다음과 같다.

$$R^{0.92495} \cdot F = 436.6807 \qquad (H)$$

그림6은 資料Ⅲ의 等級과 頻度를 各各 X, Y軸으로 한 그래프이며 그림7은 그림6의 X와 Y의 값을 로그값으로 변화시켜 구한 散布度이다.

이 散布度는 相關係數가 -0.96763으로 다른 資料보다 더 相關이 높으며 回歸方程式은 다음과 같다.

$$Y = -0.87815 \times + 3.32711$$

위의 回歸方程式은 標準誤差가 0.09922이고 決定係數가 0.9363이므로 豫測이 正確함을 알 수 있다. 이 回歸方程式에서 X대신에 logR을 Y대신에 logF를 대입하면 다음과 같다.

$$R^{0.87815} \cdot F = 2123.781 \qquad (I)$$

以上의 資料Ⅰ, Ⅱ, Ⅲ에서 誘導된 公式(G), 公式(H), 公式(I)는 Zipf 第一法則인 $R \cdot F = C$와는 다르지만 일반화된 Zipf 第一法則인 $R^B \cdot F = C$와는 일치한다.(여기서 B는 상수) 따라서 한글단어의 收錄頻度와 그 等級사이에도 一定한 統計的인 法則性이 存在함을 알 수 있다.

그러나 G.K. Zipf가 主張한 바와 같이 公式(G), 公式(H), 公式(I)는 收錄頻度가 낮은 單語에 있어서는 실제 자료에서 구한 값과 상당한 차이가 있었다. 따라서 收錄頻度가 낮은 單語들의 分布가 Zipf 第二法則과 일치하는지 比較分析해야할 必要가 있다.

나. Zipf 第二法則[29] 과의 比較分析

아래의 表7, 8, 9는 各 資料에서 구한 收錄頻度가 낮은 單語들과 Zipf 第二法則, 一般化

29) G.K. Zipf는 收錄頻度가 낮은 單語들의 分布를 正確하게 說明하기 위해 다음의 공식을 유도했다.

$$I_n / I_1 = \frac{3}{4n^2 - 1}$$

(I_n: 빈도 n인 단어수, n : 빈도)

