***Dialectologia. Special issue, VII** (2017), 187-227.*
ISSN: 2013-2247

Received 21 August 2017.

Accepted 12 October 2017.

TWO STATISTICAL TREATMENTS OF SPANISH VOCABULARY: COMPOSITE INDICES OF FREQUENCY AND DISPERSION AND PRINCIPAL COMPONENT ANALYSIS APPLIED TO ORDINAL FREQUENCIES[1]

Hiroto UEDA

University of Tokyo*

uedahiroto@jcom.home.ne.jp

Abstract

In this study, first, the characteristics of the collected materials to carry out the comparative analysis will be explained. Next, the statistical method based on the combination of two important indexes, frequency and dispersion, will be described, for which it is proposed to convert gross frequencies into ordinals. In the next section, principal component analysis will be performed to find the important factors within the multitude of current parameters. Finally conclusions will be drawn on the two methods and the application to lexical studies will be considered in general.

Keywords

Spanish frequency dictionaries, frequency, dispersion, ordinals, principal component analysis

[1] I am grateful to Maria-Pilar Perea from University of Barcelona, who helped me with the preparation of this study. This work was supported by JSPS KAKENHI Grant Number 16K02657. It has also been developed under the FFI2013-41077-P project, funded by the Spanish Ministerio de Economía y Competitividad.

* Departament of Spanish Language, Faculty of Arts and Sciences, University of Tokyo, Komaba, 3-8-1, Meguroku, Tokyo, Japan 153-8902.

DOS TRATAMIENTOS ESTADÍSTICOS DEL VOCABULARIO ESPAÑOL: ÍNDICES COMPUESTOS DE FRECUENCIA Y DISPERSIÓN Y ANÁLISIS DE COMPONENTES PRINCIPALES APLICADO A FRECUENCIAS ORDINALES

Resumen

En este estudio, primero, se explicarán las características de los materiales recogidos para llevar a cabo el análisis comparativo. A continuación, se describirá el método estadístico basado en la combinación de dos índices importantes, la frecuencia y la dispersión, para lo cual se propone convertir frecuencias absolutas en ordinales. En la sección siguiente, se realizará el análisis del componente principal para encontrar los factores importantes dentro de la multitud de parámetros actuales. Finalmente, se extraerán conclusiones sobre los dos métodos y se considerará la aplicación a los estudios léxicos en general.

Palabras clave

diccionarios de frecuencia en español, frecuencia, dispersión, ordinales, análisis de componentes principales

1. Introduction

In 1987, an internal report on the Spanish vocabulary was presented at an institute of the Tokyo University of Foreign Studies, where special emphasis was placed on frequency and dispersion in different fields (Ueda 1987). At that time, two important works were used, one by García Hoz (1953) and another by Juilland & Chang-Rodríguez (1964). In Ueda (1987), the frequency of words was counted in the Spanish manuals published in Europe and the United States. Since then, different studies have been published on the frequency of the Spanish vocabulary, of which Justicia (1995), Ávila Muñoz (1999), Almela *et al.* (2005) and Davies (2006) are important. A work specialized in the lexicon of sciences by García Hoz (1976) has also been published, which could not be used in the mentioned report.

All these materials are useful in researching the necessary statistical aspects to build a set of basic Spanish vocabulary at the service of Spanish teachers, researchers and students. However, such materials are difficult to compare because of their heterogeneity. Some works are dedicated to analyzing a specific field, for example, the terms that appear in the subjects of the high school and pre-university courses, the words used in the compositions of the students of the primary school or the words collected in the speech of a locality; while other works constitute a corpus composed of different balanced fields. Naturally each work can be used separately to study the aspects that it deals specifically. However, it would be now interesting to carry out a comparative study of the vocabularies collected in these works as well as in future studies.

The objective of this study is not to start searching for equitably distributed materials with pre-established criteria, for example, oral text—written text, colloquial speech—formal speech, literary work—non literary work, natural sciences—social sciences—human sciences, etc., each of which would create a corpus of the desired quantity, say, 100,000 words. Previously, a specific work, based on Spanish handbooks, with this same amount, was already carried out to prepare the aforementioned report of 1987, which showed that the work was laborious and difficult to develop by a single researcher. If we need 20 or 30 fields to make a general observation of the Spanish vocabulary, its realization will be only viable through a team of researchers and with highly efficient computer tools.

Before embarking on an immense task of the statistical lexical study, it is convenient to have a panoramic view of the works published so far. It is not a matter of establishing the criteria for selection of materials in advance, which is extremely difficult if we start from scratch, but to arrive at some criteria based on the data we have at hand. It is proposed to do *a posteriori* the categorization which is normally done before preparing the materials object of the study from the data analysis. This method of "post-categorization" is possible to be applied to the present study with appropriately prepared materials and the application of multivariate statistical methods.

In this study, first, the characteristics of the collected materials to carry out the comparative analysis will be explained. Next, the statistical method based on the combination of two important indexes, frequency and dispersion, will be described, for which it is proposed to convert gross frequencies into ordinals. In the next section, Principal Component Analysis will be performed to find the important factors within the multitude of current parameters. Finally conclusions will be drawn on the two methods and the application to lexical studies will be considered in general.

2. Materials

Before starting the comparative analysis of the collected data, we will briefly explain the most outstanding characteristics of the works used in this study. They are García Hoz (1953), Juilland & Chang-Rodríguez (1964), García Hoz (1976), Ueda (1987), Justicia (1995), Ávila Muñoz (1999), Almela *et al.* (2005) and Davies (2006) in chronological order of publication.

2.1 Garcia Hoz (1953)

The purpose of Garcia Hoz's work (1953) "is not purely philological or linguistic, but vocabulary is conceived as a means of psychological knowledge and content of teaching" (1953: 15). In order to determine the usual vocabulary "required in general education, which reaches every man independently of specialization" (1953: 15), the author thought of four aspects of life, "clearly differentiable, because they are different in themselves and because they have their own means of expression" (1953: 18). The four differentiated aspects and corresponding materials are (1953: 20-37):

Aspect	*Material*
Family life	Private letters from different regions of Spain
Undifferentiated social life	Newspapers (doctrinal articles, news, entertainment, advertisements)
Regulated social life	Official documents (political, religious, trade union)
Cultural life	Best selling books from the year 1943

Theater plays were excluded as a written source of language for two reasons: (1) "in the theater a particular manifestation of life is not reflected" and "becomes evident in the language of one and the other work of theater", so "if I used this theater as one more source of usual vocabulary, I would mix in the research several criteria of source selection" (1953: 21); (2) in the language of theater "life is presented, not as it is, but as interpreted by the dramatic author (...) it is not real life manifested directly, but manifested in a reflective way, through the peculiar expression of each author" (1953: 21). This is a significant negative comment on artificial language, which will also be discussed in §2.4.

The author devoted five pages to explain how the number of words to be counted was delimited. After several experiments, he concluded that the figure of 100,000 words is adequate, since "from this number the terms are repeated over a hundred times on average (...) when the words have a frequency of 100, the one that has not appeared hardly can be said usual". This amount of 100,000 to represent a vocabulary field has become almost a rule in later studies. The counting of words was done manually.

2.2 Juilland & Chang-Rodriguez (1964)

Almost a decade after the publication of García Hoz (1953), a monumental work of Juilland & Chang-Rodríguez (1964) appeared, with computer tools in the United States, which served as fundamental reference in the later studies by its systematic method and distinctions of concrete forms, with disambiguation of homonyms, in addition to the lemmas that represent its members. This is the collection of five areas divided into two genres, creative (drama, fiction and essay) and non-creative (periodical and technical), published in Spain between the two world wars (1964: XVI). The works were randomly selected from the list of *Union Catalogue of the Library of Congress*. Specifically, the five areas (*lexical worlds*) are (1964: XVII-XXI):

1. Drama: Sentences were sampled from 44 plays by 23 authors selected at random from 225 volumes of 31 authors.

2. Fiction: Sentences were sampled from 50 novels or short stories by 18 authors randomly selected at random from 250 volumes by 51 authors.

3. Essay: Sentences were sampled from 52 works by 38 authors selected at random from 569 works by 139 authors.

4. Technical documents: Sentences were sampled from 32 works of 29 authors selected at random from 225 works by 195 authors.

5. Periodical publications: Sentences were sampled from various sections (editorials, news, chronicles, advertising, etc.) of 5 newspapers and 5 magazines.

The quantitative delimitation of words in each area was set at 100,000, the same as Garcia Hoz (1953). Unlike the Spanish author, who showed all the words (12,428), these were limited to the first 5,024 with total frequency equal to or greater than 5 occurrences.

The merit of Juilland & Chang-Rodríguez (1964) was to have taken into account not only the frequency but also the calculated dispersion of each word. They presented a mathematical formula that represents the degree of Dispersion (D) (1964: LIII):

$$D = 1 - SD / (2\ m),$$

where SD is the standard deviation and m is the arithmetic mean (average). It varies from 0 (zero dispersion), for example {10, 0, 0, 0, 0}, to 1 (optimum dispersion), for example {2, 2, 2, 2, 2}.

Another merit of the work is the proposal to combine frequency (F) and dispersion (D) in the form of Usage (U):

$$U = F \times D,$$

that is to say, it is the multiplication of Frequency (F) by Dispersion (D).[2] These two formulas will be discussed in more detail in section 3 of this study (§3.2, §3.3).

2.3 García Hoz (1976)

With the interval of almost two decades since García Hoz (1953), the same author investigated the lexicon of scientific orientation (1976), in this case using the

[2] In fact, Juilland & Chang-Rodríguez (1964: LXVIII) used the formula:

$$U = F \times D / 100$$

where D (dispersion) is represented by percentage.

computer. The author explains (1976: 18): “For the determination of the General Vocabulary of Scientific Orientation at the University entrance level, it seems that the most appropriate material would consist of the texts of all the scientific subjects included in the study plans of *bachillerato* (university preparation studies in three years). For this purpose “a textbook, the most representative of each of the subjects, was selected”. The subjects covered were: Physics, Chemistry, Biology, Zoology, Botany, Geology, Literature, Grammar, History, Geography, History of Philosophy and Philosophy. Classical and modern languages were excluded because they had their own expression distinct from the Spanish language.

In column F.REL. (relative frequency) of the listing, the normalized frequency is presented, based on 100,000 words.

2.4 Ueda (1987)

For our part, we tried to calculate the frequency of the words appeared in the Spanish handbooks, published in Europe and the United States, to carry out a quantitative study in comparison to the figures presented in García Hoz (1953) and Juilland & Chang-Rodríguez (1964).

Naturally it is not a question of the Spanish language in its reality, but of a model elaborated by teachers and researchers of the language. The same reasons for excluding theater plays in García Hoz’s study (1953: 21), mentioned in §2.1, would be applied to the materials treated in this work, since in the language of the Spanish handbook life was not presented as it was, but as interpreted by the author.[3] However, the purpose was not to observe the reality of the language, but the reality of the texts treated in the handbooks to place them within the universe of the lexicons handled in previous studies. The same base of 100,000 was used to adjust the frequency to the common quantity of the two mentioned works.

Data from Spanish handbooks are included in this current study to check their situation within the multivariate universe.

[3] Indeed, opinions have been received from Morales (1989), which have been answered in Ueda (1993).

2.5 Justicia (1995)

Justicia’s work (1995) is interesting when observing the development of Spanish children vocabulary. The author investigated the free compositions written by 3,402 students of the primary school of Eastern Andalusia (Almeria, Granada, Jaen and Malaga), where more than half a million (528,542) of occurrences were been counted and 8,937 different words were recorded in total. The students were divided into three groups (6-7 years, 8-10 years, 11-13 years) and the list presented the absolute frequency of each group and the sum of the three groups.

Although it is interesting to observe the linguistic development of Eastern Andalusia children's vocabulary throughout the three stages, in this current study it has been decided to use only the sum of frequencies of the stages based on 100,000 words.

2.6 Ávila Muñoz (1999)

For obtaining the true quantitative data of lexicons of the spoken language, we have to wait for Ávila Muñoz (1999), who studied the urban varieties observed and transcribed from an oral corpus of 54 hours of recording. He classified the spoken mode in five groups (1999: 23):

Group A: Bidirectional, Face to face, Free — Conversation
Group B: Bidirectional, Face to Face, Not Free — Interview
Group C: Bidirectional, Remote, Free — Family telephone conversation
Group D: Bidirectional, Remote, Non-free — Telephone interview
Group E: Unidirectional — Conference, classes, TV, Radio

It follows the model of Juilland & Chang-Rodriguez (1964) in the mathematical treatments of the Dispersion and the Usage (1964: 78-80), where he explained the maximum value of the standard deviation (SD) in:

$$SD \leq (n - 1)^{1/2} m$$

where n is the number of data and m is the mean. This formula will be discussed in §3.2.

In the list of words with their Frequency accompanied by Dispersion and Usage, the frequency is normalized by one million words, so that it must be divided by 10 to adjust it to the common scale of 100,000 words.

2.7 Almela et al. *(2005)*

The joint work of Almela *et al.* (2005) is based on the CUMBRE corpus of 20,662,306 words, whose general design is presented in the following table (2005: 8-9):

*	*Written Language*	*Oral Language*	*Total*
Spain	70%	30%	65%
Latin America	35%	60%	35%
Total	100%	100%	100%

Within the written language we find: 1. Books (novels, poetry, short stories, history, economics, art, architecture, technique, society, psychology, philosophy, computer science, cinema, law, travel, biography, medicine/health, encyclopedias, cooking, music); 2. Magazines (general information, woman, house and kitchen, heart, technical/specialized); 3. Daily press (national and regional); 4. Educational manuals (university, primary education, secondary education, vocational training, non-regulated teaching); 5. Information/publicity leaflets (public administration, advertisements); 6. Anthologies (literary texts); 7. Humor, entertainment (short humorous writings, TBOs, jokes); 8. Written correspondence (formal, non-formal); 9. Sectoral languages (elders, adults, youth, children, men, women, fashions, politics, Spanish manuals for foreigners).

As for oral language: 1. Radio and TV (50% national level and 50% regional/local). 1.1. Conversation, middle/high stratum, formal and non-formal registers: society, culture, science, education, human sciences, history, religion, economy, environment, politics, others. 1.2. Conversation, middle/low stratum, formal and non-formal registers. 1.3. Debate, in both social strata and registers. 1.4. Group discussion, involving both strata and registers; 2. Face-to-face conversations of everyday life: greetings, health, time, money, shopping, travel, home, etc.; 3. University, secondary and primary school classes; 4. Friends/family talk about health, time, family, habitual facts, purchases, travel, plans, etc.; 5. Telephone conversations on professional and usual matters and relationships; 6. Narration of facts, etc., on various subjects of normal and daily life; 7. Real situations of daily living: medical office, law office, teacher's office, work, travel agency, street/home surveys, bank, shop, restaurant, bar, garage, in a waiting room, a means of transport, in the taxi, with the police (asking, police station, at the hotel reception, at a party, at the market and others).

In its Annex II (2005: 299-387), corpus frequencies and relative frequencies based on a million words have been listed, together with the classification by frequency bands (very high, high, remarkable).

Therefore, it is a material prepared in a general corpus, constituted by different multivariate texts. To know the common base frequency of 100,000 words, the normalized frequency per million has been divided by 10.

2.8 Davies (2006)

Davies' Frequency Dictionary (2006) is part of the collection of existing corpus with the number indicated by million (2006: 3):

*	*Spain*	*Latin America*	*Total*
Spoken	1.35	2.00	3.35
Transcripts/Play	1.67	1.73	3.40
Literature	2.42	3.96	6.38
Texts	3.20	3.67	6.87
Total	8.64	11.36	20.00

The author also takes into consideration the degree of diffusion of the word in question used through the registers within the whole corpus (2006: 6). Specifically the degree of diffusion is represented in the number of blocks within 100 blocks in total where the same word appears. Each block maintains the same amount of 200,000 words. The entire corpus is 200,000 words multiplied by 100 blocks, equal to 20,000,000 words.

In the listing of Frequency index (2006: 12-182), 5,000 words have been listed with the number of blocks and the absolute frequency within 20 million words. In our study the frequency divided by 200 has been used to obtain the normalized frequency calculated with the same common base of 100,000 words.