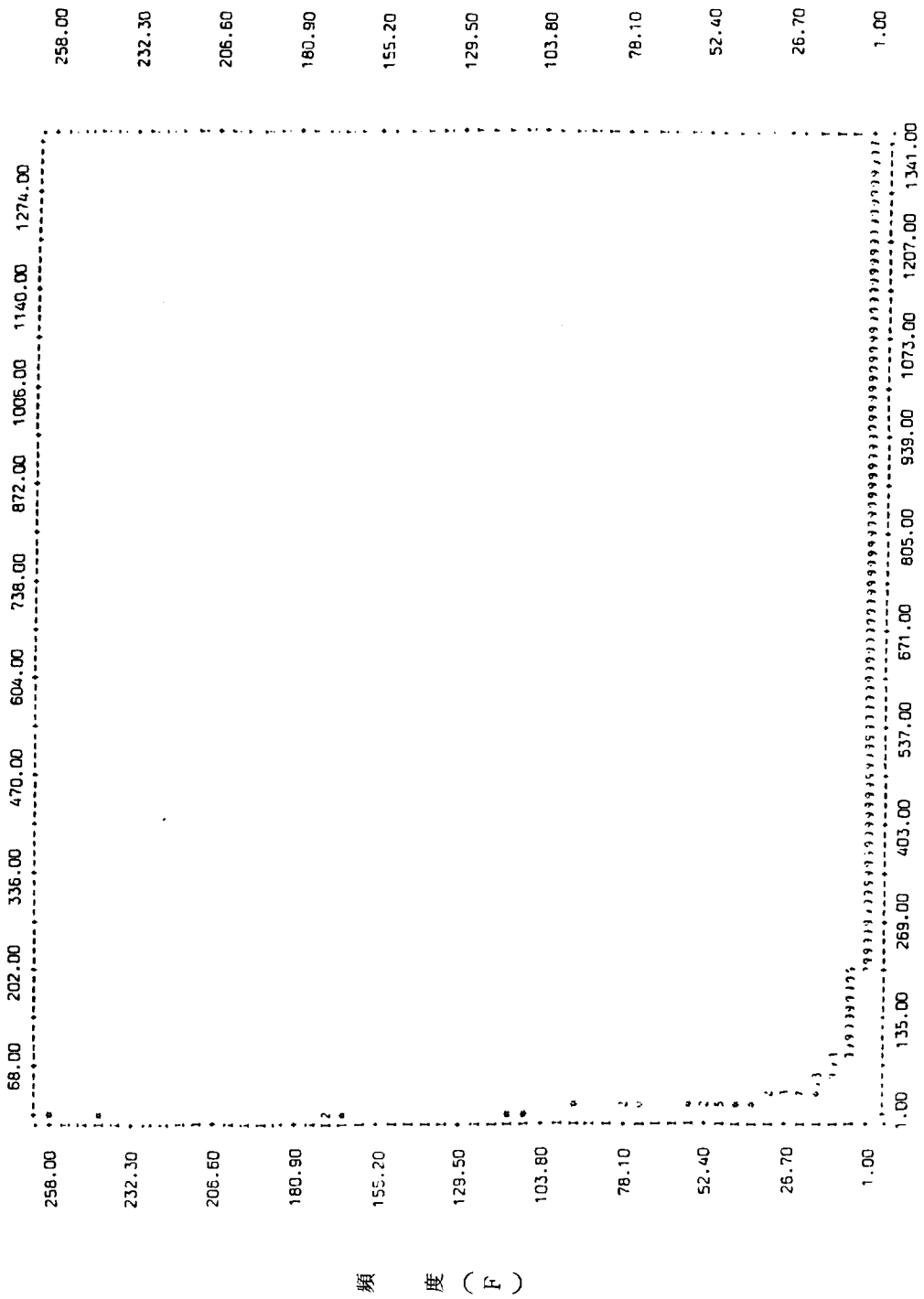


그림 2) 資料 I의 等級-頻度그래프

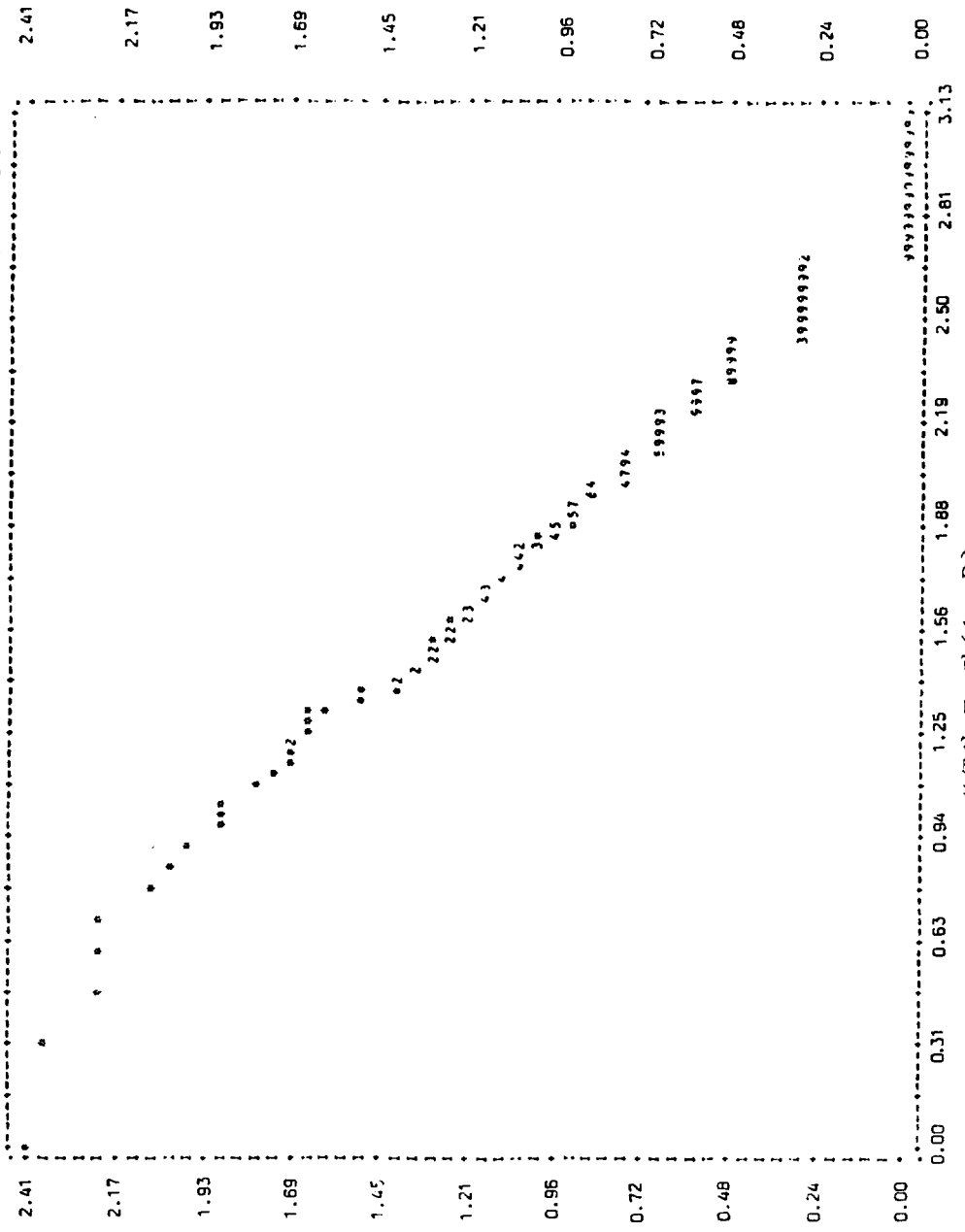


그림 3) 資料 I 의 等級의 로그값-頻度의 로그값의 散布度

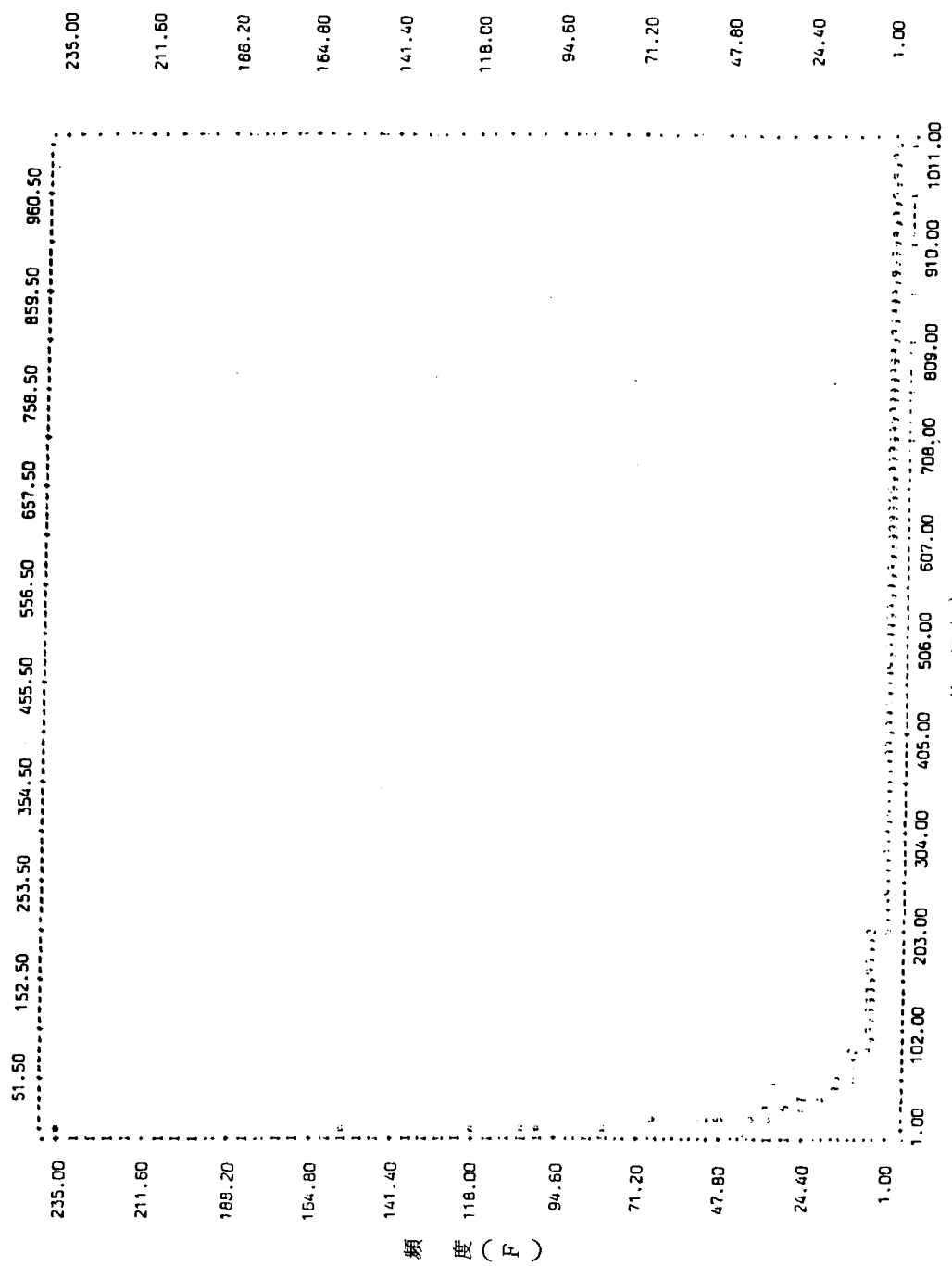


그림 4) 資料 II의 等級-頻度그래프

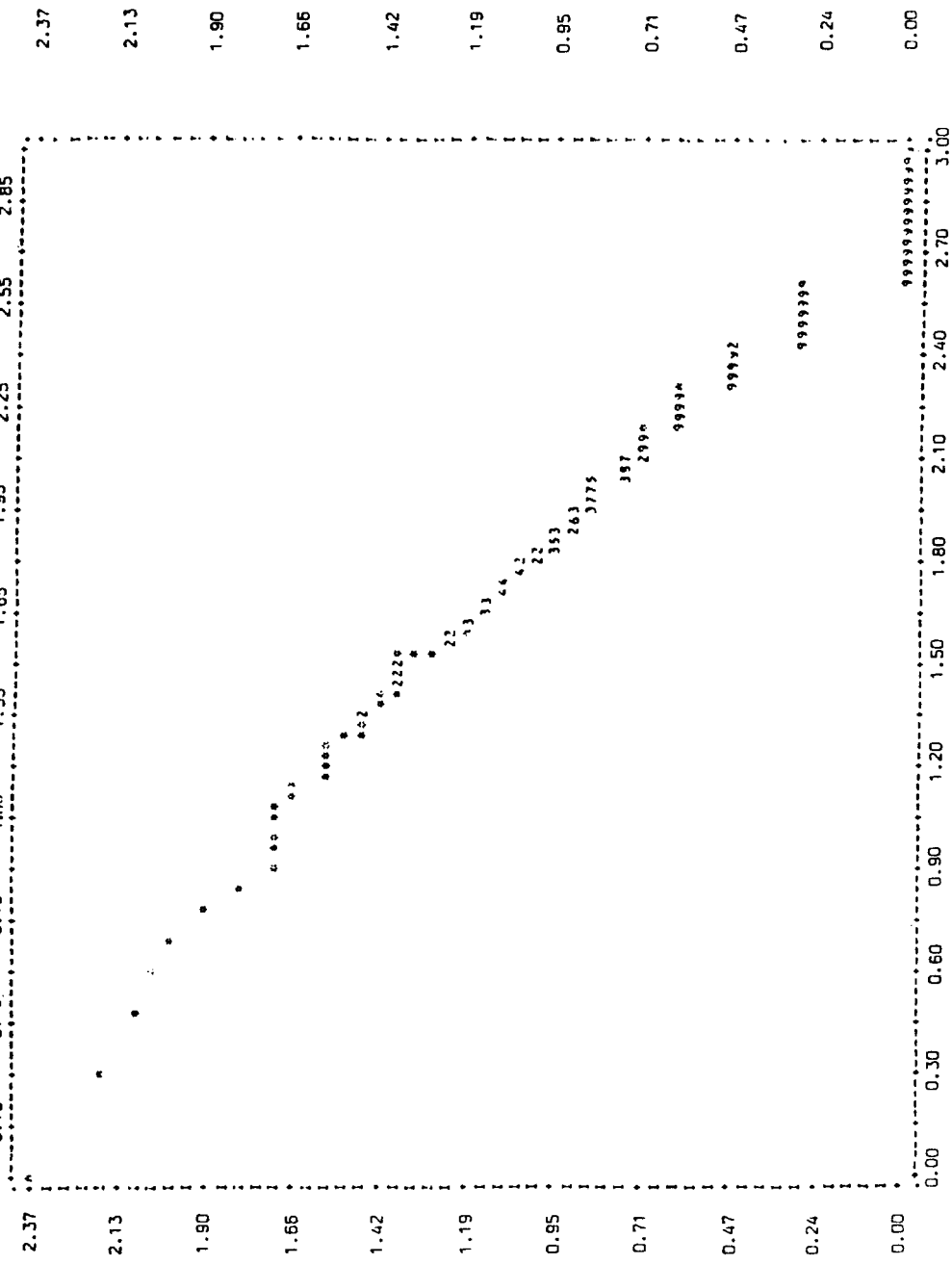


그림 5) 資料Ⅱ의 等級의 로그값-頻度의 로그값의 散布度

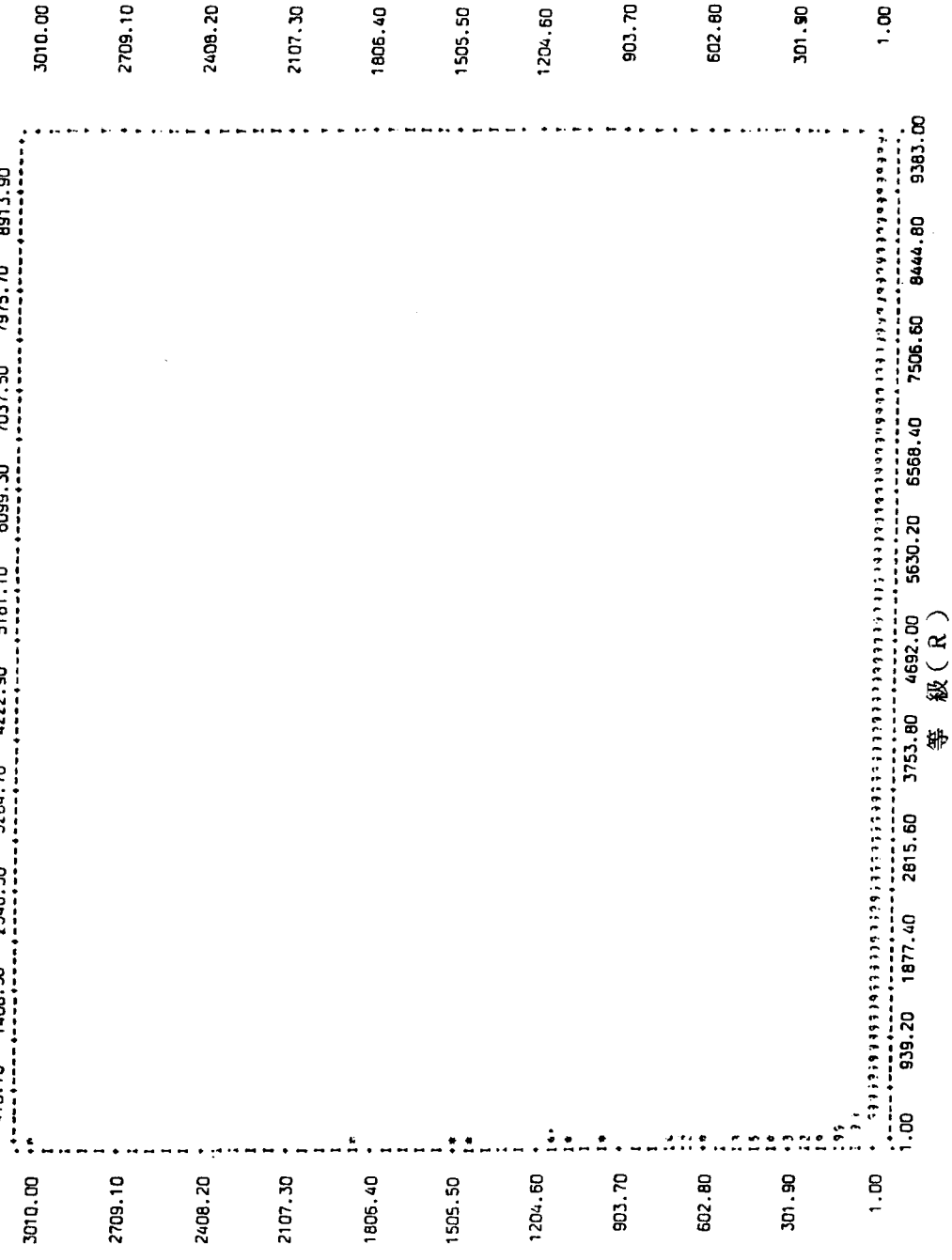


그림 6) 資料Ⅲ의 等級一散布그래프

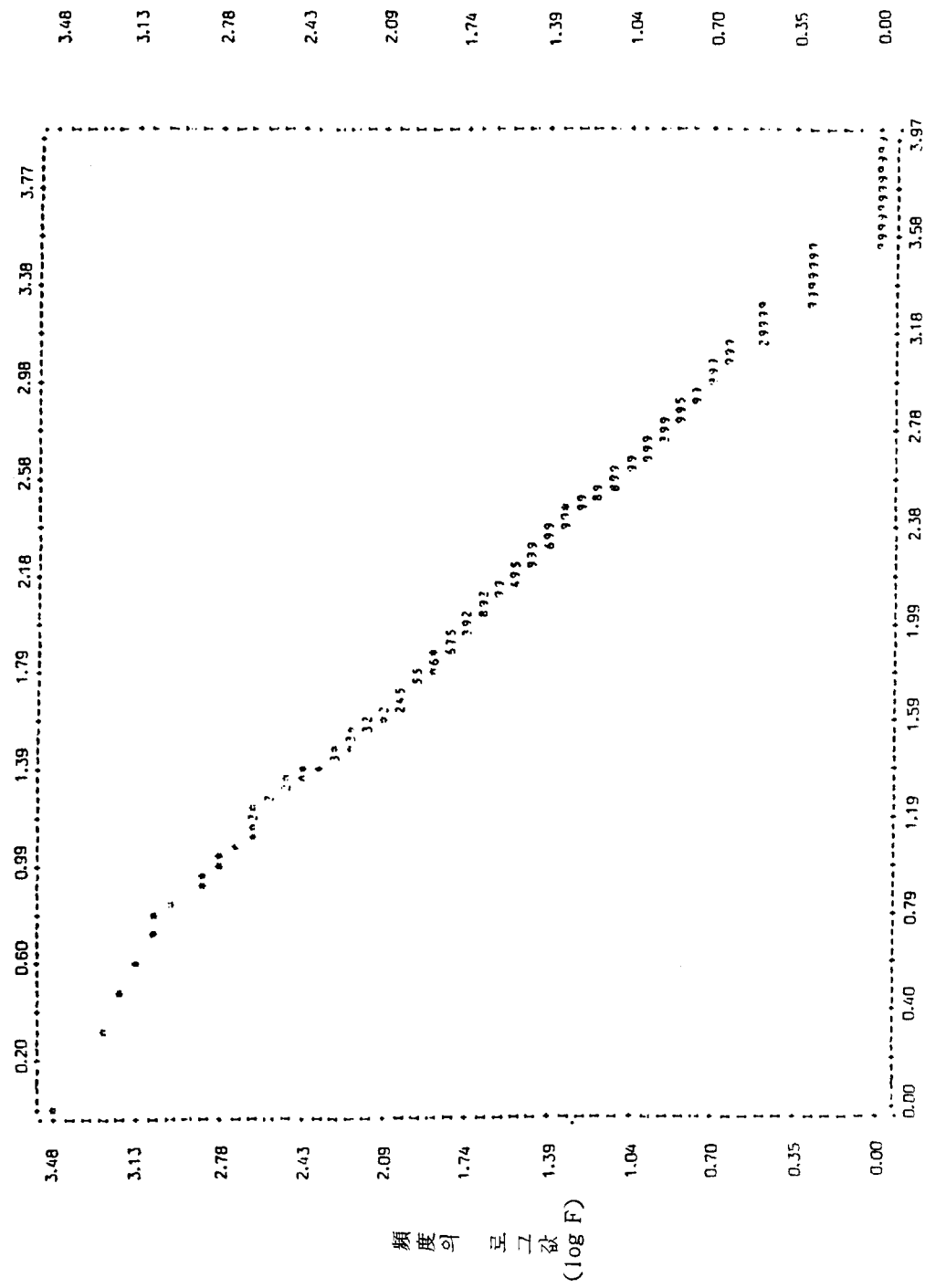


그림 7) 資料Ⅲ의 等級의 로그값-頻度의 로그값의 散布度

된 Zipf 第二法則[30]에 의해 구한 수치를 비교한 표이다.

表7에 있어서 첫번째 난중 "비율"난은 실제 자료에서 구한 수치로서 各 頻度에 해당되는 單語數를 收錄頻度가 1인 單語數로 나눈 수치이고 "단어수"난은 各 收錄頻度에 해당되는 單語數를 기입한 난이다. 두번째 난중 "비율"난은 Zipf 第二法則에 의해 구한 비율이고 "단어수"난은 이 비율에 1번 數錄된 單語數를 곱하여 구한 수치이다. 세번째 난중 "비율"난은 一般化된 Zipf 第二法則에 의한 비율을 의미하고 "단어수"난은 이 비율에 1번 數錄된 單語數를 곱하여 구한 수치이다.

표7) 자료Ⅰ의 수치와 Zipf 第二法則에 의한 수치 비교표

빈 도	자 료 Ⅰ		Zipf 第二法則		一般化된 Zipf 第二法則	
	비 율	단어수	비 율	단어수	비 율	단어수
1	1	867	1	867	1	867
2	0.22145	192	0.2	173.4	0.30136	261.2
3	0.08881	77	0.08571	74.3	0.08124	122.8
4	0.04844	42	0.04761	41.3	0.05226	70.4
5	0.04498	39	0.03030	26.3	0.05226	45.3
6	0.02768	24	0.02097	18.2	0.03624	31.4
7	0.01153	10	0.01538	13.3	0.02650	23.
8	0.01499	13	0.01176	10.2	0.02016	17.5
9	0.01036	9	0.00928	8.1	0.01582	13.7
10	0.00461	4	0.00751	6.5	0.01272	11.0
11	0.00692	6	0.00621	5.4	0.0104	9.0
12	0.00461	4	0.00521	4.5	0.0087	7.5
13	0.00461	4	0.00444	3.8	0.00735	6.4
14	0.00346	3	0.00383	3.3	0.00629	5.5
15	0.00461	4	0.00333	2.9	0.00544	4.7

表8, 9도 이상과 같은 方法으로 資料Ⅱ와 資料Ⅲ에서 구한 표이다.

30) A.D. Booth에 의해 수정된 Zipf 第二法則을 의미하며 公式은 다음과 같다.

$$I_n/I_1 = \frac{2}{n(n+1)} \qquad B = 1$$

$$I_n = (kT)^{\frac{1}{B}}\{(\frac{1}{n})^{\frac{1}{B}} - (\frac{1}{(n+1)})^{\frac{1}{B}}\}$$

$B \neq 1$ K : 상수

이것을 일반화된 Zipf 第二法則이라고 한다.