3. Frequency and dispersion

3.1 Frequency

For the limitation of space, only the initial part of the frequencies compared in 15 fields treated in the studies mentioned in the previous section (§2) is exposed:[4]

VOCABLO	*Car*	*Pe.1*	*Ofi.*	*Lib.*	*Cie.*	*Dra.*	*Fic.*	*Ens.*	*Pe.2*	*Téc.*	*Pri.*	*Man.*	*Mál.*	*Alm.*	*Dav.*
a	3690	3899	3109	3334	3968	3355	3380	3001	3641	3212	2239	3816	2513	2745	2649
abajo	5	4	4	9	11	10	17	3	0	1	18	9	31	7	11
abandonar	9	21	5	13	6	6	12	16	18	5	6	5	2	14	13
abandono	0	12	1	3	0	0	2	2	2	1	0	0	0	3	2
abarcar	1	3	1	1	4	0	3	1	4	5	0	0	1	3	3
abastecimiento	2	12	2	0	0	0	0	0	0	0	0	0	0	1	0
abatir	0	1	0	0	0	1	5	2	4	1	0	0	0	0	0
abeja	2	1	0	3	3	3	2	4	2	1	20	0	0	0	1

Table 1. Absolute frequency / Initial part in alphabetical order

[4] The columns are: Car: Letter, Pe.1: Periodical-1, Ofi.: Official documents, Lib.: Books (García Hoz 1953), Cie.: Science (García Hoz 1976), Dra.: Drama, Fic.: (Ueda 1987), Man.: Manual of Spanish (Ueda 1987), Pri.: Primary School (Justice 1995), Man.: Manual of Spanish (Ueda 1987), Mál.: Málaga (Ávila Muñoz 1999), Alm.: Almena *et al.* (2005), Dav.: Davies (2006).

From this table, different values are calculated, which represent the frequency scale:

VOCABLO	*Sum*	*Number*	*Mean*	*M.L.*	*M.O.*
a	48 551	15	32 36.7	3.510	4
abajo	140	15	9.3	.969	1006
abandonar	151	15	10.1	1.004	918
abandono	28	15	1.9	.270	3203
abarcar	30	15	2.0	.294	3108
abastecimiento	17	15	1.1	.056	3989
abatir	14	15	.9	.000	4256
abeja	42	15	2.8	.450	2553

Table 2. Sum, Number of data, Mean, Mean in logarithm (M.L.), Mean in ascending ordinal (M.O.) / Initial part

The Sum is the total of the absolute frequencies that are listed in the 15 field columns. The second column is the Number of fields, which ordinarily corresponds to the number of columns (15). However, when some data is missing in the cell, the number of data (14, 13, ...) decreases. This occurs, for example, in García Hoz (1953), which makes no distinction of *que* (conjunction) from *qué* (interrogative pronoun). Since there is no way to specify the frequency of each word, the cell is left blank and the number of data becomes 11, since García Hoz (1953) generally offers data from four fields (*Carta, Periodico-1, Oficial, Libro*).[5] The Mean is the average calculated by dividing the Sum (total frequency) by the Number of data.

Mean = Sum / Number of data

The Mean in logarithm (M.L.) is calculated by the logarithm function with base 10 applied to the Mean:

M.L. = Log(Mean) = $\log_{(10)}$ Mean

[5] For this reason, instead of the Sum, Mean is used to represent the total frequency of each word.

For example, the mean in logarithm of the preposition *a* is Log(48551) = 3.510. The reason for using the logarithm value is that the frequency distribution of the words is highly skewed, that is, there is reduced number of extraordinarily frequent words, while most words are infrequent. The following two figures show the difference between the distribution of absolute mean and that of mean in logarithm:

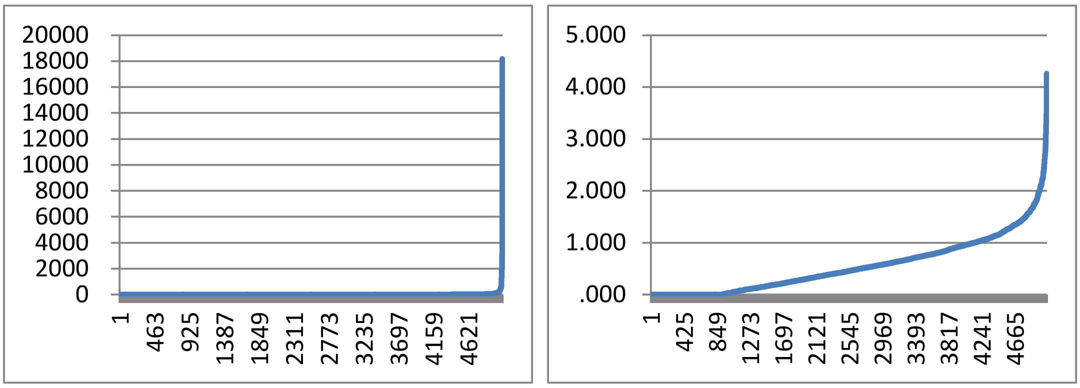

Figure 1. Distribution of absolute mean / Fig. 2. Distribution of mean in logarithm

Finally it is also useful to calculate the Mean in ordinal (M.O.).[6] The descending ordinal is acquired by setting the frequency order from the maximum to the minimum. For example, an ordered set of numbers {1, 2, 4, 5, 10} becomes the descending ordinal of {5, 4, 3, 2, 1}. For operations using the frequency ordinal, it is convenient to use the inverse way, the ascending ordinal {1, 2, 3, 4, 5}, which reflects the ascending magnitude of the absolute frequency. When dealing with data containing the same frequencies, for example {1, 2, 4, 4, 10}, the normal ascending ordinal gives {1, 2, 3, 3, 5}, instead of the ascending naive ordinal {1, 2, 3, 4, 5}. On the other hand, the ordinal that the statisticians propose is {1, 2, 3.5, 3.5, 5}, where the average of naive ordinal values are presented (Ikeda 1976: 133-144; Scholfield 1995: 168-187). In this way the equal data take the reasonable value, since the distribution of the normal ordinal {3, 3} is skewed towards the lower value (3). The graph of distribution of Mean in statistical ordinal is presented in a perfectly linear way:

[6] For the English term *rank*, which corresponds to Spanish 'rango', "ordinal" is used, because "rango" also corresponds to the English *range*, which is the difference between the maximum and the minimum.

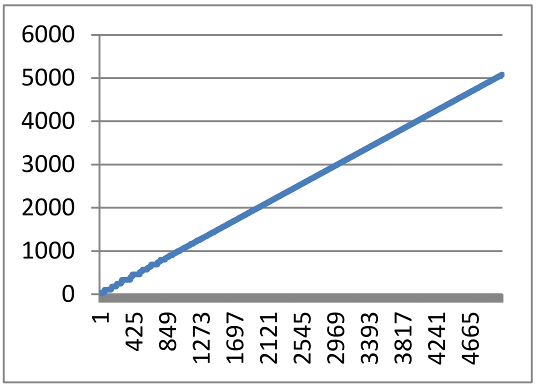

Figure 3. Distribution of mean in ascending statistical ordinal

3.2 Dispersion

Most previous lexicographical research has been devoted to observing not only the total frequency of each word but also the dispersion or degree of diffusion of the individual frequencies, considering that the amplitude of occurrence is as important as the scale of magnitude. García Hoz (1953: 385), together with the list of "usual vocabulary", the result of the frequency calculus, has prepared another one of the "common vocabulary", which is the "set of words contained in all types of vocabulary": Letters, Newspapers, Official documents and Books. Davies (2006: 6-7, 12-182), as we have previously seen (§2.2), divides the entire corpus composed of 20 million words into 100 blocks of 200 thousand words and presents the number of blocks where the word appears in next to the total frequency.

In order to know the exact method of calculating Juilland & Chang-Rodríguez's Dispersion (Disp.) (1964: LIII), we repeat the formulation discussed briefly in §2.2:

Disp. = 1 - SD / (2 m),

where SD is the Standard deviation and m is the Mean. The Dispersion varies from 0 (negative extreme dispersion), for example {10, 0, 0, 0}, to 1 (optimal dispersion), for example {2, 2, 2, 2, 2}.

In the absence of an explanation of how the aforementioned formula has been derived, Avila Muñoz (1999: 78-80), who sets out the following formula, is helpful:

$D = 1 - SD / [(n - 1)^{1/2} m]$, n: number of data; m: mean

by which it is known that the number 2 in the formula of Juilland & Chang-Rodríguez (1964) corresponds to the square root of (*n - 1*), where *n* is 5 in their work. The derivation of the square root of (*n - 1*) is explained in ADDENDA-1.

The following graph shows the distribution of the dispersion of 5,079 words in the 15 fields in ascending ordinal:

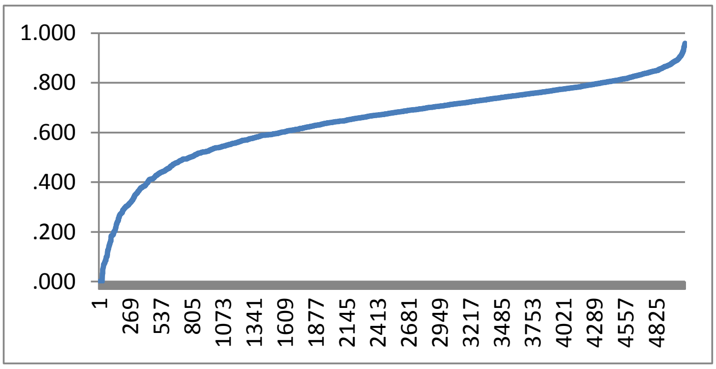

Figure 4. Distribution of the dispersion

According to the graph, the distribution of dispersion is not as biased as that of frequency shown in the previous section (§3.1, Figure 1).

3.3 Usage and Importance

Juilland & Chang-Rodriguez (1964: LXIV-LXXIV) formulated the composite index of Frequency (F) and Dispersion (D) in the form of Usage (U):

$U = F \times D$,

which is followed by Ueda (1987) and Ávila Muñoz (1999). This formula was devised "in accordance with the concept which assumes coefficients of word usage to predict or approximate word occurrence in an «ideal»", that is, perfectly representative and perfectly unbiased sample, so that "the formula takes dispersion as a corrective to frequency rather than vice versa, although it can also be said that the two factors are being coordinated" (Juilland & Chang-Rodríguez, 1964: LXVII).

By this explanation, it is understood that the Dispersion (D) functions as a corrective of the Frequency (F) and, for being corrective, the result of the multiplication in the Usage form (U) does not differ significantly from the Frequency (F), which is shown in the graph 19 of the work of the two authors (1964: LXXV).

The same is checked in the current data of 15 fields in the following table, where the Mean is used instead of the Frequency, for reasons explained in §3.1. They are now presented in descending order to see the first highest figures:

VOCABLO	*Mean*	*Disp*	*Usage*
el - la - lo	18157.3	.899	16321.1
de	7482.2	.897	6709.3
y	3244.2	.946	3068.7
a	3236.7	.958	3100.5
lo - la - le	2902.3	.888	2578.5
en	2423.8	.944	2287.7
que	2042.4	.805	1643.2
ser	1999.8	.927	1853.6
un - una - uno	1858.8	.946	1758.1
se	1544.1	.927	1431.4
qué	1390.3	.810	1126.4
no	1303.0	.816	1063.5
(…)			

Table 3. Mean, Dispersion and Usage / Initial part in descending order

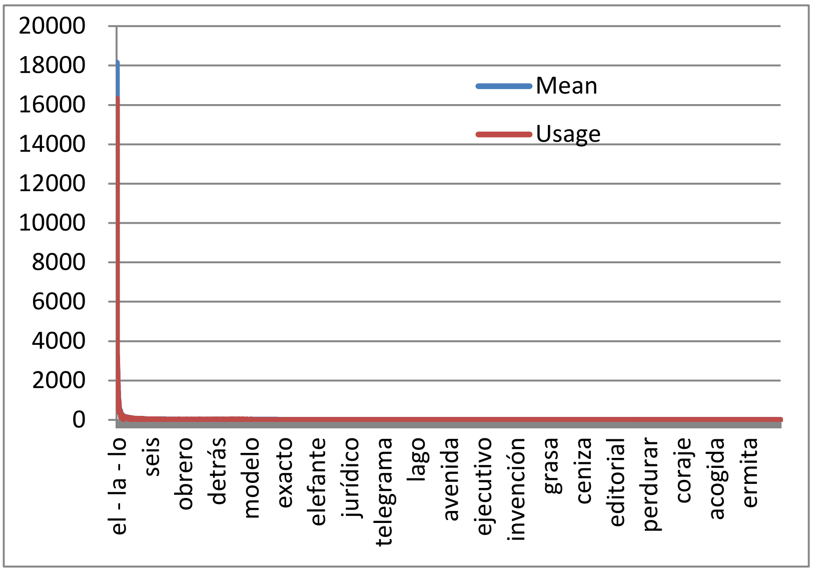

Figure 5. Mean and Usage in Comparison

Thus, despite the fact that Juilland & Chang-Rodríguez (1964) assert that "the two factors [Frequency and Dispersion] are being coordinated", it is not certain that the two factors are truly coordinated in an equitable way. In order to coordinate the Mean and Dispersion equitably, instead of the frequency itself, we propose to use the ratio with respect to the maximum of the Mean in Logarithm (M.L.max), in the form of "Mean Ratio in Logarithm" (M.R.L..):

$$\text{M.R.L..} = \text{M.L.} / \text{M.L.max.}$$

Finally, the normalized frequency index used in this paper is presented with the "Importance" (I):

$$\text{I} = (\text{M.R.L..} \times \text{D})^{1/2}$$

Unlike the Usage (U) of Juilland & Chang-Rodríguez (1964) discussed in §3.3, the Importance index (I) is calculated by multiplying two values of the same scale from 0 to 1. On the other hand, as frequency (F) and dispersion (D) are non-comparable concepts, it does not make sense to calculate the arithmetic mean in the form of (F +

D) / 2, so that the two components, frequency and dispersion, influence the result of Importance by multiplication, not by addition.[7]

The merit of the "Importance" (I) is its normalized characteristic, which fluctuates between 0 and 1. See the following table and graph:

VOCABLO	*Mean*	*Disp.*	*Us.*	*M.L.*	*M.R.L.*	*Imp.*
el - la - lo	18157.3	0.899	16321.1	4.259	1.000	.952
de	7482.2	0.897	6709.3	3.874	.910	.903
y	3244.2	0.946	3068.7	3.511	.824	.854
a	3236.7	0.958	3100.5	3.510	.824	.883
lo - la - le	2902.3	0.888	2578.5	3.463	.813	.882
en	2423.8	0.944	2287.7	3.385	.795	.866
que	2042.4	0.805	1643.2	3.310	.777	.857
ser	1999.8	0.927	1853.6	3.301	.775	.848
un - una - uno	1858.8	0.946	1758.1	3.269	.768	.844
se	1544.1	0.927	1431.4	3.189	.749	.782
qué	1390.3	0.81	1126.4	3.143	.738	.952
no	1303	0.816	1063.5	3.115	.731	.903
(…)						

Table 4. Mean, Dispersion, Usage, Mean in logarithm (M.L.), Mean Ratio in logarithm (M.R.L..), Importance (Imp.) / Initial part

[7] Generally, the addition is made for the additive values of the same concept, for example, frequency of words, hours of study, number of daily classes, etc.

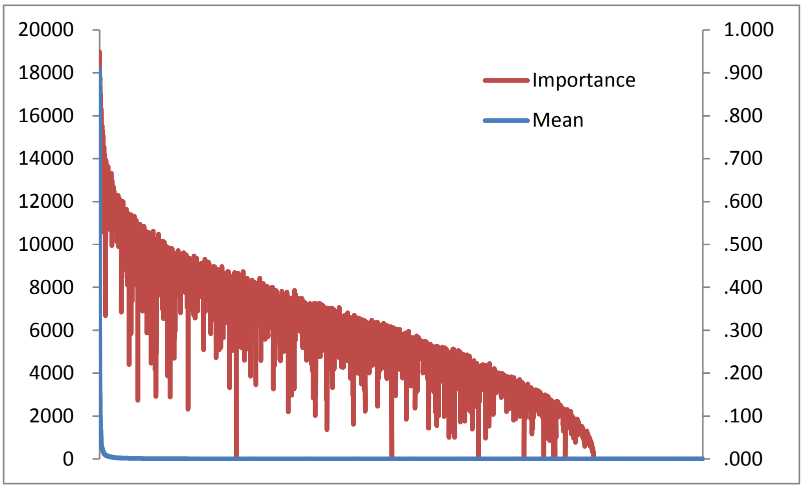

Figure 6. Mean and Importance in comparison

In this way, the mean in the form of Mean Ratio in Logarithm (M.R.L.) ranges from 0 to 1, as well as the Dispersion (D) and, consequently, the new coordinate index of frequency and dispersion in form of "Importance" (I), which does not correlate with frequency (or mean), as in Usage (U). See the following table, which shows the exact match of Usage and Mean (1.000), and the almost null correlation between Usage and Dispersion (.079). The correlations between Importance and Mean — Dispersion are inverse (.175 - .629):

Correlation	*Mean*	*Disp.*	***Usage***	*M.L.*	*M.R.L.*	***Imp.***
Mean	1.000	.078	**1.000**	.256	.269	**.175**
Disp	.078	1.000	**.079**	.481	.467	**.629**
Usage	1.000	.079	1.000	.250	.263	.172
M.L.	.256	.481	.250	1.000	.992	.954
M.R.L.	.269	.467	.263	.992	1.000	.939
Imp.	.175	.629	.172	.954	.939	1.000

Table 5. Correlation between Mean, Dispersion, Usage, Mean in logarithm (M.L.), Mean Ratio in logarithm (M.R.L..) and Importance (Imp.)