표8) 자료Ⅱ의 수치와 Zipf 第二法則에 의한 수치 비교표

빈 도	자 료 Ⅱ		Zipf 第二法則		一般化된 Zipf第二法則	
	비 율	단어수	비 율	단어수	비 율	단어수
1	1	603	1	603	1	603
2	0.2388	144	0.2	120.6	0.3181	191.8
3	0.10779	65	0.08571	51.7	0.15455	93.2
4	0.0796	48	0.04761	28.7	0.09080	54.8
5	0.03648	22	0.03030	18.3	0.05954	35.9
6	0.02985	18	0.02097	12.7	0.04195	25.3
7	0.03648	22	0.01538	9.3	0.03109	18.8
8	0.01824	11	0.01176	7.1	0.02394	14.4
9	0.01824	11	0.0928	5.6	0.01898	11.4
10	0.00663	4	0.00751	4.5	0.01540	9.3
11	0.00995	6	0.00621	3.7	0.01274	7.7
12	0.00331	2	0.00521	3.1	0.01107	6.5
13	0.00995	6	0.00444	2.7	0.00912	5.5
14	0.00995	6	0.00383	2.3	0.00785	4.7
15	0.00497	3	0.00333	2.0	0.00683	4.1

표9) 자료Ⅲ의 수치와 Zipf 第二法則에 의한 수치 비교표

빈 도	자 료 Ⅲ		Zipf 第二法則		一般化된 Zipf第二法則	
	비 율	단어수	비 율	단어수	비 율	단어수
1	1	6087	1	6087	1	6087
2	0.21981	1338	0.2	1217.4	0.3076	1872.9
3	0.08575	522	0.08571	521.7	0.1464	891.6
4	0.05191	316	0.04761	289.9	0.08478	516.1
5	0.03449	210	0.03030	184.5	0.05494	334.5
6	0.02414	147	0.02097	127.7	0.03833	233.4
7	0.01938	118	0.01538	93.6	0.02818	171.5
8	0.01182	72	0.01176	71.6	0.02153	131.1
9	0.00919	56	0.00928	56.5	0.01696	103.3
10	0.00919	56	0.00751	45.8	0.01368	83.3
11	0.00673	41	0.00621	37.8	0.01126	68.6
12	0.00229	14	0.00521	31.8	0.00942	57.3
13	0.00410	25	0.00444	27.1	0.00798	48.6
14	0.00410	25	0.00383	23.3	0.00685	41.7
15	0.00197	12	0.00333	20.3	0.00594	36.2

그리고 이들 수치중 "비율"난의 수치는 소수점이하 다섯자리까지만 취했으며 "단어수"난은 소수점이하 둘째자리에서 반올림한 값이다.

表7, 表8, 表9를 통해서 資料Ⅰ, 資料Ⅱ, 資料Ⅲ의 實際 資料에서 구한 수치와 Zipf 第二法則에 의해서 구한 수치, 그리고 一般化된 Zipf第二法則에 의해서 구한 수치사이에 상당한 차이가 있음을 알 수 있다.

그러므로 한글단어중 수록빈도가 낮은 단어들에 적용할 수 있는 公式을 誘導할 必要性이 있다. 따라서 單語의 收錄頻度와 單語數사이의 關係에 영향을 미친다고 생각하는 要因으로 單語總數(T), 1번 收錄된 單語數(W_1), 그리고 異種單總數(D)를 選擇하여 各 要因과의 關係를 分析해 보았다.

(1) 單語總數와의 關係分析

아래의 表10은 資料Ⅰ, Ⅱ, Ⅲ에서 구한 收錄頻度가 낮은 單語들의 數와 各 頻度에 해당되는 單語數를 單語總數T로 나눈 수치를 기입한 表이다. 分析範圍는 收錄頻度가 1인 單語로부터 收錄頻度가 10인 單語들 까지로 한정했다. 왜냐하면 頻度가 높아질수록 單語의 數는 同一한 값인 1을 가지게 되므로 결국 Zipf第一法則과 같이 單語數가 1인 해당빈도가 많아지게 되어 이 부분에 적합한 공식이 다시 필요해지기 때문이다.

다음의 그림 8, 10, 12는 各各 資料Ⅰ, Ⅱ, Ⅲ에 나타난 頻度를 X軸으로 하고 그리고 各 頻度에 해당되는 單語數를 單語總數로 나눈 수치(Wn/T)를 Y軸으로 한 그래프이며 그림9, 10,

표 10) 收錄頻度가 낮은 單語들의 單語總數에 대한 비율 및 單語數

빈 도	자 료 Ⅰ		자 료 Ⅱ		자 료 Ⅲ	
	비 율	단어수	비 율	단어수	비 율	단어수
단 어 총 수		5,027		4,053		50,023
1(W1/T)	0.17246	867	0.14877	603	0.12168	6,087
2(W2/T)	0.03819	192	0.03552	144	0.02674	1,338
3(W3/T)	0.01531	77	0.01603	65	0.01043	552
4(W4/T)	0.00835	42	0.01184	48	0.00631	316
5(W5/T)	0.00775	39	0.00542	22	0.00419	210
6(W6/T)	0.00477	24	0.00444	18	0.00293	147
7(W7/T)	0.00198	10	0.00542	22	0.00235	118
8(W8/T)	0.00258	13	0.00271	11	0.00143	72
9(W9/T)	0.00179	9	0.00271	11	0.00111	56
10(W10/T)	0.00079	4	0.00098	4	0.00111	56