3.4 Logarithmic frequency

Previously, in §3.1, the difference between the raw mean and the logarithmic mean, and the convenience of using the latter has been observed because of its softening characteristic of the extreme bias of the raw mean. Naturally the same can be said of the score of all fields, as presented in the following table. See Table 1 (§3.1):

VOC.	*Car*	*Pe.1*	*Ofi.*	*Lib.*	*Cie.*	*Dra.*	*Fic.*	*Ens.*	*Pe.2*	*Téc.*	*Pri.*	*Man.*	*Mál.*	*Alm.*	*Dav.*
el - la - lo	4.32	4.41	4.39	4.36	4.10	4.28	4.34	4.38	4.36	4.37	3.93	4.26	3.85	3.98	4.03
de	3.72	4.07	4.11	3.85	3.95	3.60	3.88	3.95	4.00	4.00	3.41	3.76	3.55	3.85	3.82
que	3.31	3.31	3.31	3.31	3.31	3.38	3.14	3.22	2.80	3.16	3.18	3.18	3.66	3.31	3.52
y	3.58	3.47	3.54	3.69	3.43	3.36	3.51	3.49	3.48	3.52	3.60	3.40	3.58	3.44	3.45
a	3.57	3.59	3.49	3.52	3.60	3.53	3.53	3.48	3.56	3.51	3.35	3.58	3.40	3.44	3.42
en	3.28	3.51	3.46	3.38	3.40	3.20	3.42	3.44	3.44	3.50	3.25	3.23	3.29	3.41	3.40
un - una - uno	3.23	3.20	2.98	3.32	3.25	3.27	3.41	3.29	3.22	3.19	3.36	3.28	3.21	3.32	3.36
ser	3.17	3.07	3.12	3.50	3.35	3.46	3.22	3.32	3.20	3.25	3.31	3.36	3.43	3.24	3.27
se	3.01	3.15	3.17	3.40	3.30	3.14	3.24	3.20	3.23	3.22	3.21	3.21	2.73	3.11	3.22
no	3.24	2.76	2.69	2.88	2.46	3.50	3.01	3.03	2.91	2.93	2.92	3.30	3.54	3.07	3.11

Table 6. Logarithmic frequency / Initial part

The following two tables show the values of the frequency and dispersion indices, and their correlation:

VOCABLO	*Mean*	*Disp*	*Uso*	*M.R.*	*Imp.*
el - la - lo	4.259	0.988	4.209	1.000	0.994
de	3.874	0.987	3.823	0.910	0.947
que	3.511	0.985	3.458	0.824	0.901
y	3.510	0.994	3.488	0.824	0.905
a	3.463	0.995	3.444	0.813	0.899
en	3.385	0.992	3.359	0.795	0.888
un - una - uno	3.310	0.992	3.284	0.777	0.878
ser	3.301	0.990	3.269	0.775	0.876
se	3.269	0.988	3.229	0.768	0.871
no	3.189	0.975	3.110	0.749	0.855

Correl.	*Mean*	*Disp*	***Uso***	*M.R.*	***Imp.***
Mean	1.000	.750	**.991**	1.000	**.959**
Disp	.750	1.000	**.734**	.750	**.856**
U	.991	.734	**1.000**	.991	**.932**
M.R.	1.000	.750	**.991**	1.000	**.959**
Imp	.959	.856	**.932**	.959	**1.000**

Table 7. Logarithmic Frequency: Mean, Dispersion (Disp.), Usage, Mean Ratio (M.R..), Importance (Imp) / Correlation

It should be noted that the Usage maintains an extremely high correlation with the Mean (.991), to the detriment of the correlation with the Dispersion (.734). But this trend has not been as notable as in gross frequency. The composite index of Importance (Imp.) shows a high correlation with both the Mean (.959) and the Dispersion (.856). The mean ratio (M.R.) has been calculated simply by dividing the mean by the maximum mean (4.259), without having to resort to the logarithmic mean ratio, since it is the logarithmic score, free of biased character.

3.5 Ordinal frequency

For obtaining the straight linearity of distribution of the score, we have elaborated the ascending statistical ordinal score:

WORD	*Car*	*Pe.1*	*Ofi.*	*Lib.*	*Cie.*	*Dra.*	*Fic.*	*Ens.*	*Pe.2*	*Téc.*	*Pri.*	*Man.*	*Mál.*	*Alm.*	*Dav.*	*Mean*
el - la - lo	5079	5079	5079	5079	5079	5079	5079	5079	5079	5079	5079	5079	5079	5079	5079	5079
de	5078	5078	5078	5078	5078	5078	5078	5078	5078	5078	5077	5078	5076	5078	5078	5078
que	5074	5073	5073	5070	5072	5073	5069	5071	5062	5069	5071	5068	5078	5073	5077	5077
y	5076	5074	5076	5077	5075	5072	5075	5076	5075	5077	5078	5075	5077	5076	5076	5076
a	5075	5077	5075	5075	5077	5077	5076	5075	5076	5076	5075	5077	5073	5077	5075	5075
en	5073	5076	5074	5072	5074	5068	5074	5074	5074	5075	5073	5070	5072	5075	5074	5074
un - una - uno	5070	5072	5067	5071	5070	5070	5073	5072	5071	5070	5076	5072	5070	5074	5073	5073
ser	5068	5068	5069	5074	5073	5075	5071	5073	5070	5072	5074	5074	5074	5072	5072	5072
se	5063	5070	5072	5073	5071	5066	5072	5070	5072	5071	5072	5069	5045	5070	5071	5071
no	5071	5063	5061	5065	5055	5076	5065	5066	5065	5063	5068	5073	5075	5069	5070	5070

Table 8. Ordinal Frequency / Initial Part

The frequency (mean) and dispersion indices are as follows:

WORD	*Mean*	*Disp.*	*Us.*	*M.R.*	*Imp.*
el - la - lo	5079.0	1.000	5079.0	1.000	1.000
de	5077.8	1.000	5077.7	1.000	1.000
que	5071.5	1.000	5070.5	.999	1.000
y	5075.7	1.000	5075.3	.999	1.000
a	5075.7	1.000	5075.4	.999	1.000
en	5073.2	1.000	5072.7	.999	1.000
un - una - uno	5071.4	1.000	5070.8	.999	1.000
ser	5071.9	1.000	5071.3	.999	1.000
se	5068.5	1.000	5066.7	.998	1.000
no	5067.0	1.000	5065.6	.998	1.000

Correl.	*Mean*	*Disp.*	***Us.***	*M.R.*	***Imp.***
Mean	1.000	.873	**.998**	1.000	**.961**
Disp.	.873	1.000	**.894**	.873	**.966**
Us.	.998	.894	**1.000**	.998	**.967**
M.R..	1.000	.873	**.998**	1.000	**.961**
Imp.	.961	.966	**.967**	.961	**1.000**

Table 9. Ordinal Frequency: Mean, Dispersion (Disp.), Usage, Mean Ratio (M.R.), Importance (Imp) / Correlation

The relationship between Usage and Mean/Dispersion has been greatly improved, demonstrating the convenience of using ordinal frequencies for Usage. On the other hand, the Importance (Imp.) shows the almost perfect correlation with both the Mean (.961) and the Dispersion (.966).

Correlations between the three types of frequency (gross, logarithmic and ordinal) and Usage/Importance indexes have been observed, according to which the optimum combination is the ordinal frequency with the Importance index. The selection of the ordinal frequency is more effective than that of the Importance index.

3.6 Mean and dispersion

The three mentioned works, Juilland & Chang-Rodríguez (1964), Ueda (1987) and Ávila Muñoz (1999), have observed the distribution of frequencies according to the two parameters: absolute frequency and dispersion, based on the hypothesis that the two parameters are not correlated. Indeed, if we calculate the correlation between the two, the result is 0.083, which demonstrates the practically non-correlation between them.

On the other hand, Itō (2002: 220-221) comments that the well distributed words in general have some relation with their frequency. For example, the words that appear in all four Japanese literary works that are treated in his research, and also the words that appear in three of them, are the highly frequent words in general.