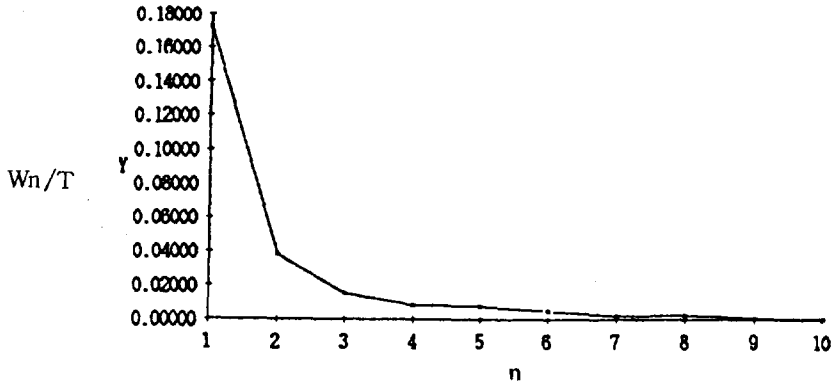


그림 8) 자료 I의 Wn/T대 n의 그래프

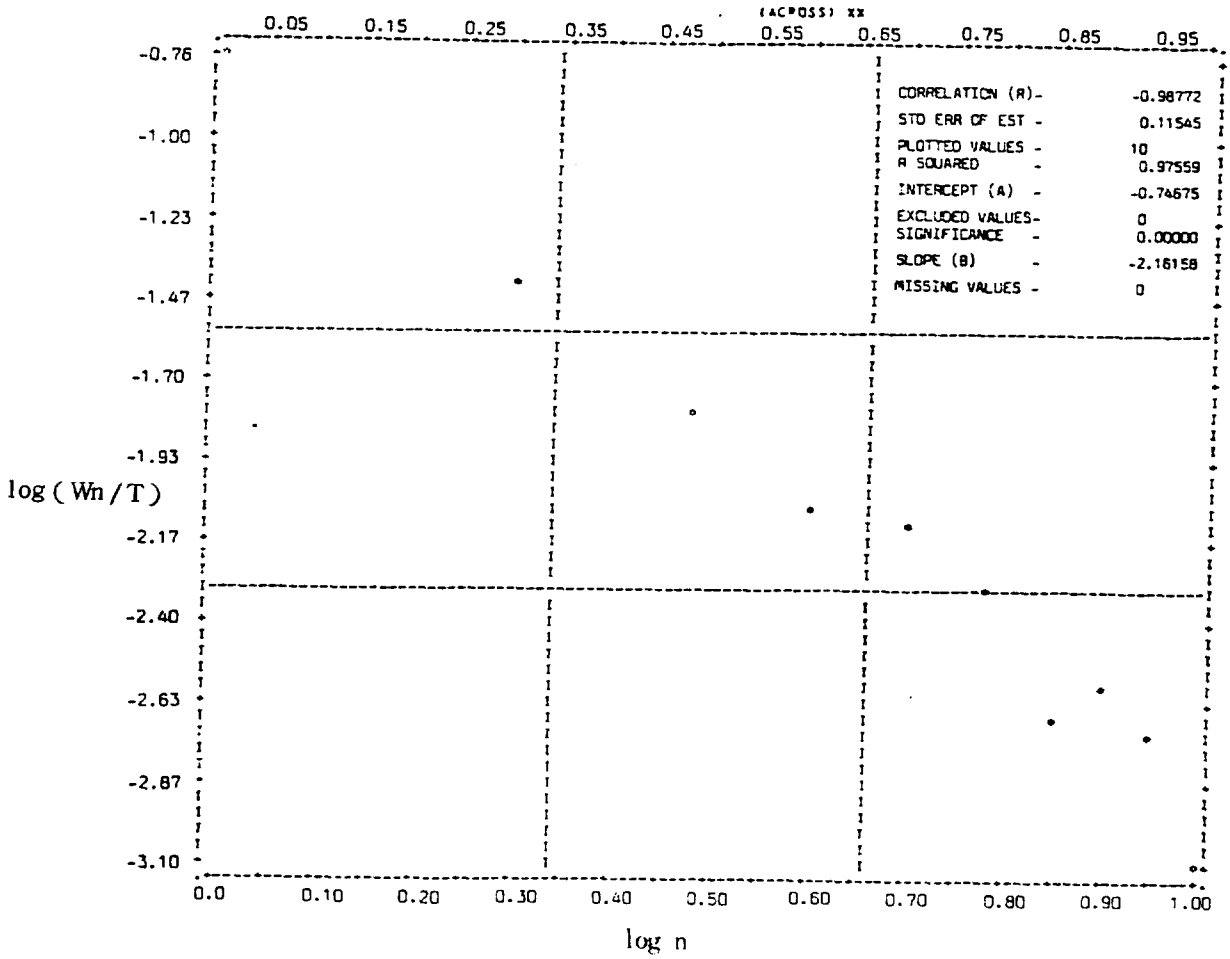


그림 9) 자료 I의 log(Wn/T)대 log n의 산표도

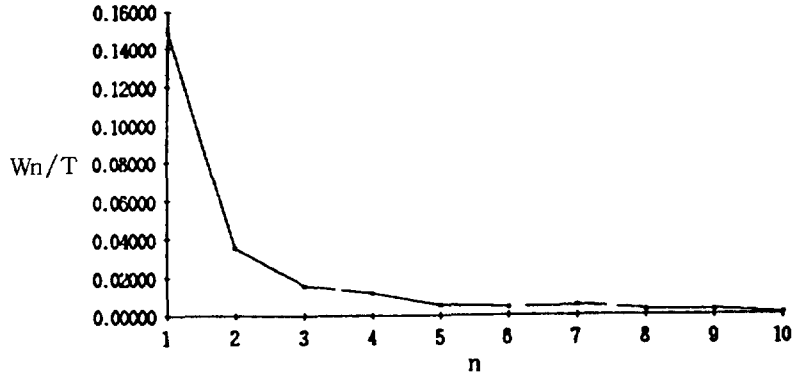


그림 10) 자료Ⅱ의 Wn/T대 n의 그래프

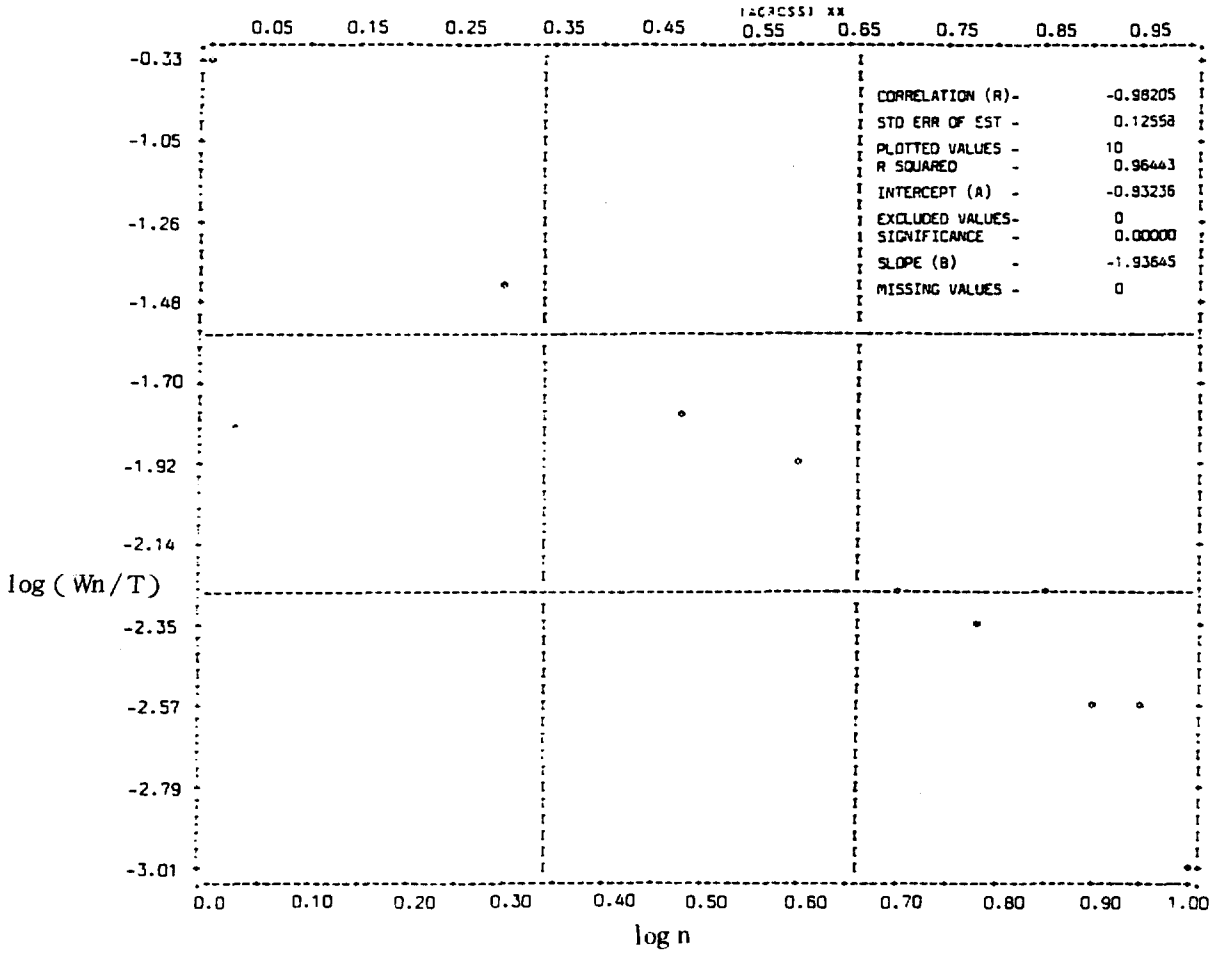


그림 11) 자료Ⅱ의 log(Wn/T)대 log n의 산포도

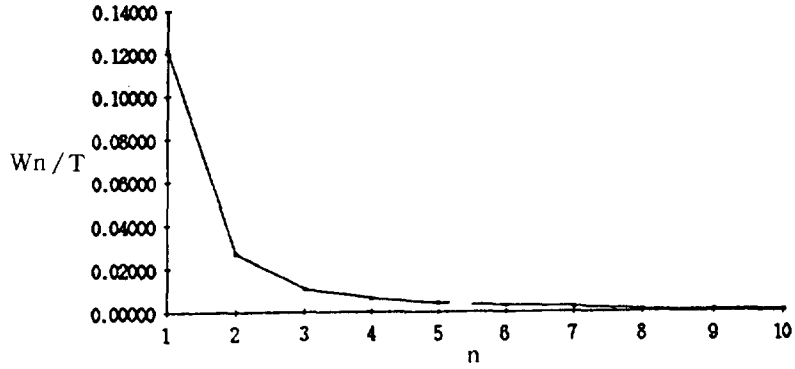


그림12) 자료Ⅲ의 Wn/T대 n의 그래프

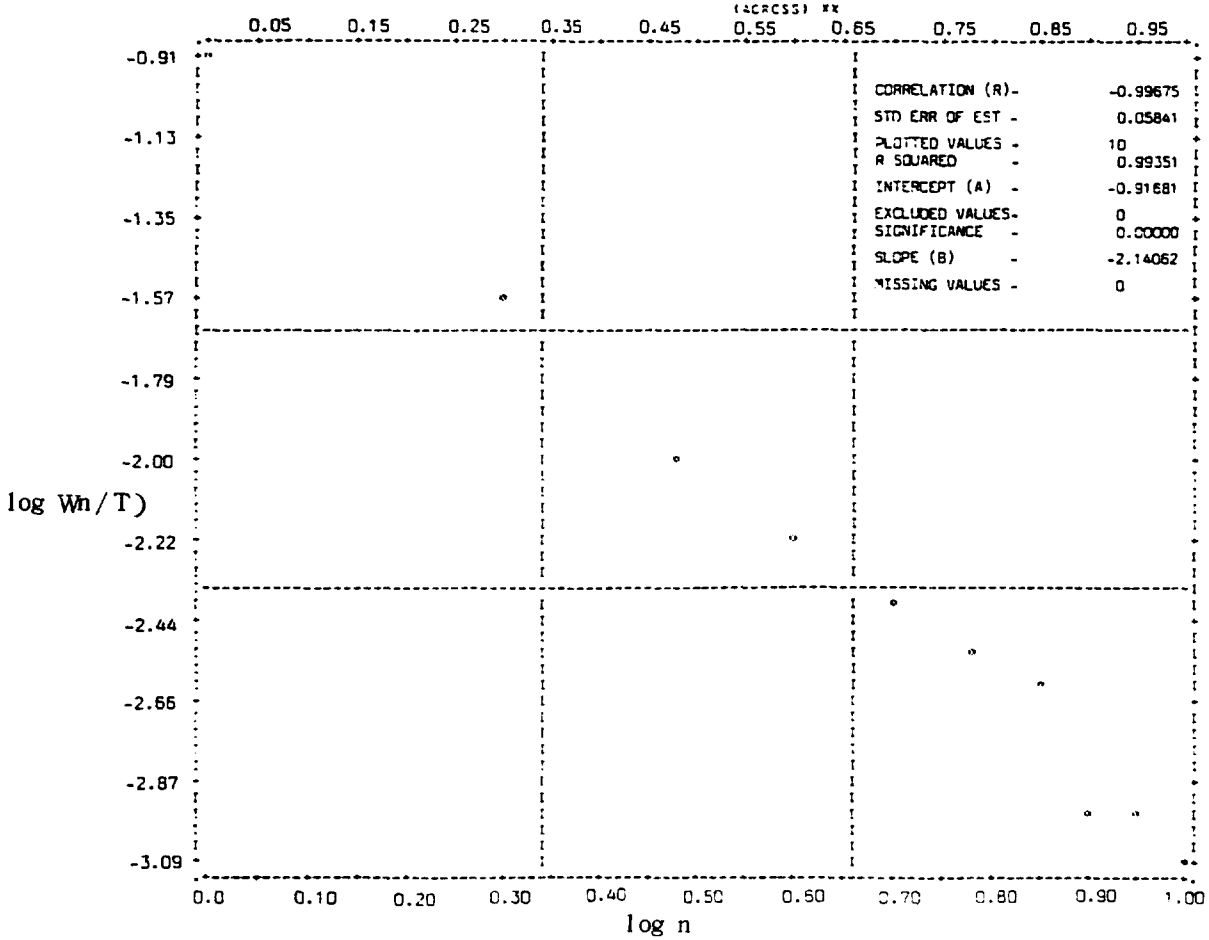


그림 13) 자료Ⅲ의 log(Wn/T)대 log n의 산포도

13은 그림 8,10,12의 X와 Y의 값의 로그값을 각각 X軸, Y軸으로 하여 그린 散布度이다.

그림 9, 11, 13의 散布度에서 구한 回歸方程式은 다음과 같다.

資料Ⅰ : $Y = -2.16158 \times -0.74675$

資料Ⅱ : $Y = -1.93645 \times -0.83236$

資料Ⅲ : $Y = -2.14062 \times -0.91681$

이 回歸方程式에서 X대신에 logn을, Y대신에 log(Wn/T)를 대입하면 다음과 같다.

資料Ⅰ : $Wn/T = 0.17915/n^{2.16158}$ (J-1)

資料Ⅱ : $Wn/T = 0.14711/n^{1.93645}$ (J-2)

資料Ⅲ : $Wn/T = 0.12111/n^{2.14062}$ (J-3)

여기서 n은 收錄頻度를, Wn은 收錄頻度가 n인 單語의 數를 그리고 T는 單語總數를 各各 意味한다.

(2) 收錄頻度 1 인 單語數와의 關係分析

아래의 表 11는 資料Ⅰ, Ⅱ, Ⅲ의 收錄頻度 1에서 10까지에 해당되는 單語數와 이 單語數를 收錄頻度가 1인 單語數로 나눈 수치를 기입한 표이다.

그리고 그림 14, 16, 18은 X軸에 頻度를, Y軸에 各 單語에 해당되는 單語의 數를 數錄頻度가 1인 單語의 數를 나눈 수치 (Wn/W_1)를 나타낸 그래프이고 그림 15, 17, 19는 그림 14, 16, 18의 X와 Y의 값의 로그값을 各各 X軸과 Y軸으로 한 散布度이다.

표 11) 收錄頻度가 낮은 單語들의 收錄頻度 1인 單語數에 대한 비율 및 單語數

빈 도	자 료 Ⅰ		자 료 Ⅱ		자 료 Ⅲ	
	비 율	단어수	비 율	단어수	비 율	단어수
1번수록된 단어수		867		603		6087
1(W1/W1)	1	867	1	608	1	6087
2(W2/W1)	0.22145	192	0.2388	144	0.21981	1338
3(W3/W1)	0.08881	77	0.10779	65	0.08575	522
4(W4/W1)	0.04844	42	0.0796	48	0.05191	316
5(W5/W1)	0.04498	39	0.03648	22	0.03449	210
6(W6/W1)	0.02768	24	0.02985	18	0.02414	147
7(W7/W1)	0.01153	10	0.03648	22	0.01938	118
8(W8/W1)	0.01499	13	0.01824	11	0.01182	72
9(W9/W1)	0.01038	9	0.01824	11	0.00919	56
10(W10/W1)	0.00461	4	0.00663	4	0.00919	56