In order to check the relationship between the Frequency (F) and the Dispersion (D) in a spread form, the values have been divided into 10 blocks according to the ascending ordinal of the two variables. A cross table has been prepared, where M:10 represents the highest Mean block and D:10, the highest Dispersion block:

Mean : Disp.	*D:1*	*D:2*	*D:3*	*D:4*	*D:5*	*D:6*	*D:7*	*D:8*	*D:9*	*D:10*	*Total*
M:1	121	145	142	62	27	7	3	0	0	0	507
M:2	164	100	72	89	48	20	13	3	0	0	509
M:3	116	68	79	78	71	46	28	13	5	3	507
M:4	33	53	44	72	92	86	55	51	20	2	508
M:5	29	37	56	52	66	69	98	58	34	9	508
M:6	18	32	36	50	54	72	84	93	50	19	508
M:7	9	33	21	42	58	71	76	77	88	33	508
M:8	12	15	33	31	43	62	61	86	99	66	508
M:9	4	19	16	30	32	55	49	78	115	110	508
M:10	1	6	5	6	17	20	41	49	97	266	508
Total	507	508	504	512	508	508	508	508	508	508	5079

Table 10. Mean (M) and Dispersion (D) in ascending ordinal blocks

The cross table shows that there is a strong tendency of correlation between the Mean (M) and the Dispersion (D). One thing is the correlation calculated with the raw frequencies and the gross dispersion (.083) and another is the observation of the panorama in blocks (Table 10).

As an example, the 266 words belonging to block M:10 and D:10; i.e. the block of the maximum mean and the maximum dispersion, are:

> *el - la - lo, de, que, y, a, en, un - una - uno, ser, se, lo - la - le, haber, por, su, con, para, como, tener, estar, más, todo, pero, hacer, este - éste, decir, poder, él - ella - ello, otro, si, ver, qué, sin, dar, vez, sobre, también, año, entre, alguno, mismo, hasta, dos, primero, nos, grande, desde, así, ni, cosa, llegar, pasar, tiempo,*

deber, poco, nuestro, tanto, poner, parte, hombre, después, parecer, hablar, dejar, quedar, vida, siempre, nuevo, aquel - aquél, cada, llevar, seguir, algo, país, encontrar, menos, volver, momento, llamar, sino, aunque, cómo, hacia, mano, antes, mundo, sentir, caso, tres, tomar, tal, mejor, cierto, conocer, mayor, durante, último, propio, problema, lugar, quien, persona, luego, tratar, casi, nunca, nosotros, mientras, ninguno, además, hora, trabajo, manera, importante, fin, cualquier(a), empezar, hoy, palabra, contar, existir, largo, lado, entrar, medio, pequeño, cambio, único, buscar, cuenta, perder, alto, bajo, dentro, permitir, frente, todavía, modo, entender, preguntar, realidad, pueblo, nombre, cuatro, según, cabeza, comenzar, considerar, posible, recordar, pedir, vario(s), ocurrir, caer, convertir, presentar, segundo, abrir, acabar, terminar, mes, mantener, calle, principio, tierra, distinto, resultar, razón, paso, guerra, conseguir, morir, tema, leer, sacar, blanco, crear, cambiar, cara, incluso, servir, necesitar, quizá(s), lograr, cinco, muerte, partir, pie, interés, escuchar, levantar, camino, ganar, condición, alcanzar, explicar, final, demasiado, fondo, demás, tocar, comprender, mal, actividad, reconocer, vista, zona, centro, aceptar, difícil, resultado, descubrir, mostrar, junto, cumplir, dónde, intentar, dirigir, mayoría, principal, pesar, menor, cuál, decidir, nacer, cuestión, edad, continuar, duda, falta, mover, cerca, apenas, diez, ofrecer, soler, línea, real, solamente, subir, libre, experiencia, acercar, usar, seis, suceder, ayudar, tercero, señalar, dedicar, representar, ocupar, pasado, siquiera, resto, responder, capaz, contrario, sufrir, repetir, imaginar, encima, minuto, joven, evitar, sangre, ocasión, corazón.

The words of Mean:10 and Dispersion:1-5 are 35:

sí, te, pues, usted(es), donde, eh, mencionar, tu, ahí, tú, niño, gustar, trabajar, claro, recibir, económico, libro, ah, campo, nacional, orden, papel, comer, animal, jugar, cultura, presidente, servicio, mesa, suelo, producción, estructura, sector, teoría, vos.

The words Mean:3-5 and Dispersion:10 are 14:

harto, amargo, sobrar, chocar, tránsito, acertar, espléndido, valiente, injusticia, sincero, guiar, orgullo, solo : sólo, labrador.

And finally, the words of Mean:1 and Dispersion:1 are 121:

auditorio, cabildo, caducidad, colecta, conciliación, consorte, decimal, decretar, deliberación, deslumbrar, disgustar, divulgación, emocionante, emperatriz, esquiar, felizmente, hebreo, misericordia, organizador, parcelación, penitencia, referéndum, refuerzo, veterano, viejecita, vuecencia, húngaro, lácteo, librito, nochebuena, parejo, ponencia, resonar, adjudicación, aeródromo, alhaja, alias, almanaque, anticipación, antojo, atrio, cafetería, cernir, cochero, comitiva, contencioso, descomposición, domiciliar, donativo, entresuelo, evadir, festividad, hurto, interrogatorio, miliciano, permutación, pesimista, plácido, potestad, predecesor, predicación, promulgar, resfriado, sacerdotal, sensacional, tardanza, recluta, reposición, sacramento, subscripción, aborrecer, angosto, bravura, coacción, epopeya, nevar, opresión, toreo, virrey, voluptuosidad, yanqui, suplicio, acompañante, contradictor, esclarecer, orbe, alteza, apañar, cabalmente, cigarrillo, conjurar, cuartito, despacito, entrometer, fronda, inacabable, innegable, labriego, mañanero, nuca, presuntuoso, turquesa, adusto, arquetipo, averiguación, béisbol, boceto, caramba, coetáneo, desinterés, ensombrecer, inofensivo, lechuza, munición, naipe, obsesionar, remotísimo, repulsión, esquí, submarino, tocadiscos.

The correlative tendency between frequency and dispersion is verified in these classified words. There are numerous correlative words of M:10 and D:10 (266) and others of M:1 and D:1 (121), while the number of discordant words is reduced: M:10 and D:1- 5 (35), and M:3-4 and D:10 (14).

In §3, the differences of three types of score have been observed: absolute frequency (§3.1.), logarithmic frequency (§3.4.) and ordinal frequency (§3.5.) in the form of mean, mean in logarithm, mean in ordinal. The raw means correspond to the frequencies without loss of statistical information, since the mean is mathematically derived from the frequency divided by the number of data. On the other hand, the frequency (or the mean) in logarithm loses the information of the intervals between the high values of the frequencies. The loss of information is greater in the case of ordinal frequencies (or means). Now, instead of gross frequencies, what is counted is simply the place occupied by each word on the ordered scale in an equidistant and sequential manner. However, the merit of the ordinal frequency or mean is that its distribution is presented in a straight line, which is convenient to perform the ranking equation, to calculate the correlation and even to carry out a multivariate analysis, which will be seen in the next section (§4).

4. Principal Component Analysis

4.1 Method

Previously, in §2, the studies referring to the frequencies of the Spanish vocabulary carried out with varied materials have been revised. Each work maintained its criterion of data selection, but when constructing a comparative table, the problem of heterogeneity and of biased and incomplete character of the data appears, despite the homogeneity and the balance maintained within each work. It is doubtful whether in the assembled fifteen fields the frequency of each word is represented in a correctly balanced manner. Another problem is the possibly high correlation among the selected fields, which would decrease the corresponding representative value.

In fact, the project of the statistical study of vocabulary is usually undertaken with pre-established criteria, as observed in §2. After ending word counting, statistical treatments are performed according to the same criteria initially established to draw a statistically reliable conclusion. However, if the pre-established criterion is used, the same criterion must be constantly applied, which makes it difficult to obtain another view of data. It is also doubtful whether the prior selection criterion was reasonable in the light of the research result.

One of the solutions to the problem of the heterogeneous data set and to the pre-established criterion, being possibly unsatisfactory in data from different sources, is the Principal Components Analysis (PCA), devised by K. Pearson and developed by H. Hotelling in early twentieth century. To this method Jolliffe (2002) has devoted a whole volume, which summarizes the method at the beginning of its introduction (ibid. 1):

> The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.

Peña (2002: 133-170) sums up what is applied in the work that is being developed (2002: 134):

> The principal component analysis has this objective: given *n* observations of *p* variables, it is analyzed if it is possible to adequately represent this information with a smaller number of variables constructed as linear combinations of the originals. For example, with variables with high dependence, it is common for a small number of new variables (less than 20% of the originals) to account for the most part (more than 80% of the original variability). (...) Its utility is double: 1. It allows us to represent optimally in a space of small dimension observations of a general *p*-dimensional space. In this sense, principal component analysis is the first step to identify possible latent or unobserved variables that generate the data. 2. It allows the original, generally correlated variables to be transformed into new variables that are uncorrelated, facilitating the interpretation of the data.

Applications of the method to linguistic issues are found in Woods, Fletcher & Hughes (1986: 273-290), Ishikawa, Maeda & Yamazaki (2010: 193-217) and Ueda

(2008). Referring to mathematical derivations to ADDENDA-2, it should be noted that in the present study we apply it to the ascending ordinal data for its convenience and the lack of convenience of other values, the absolute frequency and the frequency in logarithm, both explained in §3.1 and §3.4.[8]

After several attempts to analyze word frequency data and to compare results, it has been concluded that the analysis based on ascending ordinal frequencies is significantly feasible. The following table shows the eigenvalues of the first five components, together with their corresponding Ratio and Accumulated ratio, obtained as results of the Principal Component Analysis of 15 fields plus another, the Mean,[9] with 5,079 words:

Component	*#1*	*#2*	*#3*	*#4*	*#5*
Eigenvalue	9.205	1.336	1.091	.689	.537
Ratio	.575	.083	.068	.043	.034
Cumulative ratio	.575	.659	.727	.770	.804

Table 12. Own value, Ratio and Cumulative ratio

This table allows us to evaluate the obtained components, which offer the weights of each field that are to be assigned to the values of the words. The eigenvalue of component 1 (#1) is 9.205, which stands out among all components. It is so important that it occupies the Ratio of .575 (57.5%). It is followed by the component 2 (# 2) with its own value of 1.336. Its contribution is .083 (8.3%). When adding the two components (#1 and #2), the accumulated ratio reaches .659 (65.9%). Peña (2002) requires more than 80% of variability, while Ishikawa, Maeda & Yamazaki (2010: 203) comment that in general the first components with more than 60 ~ 80% will be included. On the other hand, the three Japanese authors point out that the eigenvalue must be greater than 1, as their sum is equal to the number of fields. Since in theory the analysis presents the same number of principal components as the fields, the

[8] Absolute frequency data are not adequate because of the highly biased distribution, which has been discussed in §3.1. The frequency data in logarithm, in addition to having the same biased characteristic, of lower degree, presents the impossibility to distinguish between the frequency 1 and 0, since it is the logarithm, which does not allow the value 0 in true number.

[9] The last field. Mean of 15 fields, is included, which serves as a midpoint reference.

component with the eigenvalue less than 1 means that it carries less information from an original field, i.e., it does not imply relative merit. Fortunately, the second component (#2) exceeds 1 (1.336), so it has been decided to include up to the second component. However, it is always convenient to keep in mind that the first component (#1) is much more important than the second component (#2).

4.2 Fields

The method of Principal Component Analysis starts from the symmetric matrix correlation of the fields, which are 16 in our case:[10]

*	*Car*	*Pe.1*	*Ofi.*	*Lib.*	*Cie.*	*Dra.*	*Fic.*	*Ens*	*P.2*	*Téc.*	*Pri.*	*Man.*	*Mál.*	*Alm.*	*Dav.*	*Mean*
Car	1.00	.50	.42	.50	.30	.52	.46	.39	.43	.33	.59	.58	.58	.54	.55	.69
Pe.1	.50	1.00	.63	.52	.43	.36	.36	.47	.58	.41	.54	.49	.52	.67	.62	.72
Of.i	.42	.63	1.00	.40	.43	.24	.19	.40	.49	.40	.40	.31	.43	.58	.51	.61
Lib	.50	.52	.40	1.00	.55	.51	.55	.57	.55	.52	.62	.51	.51	.60	.66	.77
Ci.e	.30	.43	.43	.55	1.00	.27	.36	.55	.51	.65	.48	.35	.41	.57	.65	.67
Dra	.52	.36	.24	.51	.27	1.00	.67	.51	.48	.35	.59	.57	.52	.49	.55	.68
F.ic	.46	.36	.19	.55	.36	.67	1.00	.59	.53	.44	.61	.56	.47	.50	.60	.71
Ens	.39	.47	.40	.57	.55	.51	.59	1.00	.70	.64	.55	.45	.45	.60	.66	.77
Pe.2	.43	.58	.49	.55	.51	.48	.53	.70	1.00	.59	.55	.47	.48	.64	.65	.78
T.éc	.33	.41	.40	.52	.65	.35	.44	.64	.59	1.00	.47	.36	.40	.52	.58	.69
Pr.i	.59	.54	.40	.62	.48	.59	.61	.55	.55	.47	1.00	.71	.70	.67	.72	.82
Man	.58	.49	.31	.51	.35	.57	.56	.45	.47	.36	.71	1.00	.65	.59	.63	.73
Mál.	.58	.52	.43	.51	.41	.52	.47	.45	.48	.40	.70	.65	1.00	.67	.68	.76
Alm	.54	.67	.58	.60	.57	.49	.50	.60	.64	.52	.67	.59	.67	1.00	.89	.86
Da.v	.55	.62	.51	.66	.65	.55	.60	.66	.65	.58	.72	.63	.68	.89	1.00	.89
Mean	.69	.72	.61	.77	.67	.68	.71	.77	.78	.69	.82	.73	.76	.86	.89	1.00

Table 13. Correlation between 16 fields

from which we can observe the descending ordering of the correlation coefficients with respect to Mean:

[10] For the legend of the fields, see the note of §3.1.

Mean	Dav.	Alm.	Pri.	Pe.2	Lib.	Ens.	Mál.	Man.	Pe.1	Fic.	Téc.	Car	Dra.	Cie.	Ofi.
1.00	.89	.86	.82	.78	.77	.77	.76	.73	.72	.71	.69	.69	.68	.67	.61

by which we know that the correlation with Mean is decreasing from Davies (Dav.) to Official document (Ofi.). This analysis is valid, if our interest is only in the variable Mean. In contrast, Principal Component Analysis does not extract only one variable, but treats all variables in order to know the strongest relationships between them.

The Table 14 shows the numerical constitution of the weights of the first two components (#1, #2), with respect to the 16 fields.

Co.	*Car*	*Pe.1*	*Ofi.*	*Lib.*	*Cie.*	*Dra.*	*Fic.*	*Ens*	*P.2*	*Téc.*	*Pri.*	*Man.*	*Mál.*	*Alm.*	*Dav.*	*Mean*
#1	.220	.240	.198	.252	.220	.223	.230	.251	.253	.223	.272	.242	.251	.288	.302	.304
#2	-.300	.160	.344	.013	.380	-.400	-.310	.148	.169	.328	-.218	-.342	-.200	.094	.06	.069

Table 14. First and second components (Co.)

To interpret the distribution of each component it is convenient to prepare the independent graphs:

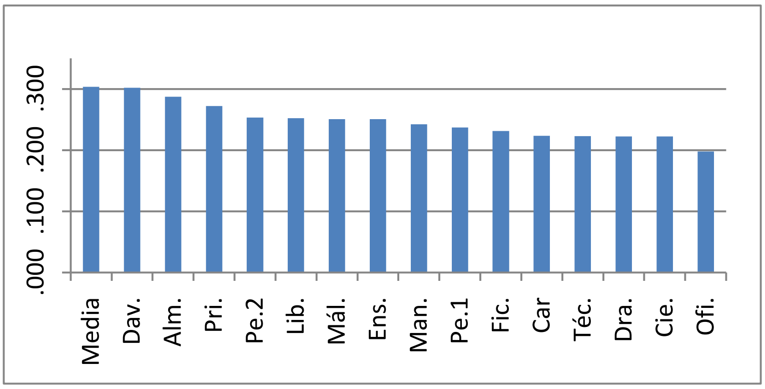

Figure 11. First component: generality

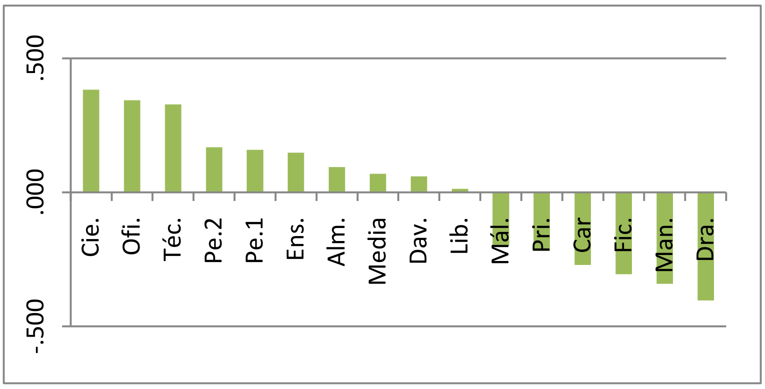

Figure 12. Second component: formality

The ordering of the first component: Mean, Davies, Almela, Primary, Periodical-2, Book, Malaga, Essay, Manual, Periodical-1, Fiction, Letter, Technical, Drama, Science, Official, is interpreted as the degree of "generality"; that is to say, the Mean represents the generality par excellence, followed by Davies (Dav.) and Almela (Alm.) who use the balanced compound data. The essays of the students of the primary school (Pri.) would use some common words. Periodical-2 (Pe.2) is of Juilland & Chang-Rodríguez (1964), which includes a wide range of periodicals, while Periodical-1 (Pe.1), by García Hoz (1953), is limited to daily newspapers of Spain from 1943 to 1948. Books (Lib.), the speech of Malaga (Mál.), Essays (Ens.) and Manuals of Spanish (Man.) are materials that manifest the variety of Spanish, neither general nor specific. The remaining ones, Fiction (Fiction), Letter (Car.), Technical documents (Tech.), Drama (Dra.), Science (Cie.) and Official documents (Ofi.) are in the line that goes from "generality" to end of 'specialty'.