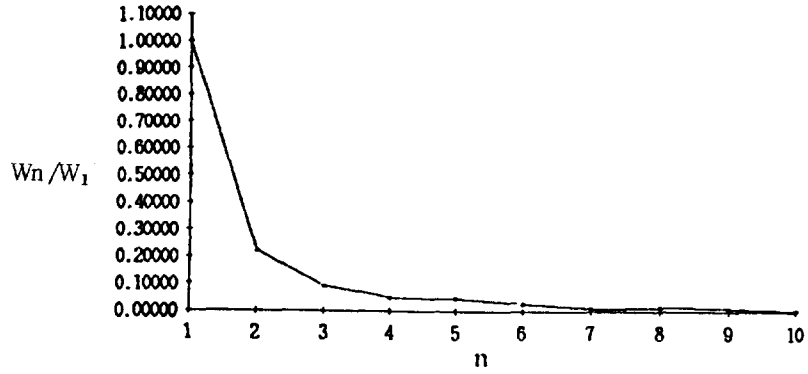


그림 14) 자료 I의 Wn/W_1 대 n의 그래프

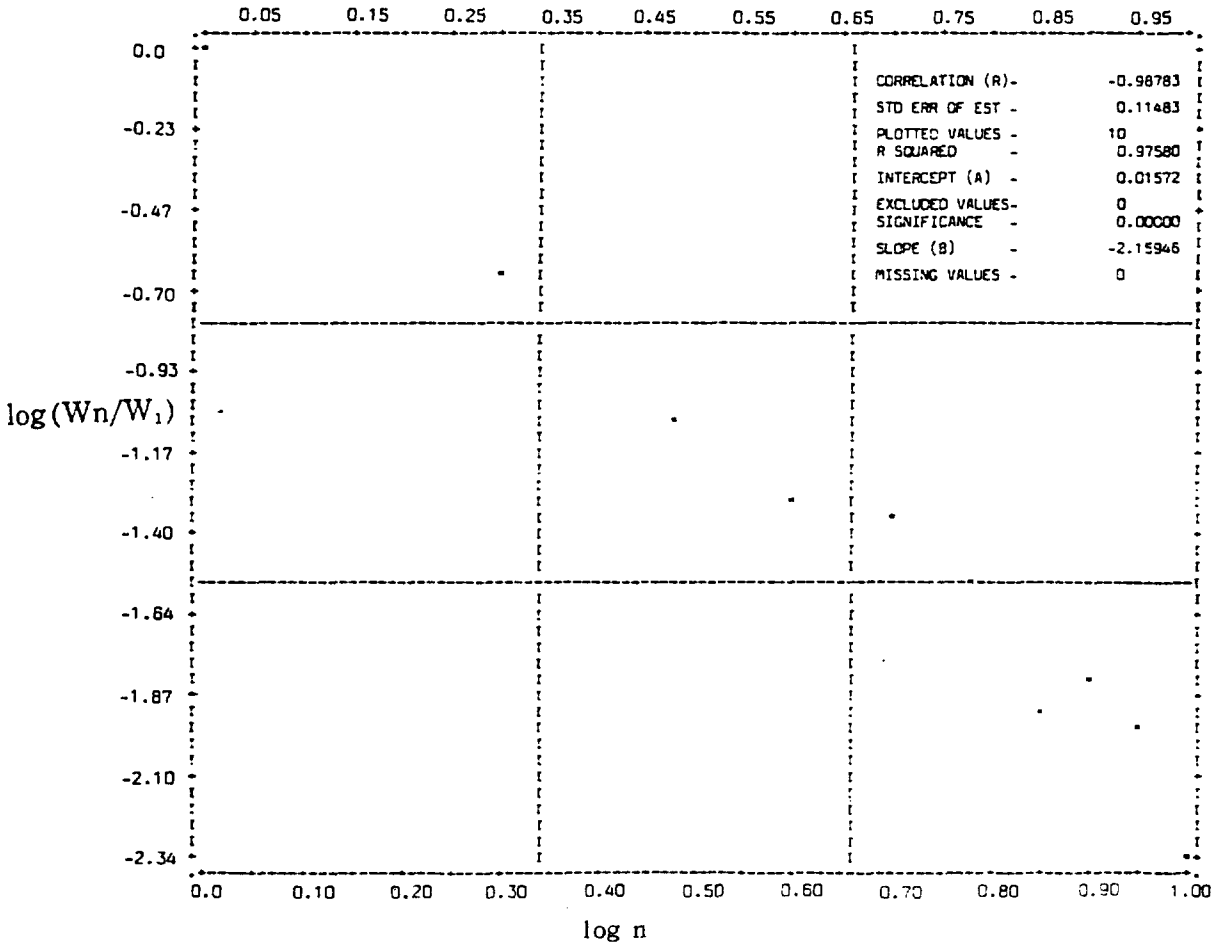


그림 15) 자료 I의 $\log(Wn/W_1)$ 대 log n의 산포도

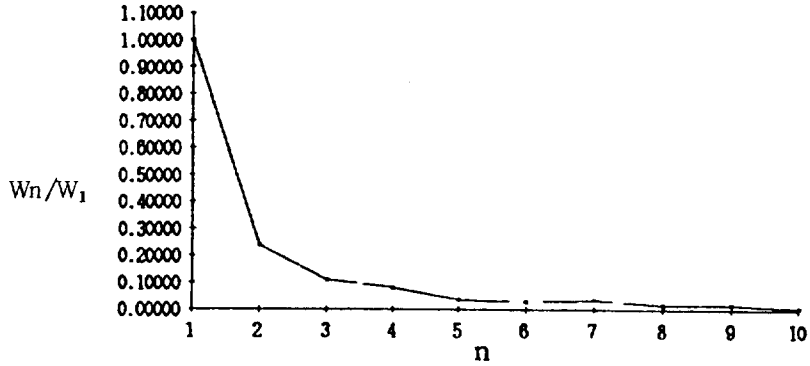


그림 16) 자료Ⅱ의 Wn/W_1 , 대 n의 그래프

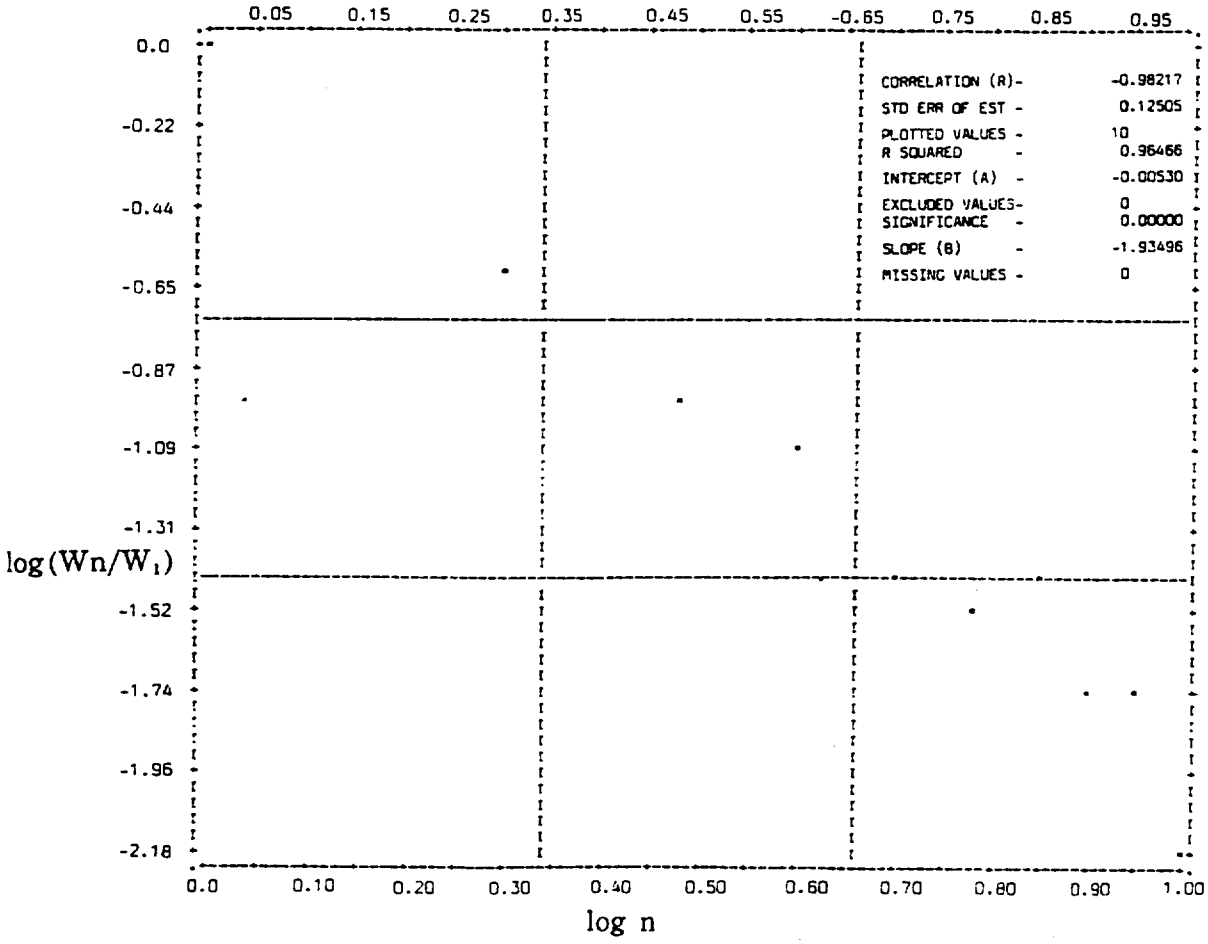


그림 17) 자료Ⅱ의 $\log(Wn/W_1)$ 대 log n의 산포도

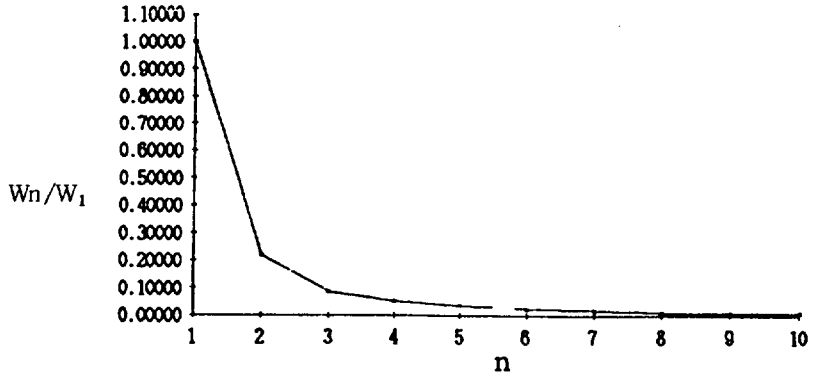


그림 18) 자료Ⅲ의 Wn/W_1 , n의 그래프

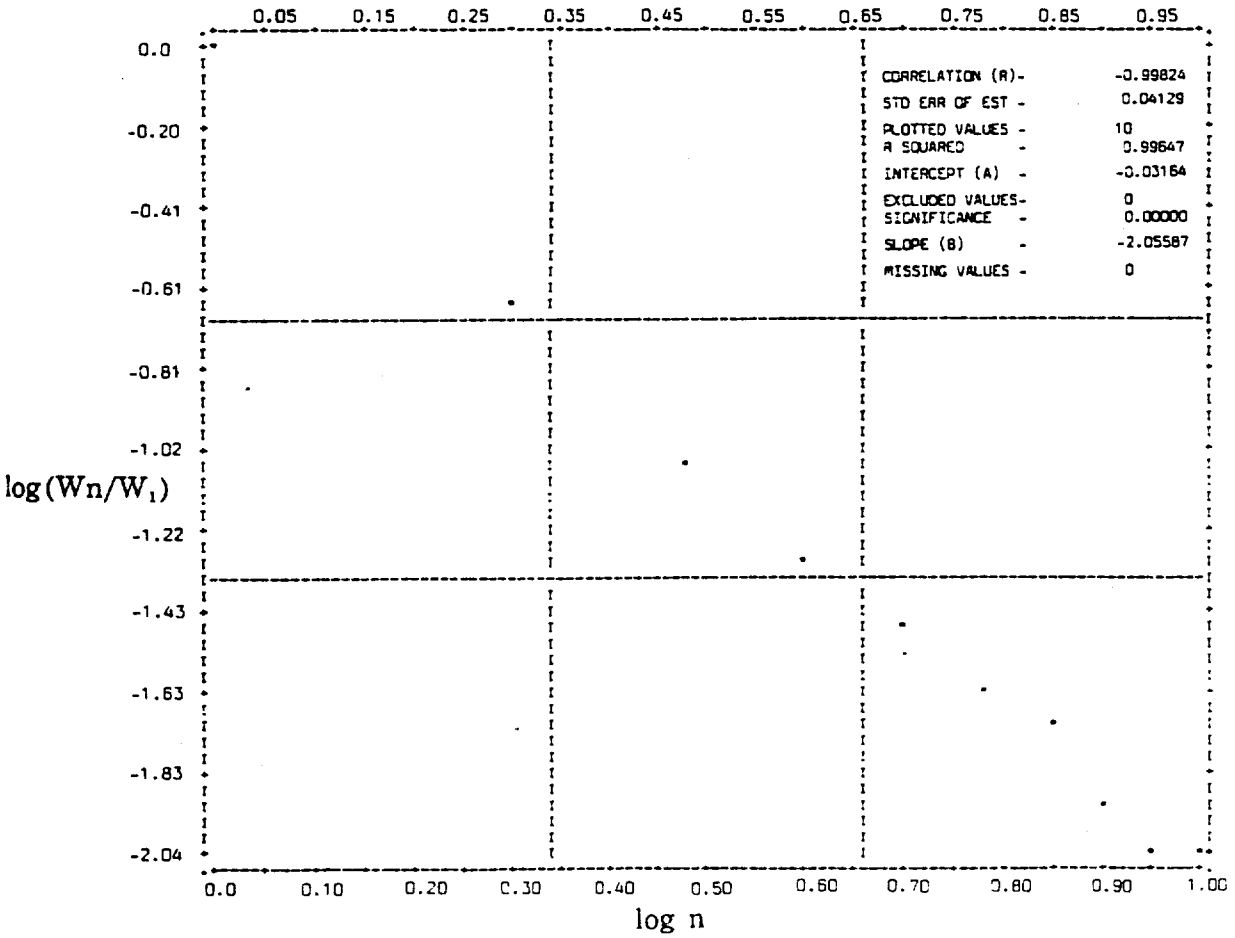


그림 19) 자료Ⅲ의 $\log(Wn/W_1)$대 log n의 산포도

그림 15, 17, 19의 散布度에서 구한 回歸方程式은 다음과 같다.

資料Ⅰ : $Y = -2.15946 \times + 0.01572$
資料Ⅱ : $Y = -1.93496 \times - 0.0053$
資料Ⅲ : $Y = -2.05587 \times - 0.03164$

이 回歸方程式에서 X대신에 logn을, Y대신에 $\log(Wn/W_1)$을 대입하면 다음과 같다.

資料Ⅰ : $Wn/W_1 = 1.03686/n^{2.15946}$ (K-1)
資料Ⅱ : $Wn/W_1 = 0.98787/n^{1.93496}$ (K-2)
資料Ⅲ : $Wn/W_1 = 0.92973/n^{2.05587}$ (K-3)

③ 異種單語總數(D)와의 關係分析

아래의 表 12는 資料Ⅰ, Ⅱ, Ⅲ의 收錄頻度가 낮은 單語들의 數와 各 頻度에 해당되는 單語數을 異種單語總數로 나눈 수치를 나타낸 표이다.

그림 20, 22, 24는 各各 資料Ⅰ, Ⅱ, Ⅲ의 收錄頻度를 X축에, 그리고 各 收錄頻度에 해당되는 單語數를 異種單語總數로 나눈 수치를 Y축에 나타낸 그래프이며 그림 21, 23, 25는 그림 20, 22, 24의 X와 Y값의 로그값을 各各 X軸, Y軸에 나타낸 散布度이다.

표 12) 收錄頻度가 낮은 單語들의 異種單語總數에 대한 비율 및 單語數

자료 / 빈도	자료 Ⅰ		자료 Ⅱ		자료 Ⅱ	
	비율	단어수	비율	단어수	비율	단어수
이종단어총수		1,341		1,011		9,383
1(W1/D)	0.64653	867	0.59643	603	0.64872	6,087
2(W2/D)	0.14317	192	0.14243	144	0.14259	1,338
3(W3/D)	0.05741	77	0.06429	65	0.05563	522
4(W4/D)	0.03131	42	0.04747	48	0.03367	316
5(W5/D)	0.02908	39	0.02176	22	0.02238	210
6(W6/D)	0.01789	24	0.01780	18	0.01566	147
7(W7/D)	0.00745	10	0.02176	22	0.01257	118
8(W8/D)	0.00969	13	0.01088	11	0.00767	72
9(W9/D)	0.00671	9	0.01088	11	0.00596	56
10(W10/D)	0.00298	4	0.00395	4	0.00596	56

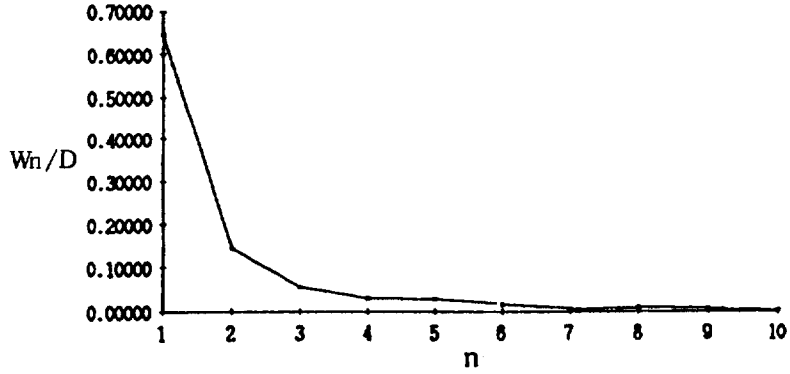


그림 20) 자료 I의 Wn/D 대 n의 그래프

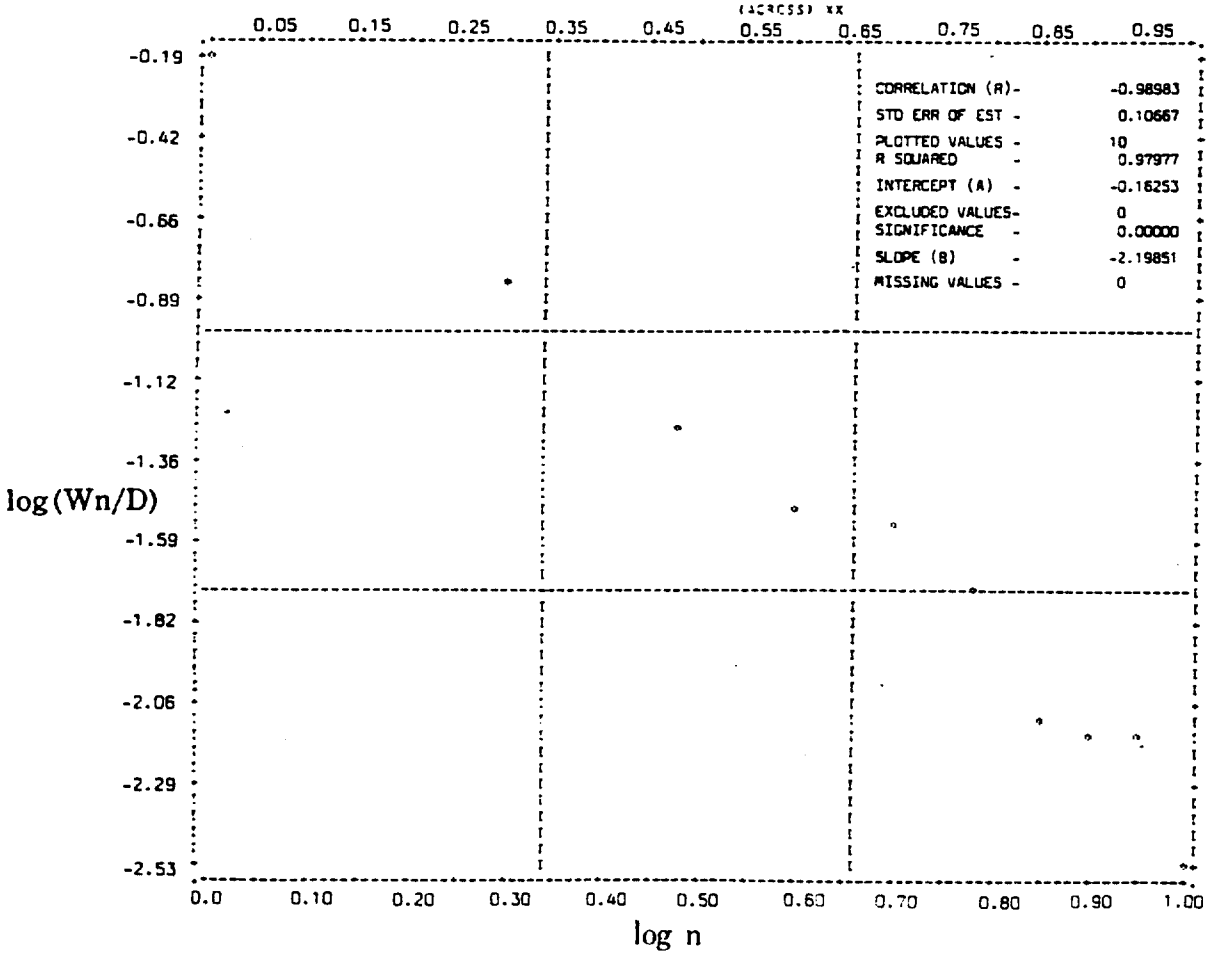


그림 21) 자료 I의 log(Wn/D)대 log n의 산포도

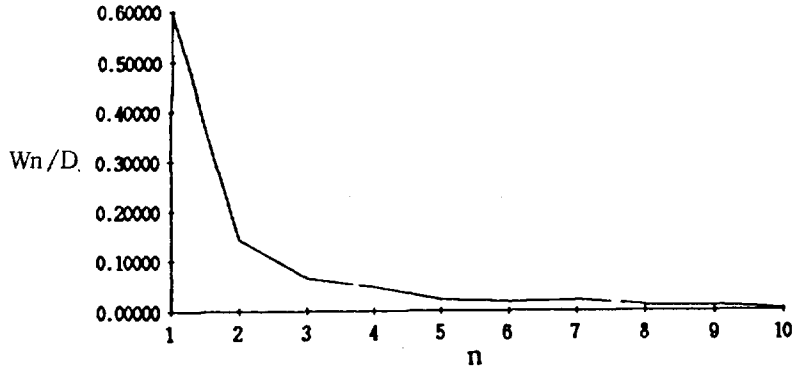


그림 22) 자료Ⅱ의 Wn/D 대 n의 그래프

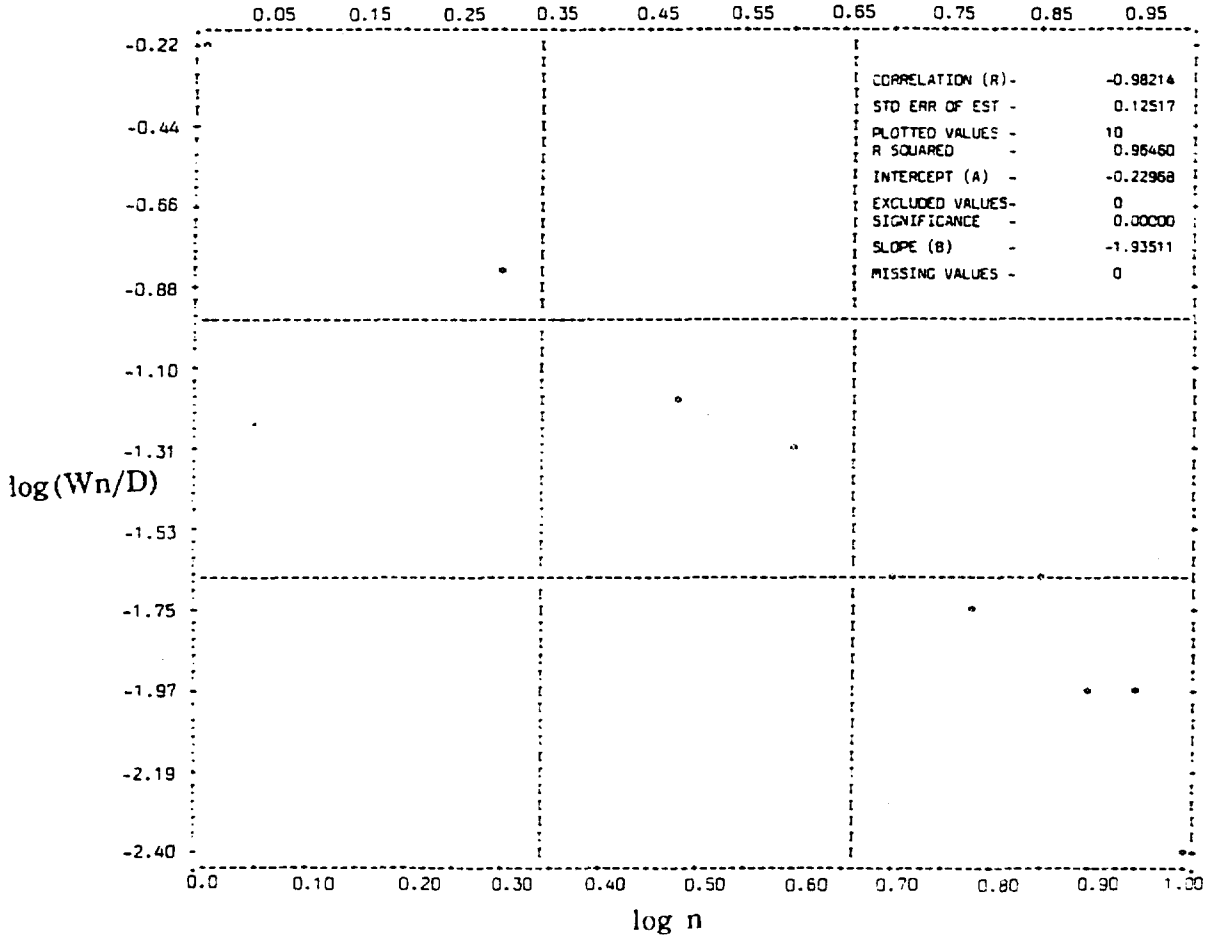


그림 23) 자료Ⅱ의 log(Wn/D)대 log n의 산포도

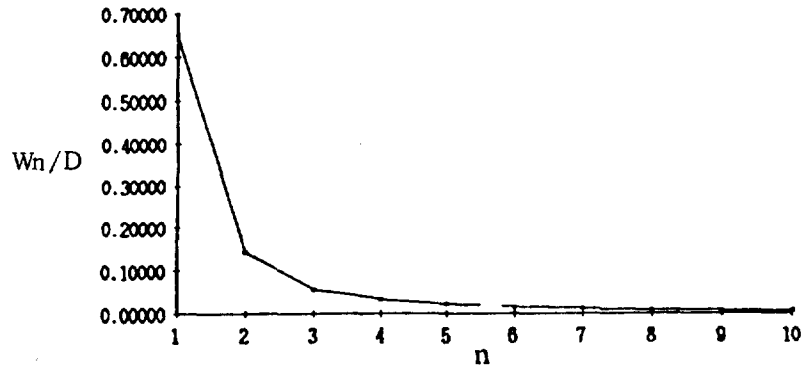


그림 24) 자료Ⅲ의 Wn/D대 n의 그래프

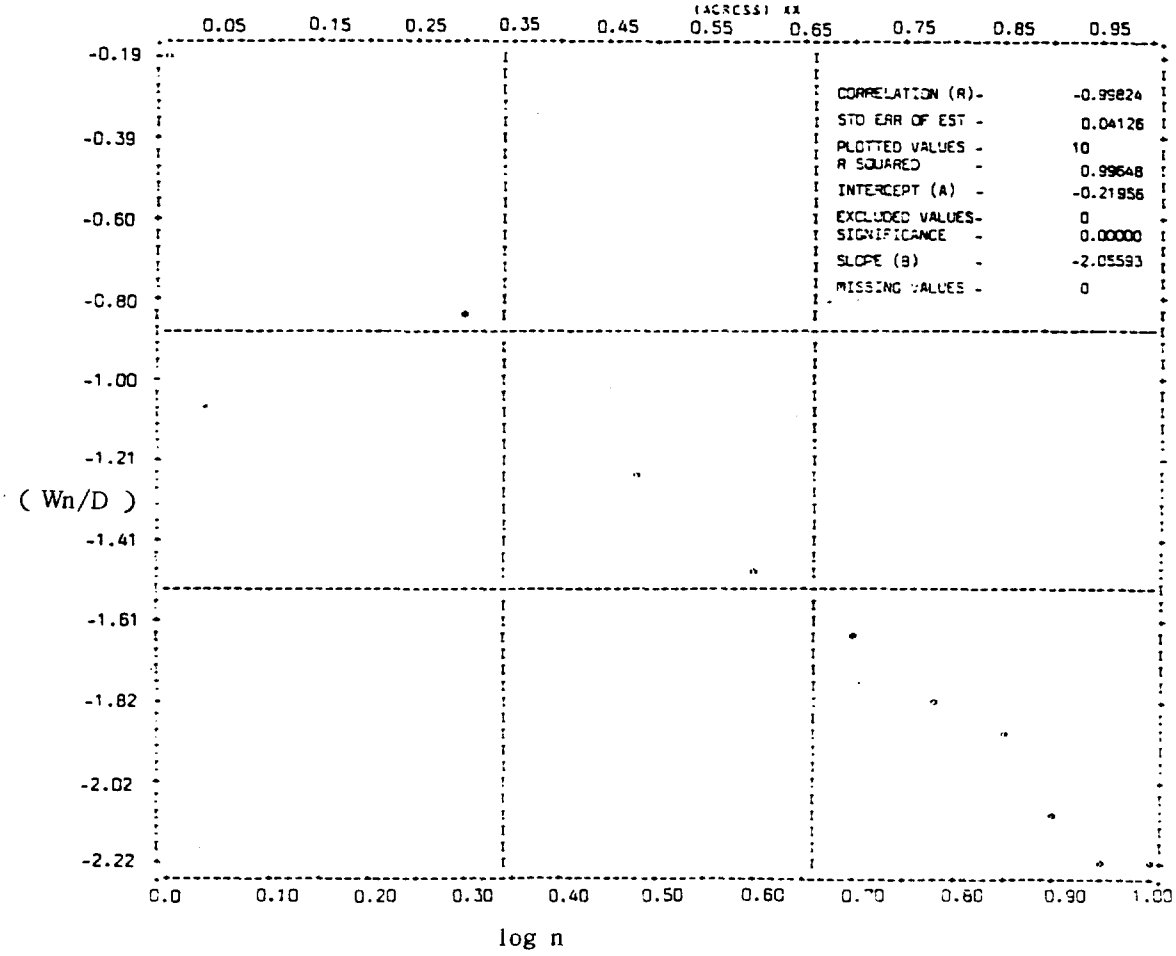


그림 25) 자료Ⅲ의 log (Wn/D)대 log n의 산포도

그림 21, 23, 25로 부터 구한 回歸方程式은 다음과 같다.

資料Ⅰ : $Y = -2.19851 \times -0.16253$
資料Ⅱ : $Y = -1.93511 \times -0.22968$
資料Ⅲ : $Y = -2.05593 \times -0.21956$

위의 回歸方程式에서 X대신에 logn을 Y대신에 log(Wn/D)를 대입하면 다음과 같다.

資料Ⅰ : $W_1/D = 0.6878/n^{2.19851}$ (L-1)
資料Ⅱ : $W_1/D = 0.58927/n^{1.93511}$ (L-2)
資料Ⅲ : $W_1/D = 0.60316/n^{2.05593}$ (L-3)

④ 各 要因에 의해 誘導된 公式의 比較分析

이상에서 收錄頻度가 낮은 單語들에 適合한 公式을 誘導하기 위해 各 要因과 收錄頻度가 낮은 單語들과의 關係를 公式으로 誘導해 보았다. 그러므로 이들 公式中에서 收錄頻度가 낮은 單語들의 分布를 가장 正確하게 나타낼 수 있는 公式이 어느 것인지 알아보기 위해 實驗資料의 公式에 의해 구해지는 수치를 比較해 보았다.

다음의 表13, 14, 15는 資料Ⅰ, Ⅱ, Ⅲ에 해당되는 各 公式들에 의해 구해진 單語數와 實際로 收錄된 單語數를 나타낸 表이다.

表13은 資料Ⅰ에서 구한 收錄頻度가 낮은 單語들의 各 頻度에 따른 單語數와 앞에서 誘導한 單語總數와의 分析에 의한 公式(J-1), 收錄頻度1인 單語數와의 分析에 의한 公式(K-1), 그리고 異種單語總數와의 分析에 의한 公式(L-1)에 頻度를 대입해 구한 單語數들을 기입한 表이다.

표 13) 자료Ⅰ의 수치비교표

빈 도	실제단어수	공식 J-1)	공식 K-1)	공식 L-1)
1	867	900.6	899.	922.3
2	192	201.3	201.2	200.9
3	77	83.8	83.8	82.4
4	42	45.	45.	43.8
5	39	27.8	27.8	26.8
6	25	18.7	18.8	17.9
7	10	13.4	13.5	12.8
8	13	10.1	10.1	9.5
9	9	7.8	7.8	7.3
10	4	6.2	6.2	5.8

표 14) 자료Ⅱ의 수치비교표

빈 도	실제단어수	공식 J-2)	공식 K-2)	공식 L-2)
1	603	596.2	595.7	595.8
2	144	155.8	155.8	155 8
3	65	71.0	71.1	71.1
4	48	40.7	40.8	40.7
5	22	26.4	26.5	26.5
6	18	18.6	18.6	18.6
7	22	13.8	13.8	13.8
8	11	10.6	10.7	10.7
9	11	8.5	8.5	8.5
10	4	6.9	6.9	6.9

표 14 는 資料Ⅱ에서 구한 頻度에 따른 單語數와 各 要因과의 分析에서 구한 公式(J-2), 公式(K-2), 그리고 公式(L-2)에 頻度를 구입해 구한 單語數를 기입한 표이다.

아래의 表15도 위와 같은 方法으로 資料Ⅲ을 대상으로 구한 수치와 公式(J-3), 公式(K-3), 그리고 公式(I-3)에 의해 구한 수치를 나타낸 표이다.

표 15) 자료Ⅲ의 수치비교표

빈 도	실제단어수	공식 J-3)	공식 K-3)	공식 L-3)
1	6,087	6058.3	5659.3	5659.5
2	1,338	1373.9	1361.1	1361.1
3	522	576.8	591.4	591.4
4	316	311.6	327.3	327.3
5	210	193.3	206.9	206.9
6	147	130.8	141.2	142.2
7	118	94.0	103.6	103.6
8	72	70.7	78.7	78.7
9	56	54.9	61.8	61.8
10	56	43.8	49.8	59.8

따라서 이들 表 13, 14, 15를 比較해 본 結果 收錄頻度가 낮은 單語들의 分布는 單語總數와의 關係에서 誘導되어야 할 것이다.

그러므로 收錄頻度가 낮은單語들의 경우 公式은 다음과 같다.

$$n \cdot Wn = k \cdot T \qquad (M)$$

(n : 頻度, Wn : 頻度 n 인 單語數, k : 특정 책자에 해당되는 상수, T : 單語總數)

公式(M)에서 k의 값은 收錄頻度가 1인 單語의 數를 單語總數로 나눈 수치에 매우 근사한 값이다.

다음의 表 16은 李祥喆이 쓴 碩士學位論文인 "한글自然語 Keyword 檢索시스템의 設計基準 및 한글의 Redundancy에 관한 硏究" (서울 : 숭전대학교 대학원 산업공학과, 1982)에 수록되어 있는 收錄頻度가 낮은 單語들의 表이다.

表 16) 수록빈도가 낮은 단어들의 등급-빈도표 (자료 4 ~ 10)

구 분	Ⅳ	Ⅴ	Ⅵ	Ⅶ	Ⅷ	Ⅸ	Ⅹ
收錄單語總數 T	4,366	5,747	5,895	10,113	11,642	10,261	16,008
異種單語總數 D	926	984	1,145	1,544	1,707	1,656	2,130
頻度1에 해당되는 單語數 W_1	560	583	722	903	1,010	976	1,215
頻度2에 해당되는 單語數 W_2	138	141	177	229	256	266	341
頻度3에 해당되는 單語數 W_3	48	58	66	100	107	101	133
頻度4에 해당되는 單語數 W_4	37	36	34	55	77	70	92
頻度5에 해당되는 單語數 W_5	26	21	24	37	31	43	57
頻度6에 해당되는 單語數 W_6	18	21	15	28	29	20	45
頻度7에 해당되는 單語數 W_7	10	7	9	20	20	16	19
頻度8에 해당되는 單語數 W_8	6	11	9	15	18	22	18
頻度9에 해당되는 單語數 W_9	7	8	9	13	15	13	20
頻度10에 해당되는 單語數 W_{10}	7	8	7	10	10	9	15

이 表의 資料 Ⅳ, Ⅴ, Ⅵ, Ⅶ, Ⅷ, Ⅸ, Ⅹ은 國內 科學技術硏究課題總覽의 農學分野資料를 대상으로 하였으며 各 資料의 範圍는 아래와 같다.

資料Ⅳ : 1977 ~ 1978 年
資料Ⅴ : 1978 ~ 1979 年
資料Ⅵ : 1979 ~ 1980 年
資料Ⅶ : 1977 ~ 1979 年
資料Ⅷ : 1978 ~ 1980 年
資料Ⅸ : 1977 ~ 1978 年과 1979 ~ 1980 年
資料Ⅹ : 1977 ~ 1980 年

表 16에서는 資料에서 구한 單語總數, 異種單語總數 및 各 頻度에 해당되는 各 資料의 單語數를 나타내고 있다.

表 16의 데이터를 전과 동일한 방법으로 처리

한 결과 다음의 公式을 구했다.

資料Ⅳ : $Wn/T = 0.12269/n^{1.98443}$ (N-1)

資料Ⅴ : $Wn/T = 0.0921/n^{1.93174}$ (N-2)

資料Ⅵ : $Wn/T = 0.11455/n^{2.07072}$ (N-3)

資料Ⅶ : $Wn/T = 0.08619/n^{1.94424}$ (N-4)

資料Ⅷ : $Wn/T = 0.0848/n^{1.97295}$ (N-5)

資料Ⅸ : $Wn/T = 0.09624/n^{1.99596}$ (N-6)

資料Ⅹ : $Wn/T = 0.07825/n^{1.96122}$ (N-7)

以上의 公式으로부터 公式(M)의 指數 α 가 2에 가까운 수치임을 알 수 있다.

그리고 各 資料의 單語總數에 대한 各 頻度에 해당되는 單語數의 比率間에도 어떤 一定한 法則性에 존재하는 것 같았으며 이 比率이 各 資料의 單語總數에 대한 異種單語總數의 比率에 의해 영향을 받는 것 같았다.

따라서 한 資料의 單語總數를 T, 異種單語의 總數를 D라고 하고 다른 어떤 資料의 單語總數를 T′, 異種單語總數를 D′라고 할 때 다음과 같은 관계가 성립된다.

$$Wn/T = \frac{D/T}{D'/T'} \times Wn'/T'$$

여기서 Wn은 한 資料의 頻度 n인 單語數를 意味하며 Wn′은 다른 어떤 資料의 頻度 n인 單語數를 各各 意味한다.

따라서 위의 공식을 정리하면 다음과 같다.

$$Wn = \frac{D}{D'} \times Wn' \qquad (O)$$

다음의 表17은 資料Ⅰ을 基準으로 하여 그 외 各 頻度에 해당되는 單語數를 公式(O)을 利用해 구한 수치를 나타낸 것이다.

表17의 Ⅱ난에서 Ⅹ난까지의 수치는 소수점 둘째자리에서 반올림한 수치이다.

表17) 各 資料의 異種單語總數를 같은 수치로 환산했을 때 頻度에 따른 單語數

자료 / Wn	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	Ⅶ	Ⅷ	Ⅸ	Ⅹ
W_1	867	799.8	869.9	810.9	794.5	845.5	784.2	793.4	790.3	764.9
W_2	192	191.0	191.2	199.8	192.1	207.2	198.8	201.1	215.4	214.6
W_3	77	86.2	74.6	69.5	79.0	77.2	86.8	84.0	81.7	83.7
W_4	42	63.6	45.1	53.4	49.0	39.8	47.7	60.4	56.6	57.9
W_5	39	29.1	30.0	37.6	28.6	28.1	32.1	24.3	34.8	35.8
W_6	24	23.8	21.0	26.0	28.6	17.5	24.3	22.7	16.1	28.3
W_7	10	29.1	16.8	14.4	9.5	10.5	17.3	15.7	12.9	11.9
W_8	13	14.5	10.2	18.6	14.9	10.5	13.0	14.1	17.8	11.3
W_9	9	14.5	8.0	10.1	10.9	10.5	11.2	11.7	10.5	12.5
W_{10}	4	5.3	8.0	10.1	10.9	8.1	9.5	7.8	7.2	9.4

第4章 結 論

한글문헌 3권을 대상으로 單語의 收錄頻度 分散을 調査한 結果 다음의 結論을 얻었다.

1) 한글문헌에 있어서도 單語의 收錄頻度와 等級사이에 一定한 統計的인 法則性이 存在하며 G.K. Zipf가 誘導한 公式과 일치한다.

2) Zipf第二法則은 한글문헌에 적용되지 않았으며 收錄頻度가 낮은 單語들의 分散은 單語總數의 側面에서 說明되어야 하며 本 硏究를 통해서 誘導한 公式은 다음과 같다.

$$n \cdot Wn = k \cdot T$$

3) 한글문헌에 있어서 收錄頻度가 낮은 單語들의 各 頻度에 해당되는 單語數는 各 文獻의 異種單語總數에 영향을 받는다. 따라서 어떤 文獻과 다른 文獻사이의 收錄頻度가 낮은 單語들 사이에는 다음 關係가 성립된다.

$$Wn = \frac{D}{D'} \times Wn'$$

以上의 硏究結果들은 한글문헌 3권을 대상으로 시도한 分析의 結果이므로 한글전반에 적용되는 일관성있는 法則을 誘導하기 위해서는 앞으로도 많은 實驗·分析이 있어야 하겠다.

參 考 文 獻

Bookstein, Abraham, "The bibliometric distributions", Listribution

Bookstein, Abraham, "The bibliometric distributions," Library Quarterly, Vol. 46, No. 4, 1976. pp. 416-423.

Booth, Andrew D., "A law of occurrences for words of low frequency," In: Introduction to information science. ed. by Tefko Saracevic. New York: R. R. Bowker Co., 1970. pp. 219-222.

Brookes, B. C. "The complete Bradford-Zipf bibliograph," Journal of Documentation Vol. 25, No. 1, 1969. pp. 58-60.

Brookes, B. C. "The derivation and application of the Bradford-Zipf distribution," Journal of Documentation, Vol. 24, No. 4, 1968. pp. 248-265.

Brookes, B. C. and Griffith B. C., "Frepuency-rank distribution," Journal of the American Society for Information Science, Vol. 29, No. 1, 1978. pp. 5-13.

Brookess, B. C., "Towards informetrics: Haitun, Laplace, Zipf, Bradford and the alvey programme," Journal of Documentation, Vol. 40, No. 2, 1984. pp. 120-143.

Bulick, Stephen, "Book use as a Bradford-Zipf phemomenon," College & Research Library, Vol. 39, No. 4, 1978. pp. 215-219.

Buckland, M. K. and Hindle, A., "Library Zipf," Journal of Documentation, Vol. 25, No. 1 1969. pp. 52-57.

Dacey, M. F., "A Growth process for Zipf's and Yule's city size laws," Enviroment & Planning, Vol. 11, No. 4, 1979. pp. 361-372.

Bairthone, R. A., "Empirical hyperbolic distributions (Bradford-Zipf-Ma-ndelbrot) for bibliometric description and prediction," Journal of Documentation, Vol. 25, No. 4, 1969. pp. 319-343.

Fedorowicz, Jane, "The theoretical foundation of Zipf's law and its application to the bibliographic database environment," Journal of the American society for Information Science, Vol. 33, No. 5, 1982. pp. 285-293.

Griffith, B. C. and Krevitt, B. A., "A comparison of serveral Zipf type distributions in their goodness to fit language data," Journal of the Americal Society for Information Science, Vol. 23, No. 3, 1972. pp. 220-221.

Haitun, S. D., "The role Zipf distribution," Scientometrics, Vol. 4, No. 3, 1982. pp. 181-194.

Herdan, G., "A critical examination of simon's model of certain distribution functions in linguistics," Appli d Statistics, Vol. 10, Nol. 10, No. 2, 1961. pp. 65-76.

Hill, B. M., "The rank-frequency form of Zipf's

law,"Journal of the American Statistical Association, Vol. 69, 1974. pp. 1974. pp. 1017-1026.

Hill, B. M. and Woodroofr, Michael, "Stronger forms of Zipf's law," Journal of the American S Statistical Association, Vol. 70, 1975. pp. 212-219.

Hill, B. M., "Zipf's law and prior distributions for the composition of a population," Journal of the American Statistical Association," Vol. 65, 1970. pp. 1220-1232.

Lawani, S. M., "Bibliometrics: its theoretical foundations, methods and applications," Libri., Vol. 31, No.4, 1981. pp. 294-315.

Narin, Francis and Moll, Joy K., "Bibliometrics," Annual Review of Information Science and Technology, Vol. 12, 1977. pp. 35-58.

O'Conner, Daniel O. and Voos, Henry, "Empirical laws, theory, Construction and bibliometrics, Library Trends, Vol. 30, No. 1, 1981. pp. 9-20.

Orlov, Ju. K., "Why, how and when does Zipf-Mandelbart law fail" J. Ling. Calc., No. 4, 1977. pp. 5-27.

Price, Derek de Solla, "A general theory of bibliometric and other cumulative advantate processes,"Journal of the American Society for Information Science, Vol. 27, No. 5-6, 1976. pp. 292-306.

Rapoport, A., "Zipf's law revisited,"In: Studies on Zipf's law, ed. by H. Guiter and V. Arapov. Bochum: Studienverlag Brokmeyer, 1982. (Quantitative liguistics; V. 16) pp. 1-28.

Rouault, A., "Zipf's law and Markovian sources," Annals De I'Institut Henri Poincare, Vol. 14, No. 2, 1987. pp. 169-188.

Scarrott, G., "Will Zipf join Gauss," New Scientist, Vol. 16, 1974. pp. 402-404.

Simon, H. R., "Why analyze bibliographies?," Library Trends, Vol. 22, No. 1, 1973. pp. 3-8.

Woodroofe, M. & Hill, B. M., "On Zipf's law," Journal of Applied probability, Vol. 12, No. 3, 1975, pp. 425-434.

Wyllys, Ronald E., "Empirical and theortical bases of Zipf's law," Library Trends, Vol. 30, No. 1, 1981. pp. 53-64.

Yavuz, D., "Zipf's law and entropy," IEEE Trans. Inform. Theory, Vol. It-20, No. 5, 1974. p. 650.

Yule, G. U., The Statistical study of literary vocabulary, ( ) : Archon Books, 1968.

Zipf, G. K, Human behavior and the principle of least effort. Reading: Massachusetts, 1949.

Zipf, G. K., The psycho-biology of language, Boston: Houghton Mifflin Co., 1935.

附 錄 1

Derek de Solla Price의 誘導過程 [31]

어떤 母集團을 P라 하고 이 集團의 어느 部分 f(n)은 n狀態에 있으며 n을 f(n)에 있는 個個人에 의해 各各 成就되어지는 成功의 總數라고 假定하면

$$\sum_{1}^{\infty} f(n) = 1$$

이다. 그리고 以前 成功의 平均數를

$$\sum_{1}^{\infty} n \cdot f(n) = R$$

이라고 하자. 變化는 항상 增加하는 方向으로만 가능하며 逆方向으로는 불가능하다.

새로이 소수의 개개인 dp가 추가되고 그리고 이에 따라 새로운 成功R·dp가 모든 구성요소에 추가된다고 하면 以前의 成功當 새로운 成功의 比率은 $R \cdot dp/RP = dp/p$이다. 그리고 以前의 成功 n을 지닌 개개인들 $P \cdot f(n)$의 계층을 위한 새로운 성공은 $P \cdot n \cdot f(n) \cdot dp/p = n \cdot f(n) \cdot dp$이며 따라서 n번째 상태에서 n+1번째 상태로 변천할 것이다. 그러므로 n번째 상태의 개개인의 수에 있어서 변천은

$$\frac{d}{dp} \cdot p \cdot f(n) = \begin{cases} -nf(n)+(n-1)f(n-1) & n>1 \\ -f(1)+1 & n=1 \end{cases} \quad (A1-1)$$

이다. 따라서

$$p \cdot \frac{d}{dp} \cdot f(n) = \begin{cases} -(n+1)+(n-1)f(n-1) & n>1 \\ -2f(1)+1 & n=1 \end{cases} \quad (A1-2)$$

방정식 (A1-2)에서 좌변은 거의 0에 가까운 값을 가지므로 0을 대입하면,

$$f(n) = \frac{n-1}{n+1} f(n-1) \qquad (A1-3)$$

이 된다. 따라서

$$f(n) = \frac{1}{n(n+1)} \qquad (A1-4)$$

이다.

만약 분산이 P와 함께 서서히 변하고 그리고 n과는 독립인 P의 함수의 생산물과 P와는 독립인 n의 함수의 생산물로 분리할 수 있다고 하면

$$f(n) = F(p) \cdot g(n) \qquad (A1-5)$$

이된다.

따라서 공식(A2-5)를 공식(A1-2)에 대입하면

$$\frac{P}{F(p)} \cdot \frac{dF(p)}{dp} = \frac{-(n+1)g(n)+(n-1)g(n-1)}{g(n)} = \begin{cases} -2 & n>1 \\ \dfrac{-2g(1)+1}{g(1)} & n=1 \end{cases} \quad (A-6)$$

이다. 그리고 변수들이 분리될 수 있으므로 공식(A2-6)은 모든 n을 위한 그리고 모든 P의 상수이다. 이 상수를 m이라 하면

31) Derek de Solla Price, "A General theory of bibliometrical other cumulative advantage distribution", Journal of the American Society for Information Seience, Vol.27,No.5-6, 1976, pp.294-295.

$$\frac{P}{F(p)} \cdot \frac{dF(p)}{dp} = m \qquad (A2-7)$$

이 된다. 그러므로

$$F(p) = C \cdot p^m \qquad (A2-8)$$

이다. 따라서

$$g(n) = \frac{n-1}{n+1+m} g(n-1)$$

$$= \frac{(n-1)./(m+1)}{(n+1+m)./} \qquad (A2-9)$$

공식 (A2-9)는 Euler's first Integral이라고 알려진 Beta 함수에 의해 적절히 표현될 수 있다.

$$\beta = (a,b) = \beta(b,a) = \int_0^1 x^{a-1}(1-x)^{b-1}$$

$$= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$= \frac{(a-1)./(b-1)./}{(a+b-1)./} \qquad (A2-10)$$

따라서 방정식 (A2-5), (A2-7), (A2-9)에서 다음 공식을 유도할 수 있다.

$$f(n) = CP^m \cdot \beta(n, m+2) \qquad (A2-11)$$

여기서 P의 어떤 특정가치를 위해 분산이 Beta 함수 (n, m + 2)의 가치에 비례함을 알 수 있다. 따라서 누가편익분포를 위한 density (밀도)는 다음과 같다.

$$f(n) = (m+1)\,\beta(n, m+2)$$

여기서 $m+1 = e$ 라고 하면

$$f(n) = e \cdot \beta(n, e+1) \qquad (A2-12)$$

이 된다.

附 錄 2

H. Simon의 유도과정[32]

Simon은 문헌에 있어서 단어들의 분산이 다음 공식을 따라 행동한다고 주장했다.

$$f(r) = A \cdot \beta(r, e+1)$$

여기서 $\sum_{n=1}^{\infty} f(r) = 1$이고 A와 e는 상수이며 $\beta(r, e+1)$은 r과 e + 1의 Beta 함수이다.

$$\beta(r \cdot e+1) = \int_0^1 \lambda^{r+1}(1-\lambda)^e d\lambda$$

$$= \frac{\Gamma(r)\Gamma(e+1)}{\Gamma(r+e+1)}$$

$$(0 < r,\ 0 < e < \infty)\ 0$$

여기서 $r \to \infty$일때

$$\frac{\Gamma(r)}{\Gamma(r+e+1)} \sim r^{-(e+1)}$$

이 된다. 따라서 f(r)은 다음과 같다.

$$f(r) = A \cdot \Gamma(e+1)\, r^{-(e+1)}$$

여기서 $A\Gamma(e+1)$대신에 C를 e+1 대신에 m을 대입하면

$$f(r) = C \cdot r^{-m}$$

이 된다. 따라서 일반화된 Zipf 第一法則과 같다.

32) G. Herdan, "A Critical examination of simon's model of Certain distribution functions in linguistics", Applied Statistics, Vol.10 No. 2(1961), p.67.

附 錄 3

Herdan-Waring공식의 유도과정 [33]

Herdan은 Waring의 공식 $\frac{1}{p-q}$ 을 이용했다.

$$\frac{1}{p-q}=\frac{1}{p}+\frac{q}{p(p+1)}+\frac{q(q+1)}{(p+1)(p+2)}+\cdots\cdots+\frac{q[r]}{p[r+1]}+\cdots\cdots \quad (p>q>0)$$

p-q를 양변에 곱하면

$$1=(p-q)\left[\frac{1}{p}+\frac{q}{p(p+1)}+\frac{(q+1)}{p(p+1)(p+2)}\cdots\right]$$

이 된다. 여기서 r번째 용어는 빈도분포에서 f(r)을 의미한다. 따라서 빈도의 평균과 분산은 다음과 같다.

$$\mu=\frac{q}{p-q-1}$$

$$\sigma^2=\frac{q(p-1)(p-q)}{(p-q-1)^2(p-q-2)}$$

附 錄 4

B.M. Hill의 유도과정 [34]

T개의 종(species)이 nT개의 속(genus)에 할당되어진다고 가정하자(단 각각의 속에는 최소한 한개 이상의 종이 할당되어진다). r번째 속에 있는 종의 수를 f(r)이고 n개의 종을 지닌 속의 수를 In이라고 가정하면 r=1, 2, 3…… nT일때

$$\sum_{r=1}^{nT} f(r)=T \qquad (D'-1)$$

이다.

그리고 속(species)에 대한 종(genus)의 할당방법이 보오스-아인쉬타인 분산형태라고 가정하면

$$\Pr\{f(r)|nT, T\}=\binom{T-1}{nT-1}^{-1} \qquad (D'-2)$$

이다. 여기서 f(r)=〔f(1)…… f(nT)〕이다.

그리고 nT가 T에 의존하는 분산의 임의변수이고 In이 nT와 T에 의존하는 임의변수라고 가정하자. 또한 $\theta=nT/T$의 분포가 $T\to\infty$일때 F(0)=0인 분산함수F(x)에 수렴한다고 가정하자. 그러면 $n\geq 1$이고 $T\to\infty$일때 n개 종을 지닌 속의 비율인 In/nT가 $\theta(1-\theta)^{n-1}$의 분산에 수렴하다. 그리고

$$E\{In/nT|T\}=\int_0^1 y(1-y)^{n-1}dy \qquad (D'-3)$$

이다.

만약 분산함수 F가 매개변수 α와 β가 $\alpha>0$, $\beta>0$인 베타함수(Beta function)라면 공식(D'-3)의 우측식은 다음과 같다.

$$F'(x)=\Gamma(\alpha+\beta)[\Gamma(\alpha)\Gamma(\beta)]^{-1}x^{\alpha-1}(1-x)^{\beta-1}$$

여기서 기호 "Γ"는 감마함수(gamma function)이며 $0\leq x\leq 1$이다. 따라서 $n\to\infty$일때

33) Jane Fedorowicz, "The theoretical foundation of zipf's law and its application to the bibliographic data base environment" Journal of the American Society for Information Science, Vol.33, N.5, 1982, p.289.

34) B.M.Hill and Michael Woodroofe "Stronger forms of zipf's law" Journal of the American Statistical Association, Vol.70 N.349. 1975, pp.212-219.

$$E\{\theta(1-\theta)^{n-1}\} \sim \alpha\Gamma(\alpha+\beta)[\Gamma(\beta)]^{-1} \left(\frac{1}{n}\right)^{(1+\alpha)}$$

이다.

여기서 기호 "~"는 양쪽의 비율이 일치함을 의미한다. 따라서, θ가 한단위간격(unit interval)으로 동일한 분포를 가진다면

$$E\{\theta(1-\theta)^{n-1}\} = [n(n+1)]^{-1}$$

이 된다. 이 공식이 매우 다양한 자료에 일치되는 Zipf 법칙의 간단한 형태이다.

그러나 위의 모델에 있어서 적절한 θ분포의 선택에 의해 generic-specific from과 유사한 공식을 유도했다. 그래서 위의 모델을 Zipf 법칙의 generic-specific form이 지니는 weak form이라고 한다.

generic-specifi form의 stronger form은 In/nT 그 자체가 $n^{-(1+a)}$에 거의 밀접하게 비례하는 경향이 있음을 의미하는 것으로 다음과 같이 유도할 수 있다.

전체 과(family)의 수를 K, 전체과에 할당되는 종(species)의 수를 T이라고 하는 그리고 i 번째 과(family)에 있는 Ti종(species)이 보오스-아인쉬타인 분산에 따라 속(genus)에 구분되어진다고 가정하고 정확히 n개 종을 지닌 i 번째 과내에 있는 속의 수를 Ini이라고 가정하자

그리고

$$Ini = I_{n1} + I_{n2} + I_{n3} \cdots\cdots I_{nk}$$

그리고

$$nT = nT_1 + nT_2 + nT_3 \cdots\cdots + nT_k$$

라고 하자. 따라서, 정확히 n개 종을 지닌 속의 비율은 가중평균(weight average)인 In/nT이다.

$$\frac{1}{nT} \cdot In = \sum_{i=1}^{k} \left(\frac{Mi}{M}\right)\left(\frac{Ini}{Mi}\right)$$

i의 값을 1로 두면 $T_1 \to \infty$일때

$$I_{n1}/nT_1 = \theta_1(1-\theta_1)^{n-1} + \delta T_1$$

이다. 그리고 $\theta_1 = \frac{nT_1}{T}$ 이며 δT_1은 확률상 0에 수렴한다. 따라서

$$Ini/nTi = \theta i(1-\theta i)^{n-1} + \delta Ti$$

이며 그리고 In/nT가 일정한 P(n)에 수렴함을 알 수 있다.

여기서 $p(n) > 0$이고 $\sum_{n=1}^{\infty} p(n) = 1$이다. 따라서

$$p(s) = \int_0^1 y(1-y)^{n-1} dH(y)$$

이다. 여기서 $n \geq 1$이며 H는 특정분포함수이다. 만약 $n \to \infty$이면

$$p(n) \sim C n^{-(1+a)} \qquad (c는 어떤 상수)$$

이다. 이 형태를 Stronger form이라고 한다.

附 錄 5

Zipf 第二法則[35]

만약 모든 異種單語를 서로 중복되지 않고 완벽하게 等級을 매길 수 있다면 그리고 等級 r에 해당하는 한 單語가 수록될 확률을 p(r)

35) A.D. Booth, "A law of occurrences for words of low frequency", In; Introduction to information science, edited by Tefko Saracevic. New York: R.R. Bowker Co., 1970, p.220.

이라고 한다면 收錄頻度의 總數가 T인 어떤 책자의 경우 單語의 收錄頻度를 다음과 같이 나타낼 수 있다.

T·P(1) : 等級 1 인 單語의 收錄頻度
T·P(2) : 等級 2 인 單語의 收錄頻度
T·P(3) : 等級 3 인 單語의 收錄頻度
⋮ ⋮
T·P(r) : 等級 r 인 單語의 收錄頻度

따라서 실제로 한번 收錄되는 單語의 總數(I_1)은 다음과 같다.

$$1.5 > T \cdot P(r) \geqq .5 \qquad (E\text{-}1)$$

Zipf 第一法則에 의하면 $p(r) = \frac{K}{r}$ (K : 특정책자에 해당되는 상수) 이므로 이를 공식 E-1) 에 대입하면

$$1.5 > T \cdot \frac{K}{r} \geqq .5 \qquad (E\text{-}2)$$

가 된다. 따라서 r의 최대값과 최소값은 다음과 같다.

$$r_{max} = \frac{KT}{.5}, \quad r_{min} = \frac{KT}{1.5}$$

따라서 r의 값 즉 I_1은 다음과 같다.

$$I_1 = \frac{4}{3}KT \qquad (E\text{-}3)$$

전술한 가정에 의하면 異種單語의 總數(D)는 等級이 가장 높은 單語의 等級과 같은 수치이므로 다음과 같이 구할 수 있다.

$$T \cdot P(D) \geqq .5 \qquad (E\text{-}4)$$

공식 E-4) 에 $P(D) = \frac{K}{D}$ 를 대입하면 다음과 같다.

$$T \cdot \frac{K}{D} \geqq .5$$

따라서

$$I_1/D = \frac{2}{3} \qquad (E\text{-}5)$$

이다.

이상과 같은 방법으로 頻度가 n인 單語의 總數 In의 공식을 구할 수 있다.

공식 (E-1) 과 같이 한 單語가 n번 收錄되려면 다음의 공식을 만족시켜야 한다.

$$n + \frac{1}{2} > T \cdot P(r) \geqq n - \frac{1}{2}$$

$P(r) = \frac{K}{r}$ 을 대입하면

$$n + \frac{1}{2} > T \cdot \frac{K}{r} \geqq n - \frac{1}{2}$$

이 된다. 따라서 In은 다음과 같다.

$$In = r_{max} - r_{min} = \frac{4KT}{4n^2-1} \qquad (E\text{-}6)$$

그러므로 공식 (E-3) 과 (E-6) 에 의해

$$In/I_1 = \frac{3}{4n^2-1} \qquad (E\text{-}7)$$

이 된다.

공식 (E-7) 이 Zipf 第二法則이다.

附 錄 6

일반화된 Zipf 第二法則[36]

일반화된 Zipf 第二法則은 일반화된 Zipf 第一法則을 이용하여 유도한 것이다. 즉 $\rho(r) = \frac{K}{r}$ 대신에 $\rho(r) = \frac{K}{r^B}$ 를 대입시켜 구한 공식이다.

어떤 단어가 한번 수록되기 위한 조건을

$$2 > T \cdot P(r) \geq 1 \qquad (F\text{-}1)$$

이라하고 그리고 어떤 단어가 n번발생되기 위위한 조건은

$$n+1 > T \cdot P(r) \geq n \qquad (F\text{-}2)$$

이라고 하면 공식 (F-1)과 공식 (F-2)에 의해

$$In = (kT)^{\frac{1}{B}} \left[\left(\frac{1}{n}\right)^{\frac{1}{B}} - \left(\frac{1}{(n+1)}\right)^{\frac{1}{B}} \right] \qquad (F\text{-}3)$$

이 된다. 따라서 이종단어의 총수 D는

$$T \cdot P(D) \geq 1 \qquad (F\text{-}4)$$

이다. 공식 (F-4)에 $P(D) = \frac{K}{D^B}$ 를 대입하면

$$D = (Tk)^{\frac{1}{B}} \qquad (F\text{-}5)$$

이 된다. 따라서 공식 (F-2)와 (F-5)에 의해

$$I_1/D = 1 - \left(\frac{1}{2}\right)^{\frac{1}{B}}$$

이 된다. 그리고 In/I_1 은 다음과 같다.(B = 1 일때)

$$In/I_1 = \frac{2}{n(n+1)}$$

36) A.D. Booth, op.cit., p.221.