According to the second component (#2), the fields are aligned in the gradation from "formality" to "familiarity": Science, Official, Technical, Periodical-2, Periodical-1, Essay, Almela, Mean, Davies, Letter, Fiction, Manual and Drama. The same gradation from "formality" to "familiarity" (# 2: -400 ~ .380) is more notable than "generality" to "specialty" (#1: .198 ~ .304), which can be observed in Fig.11 and 12.

In §4.1 it has been observed that the first component carries the largest amount of information (57.5%), followed by the second with a much smaller amount of information (8.3%), which can be seen in the following graph:

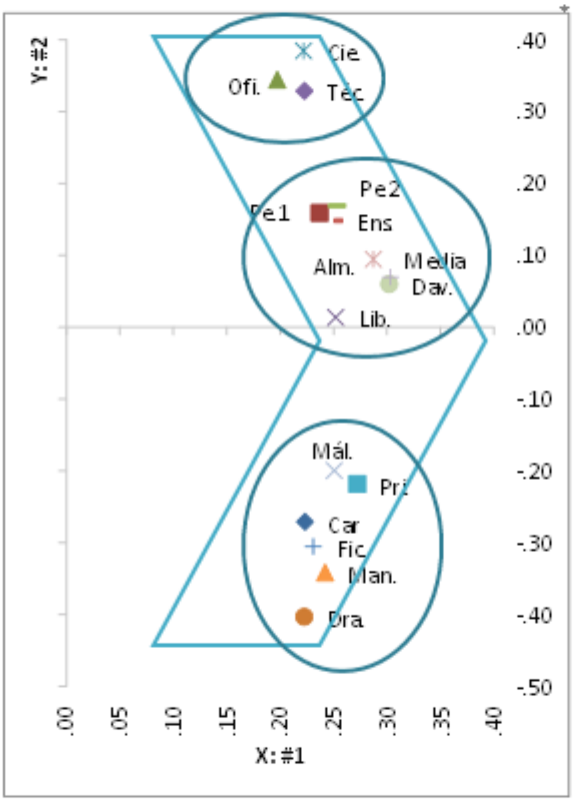

Figure 13. First and second components

It is not a question of classification into four groups: “general-formal”, “general-familial”, “special-formal” and “special-familial”, but the tripartite division. In the first place, it is divided between “general” (Mean, Davies, Almela, Periodical-2, Book, Periodical-1) and “special”, and secondly, “special” is subdivided into “formal” (Official, Cience, Technical) and “familiar” (Drama, Manual, Fiction, Letter, Primary, Malaga). Some fields are in the overlap zone, which is resolved in the two-dimensional space as seen in the previous graph, where grouping of fields is easy to recognize.

4.3 Words

The Principal Component Analysis facilitates not only the general overview of fields but also individual data, that is, words. Along with the field classification that has just been made in §4.2., it is also possible to carry out the word classification.

In the first place, the “general” words are distinguished, which take on greater weight from the “general” fields. For lack of space, only the first 100 "general" words are listed:

> *el - la - lo, de, y, a, en, lo - la - le, ser, que, un - una - uno, se, haber, no, por, qué, con, para, su, tener, estar, este - éste, todo, más, hacer, como, decir, poder, él - ella - ello, pero, ir, otro, ya, dar, mucho, o, ver, si, muy, nos, tanto, dos, me, también, vez, año, porque, sin, grande, día, mismo, hasta, bien, saber, ese - ése, pasar, sobre, llegar, primero, así, nuestro, querer, aquel - aquél, poner, sí, tiempo, desde, hombre, poco, alguno, entre, después, ni, siempre, cosa, cómo, llevar, pues, vida, quedar, tres, bueno, parte, encontrar, venir, nuevo, ahora, seguir, llamar, dónde, cada, dejar, mi, parecer, menos, mundo, hablar, salir, conocer, antes, casa, sino.*

Within the other “special” field, the hundred words belonging to the “formal” group are:

> *desarrollo, efectuar, técnica, universal, integrar, unidad, originar, producción, industria, caracterizar, derivar, orgánico, cifra, existente, occidental, fase, precedente, decreto, organismo, realización, constituir, estructura, previo, totalidad, vital, decisivo, sucesivo, apertura, formación, concurrir, aportación, posterior, interpretación, análogo, intelectual, directo, distribución, geográfico, respectivamente, longitud, proyección, manifiesto, definir, individuo, institución, fundamental, preparación, sucesión, administrativo, eliminar, auténtico, minero, contenido, origen, tendencia, característica, crecimiento, actualidad, relativo, modificación, célula, ambos, objetivo, efectivo, concesión, explotación, perí[í]odo, beneficiar, potencia, virtud, adoptar, poseer, diverso, proceso, atribuir, correspondiente, dominio, conclusión, defensa, vigente, capacidad, continuación, contemporáneo, núcleo, región, doctrina, denominar, aplicación, elemento, método, reaccionar, lesión, eclesiástico, marítimo, órgano, dotar, experimentar, norma, secundario, químico.*

The hundred words of the third group “familiar” are:

adiós, guapo, cenar, alegrar, señorita, tío, ti, beso, tú, chico, broma, novio, bailar, tontería, sobrino, bonito, contigo, hola, enfadar, gana, coger, sopa, lástima, ah, mamá, despacio, desayunar, boda, cansar, bolsillo, llave, coche, cama, avisar, acostar, doler, divertir, adonde, zapato, baile, molestar, abuelo, carro, reír, orgulloso, agradar, encantar, asustar, sueño, paciencia, rato, pollo, perdonar, tirar, teléfono, contento, billete, extrañar, estupendo, mozo, saludar, tonto, padrino, comedor, conmigo, quejar, oh, gordo, pasear, ofender, ay, aburrir, pañuelo, vergüenza, restaurante, coser, prisa, despedir, charlar, llover, miedo, abrigo, desayuno, almorzar, papá, pegar, risa, cuesta, arreglar, paseo, llorar, caro, tienda, meter, distraer, perdón, gustar, susto, vera, poquito.

5. Conclusion

In the introduction (§1), the “post-categorization” method has been discussed, which consists in searching the category from the analyzed data, rather than pre-setting the classification that would serve to carry out the analysis. It is a question of looking for a small number of new variables that would explain most of the variation of the data. Since one of the most common methods that are applied for the same purpose is Principal Component Analysis, we have elaborated our own program (see ADDENDA-2).

The same method starts from the symmetric correlation matrix. However, the gross (absolute) frequency data of the words are not adequate as they have a highly skewed distribution. The transformation into logarithm also does not serve to obtain an ideally balanced distribution. The only conversion that is convincing has been to ascending ordinal values. The same conversion has also been useful in classifying the vocabulary according to frequency and dispersion (§3).

As regards “post-categorization”, referred to in §1, it has been concluded that the words treated in various previous studies (§2) are classified according to the three mutually uncorrelated components: “general”, “special-formal”, “special-familiar”, whose constituents represent the specific weights assigned to each field (§4).

It is convenient to see the validity of the analysis not only within the data treated (§2) but also with the data that will be presented in future. Currently computer tools are being developed that facilitate the collection of lexical data in different fields both in synchrony and in diachrony of the various languages.[11] The relevant information about this and other subjects will be given in later works.[12]

References

ALMELA, Ramón; Pascual CANTOS, Aquilino SÁNCHEZ, Ramón SARMIENTO & Moisés ALMELA (2005) *Frecuencias del español. Diccionario y estudios léxicos y morfológicos*, Madrid: Editorial Universitas.

ÁVILA MUÑOZ, Antonio Manuel (1999) *Léxico de frecuencia del español hablado en la ciudad de Málaga*, Málaga: Universidad de Málaga.

DAVIES, Mark (2006) *A frequency dictionary of Spanish. Core vocabulary for learners,* New York: Routledge.

GARCÍA HOZ, Víctor (1953) *Vocabulario usual, vocabulario común y vocabulario fundamental*, Madrid: C.S.I.C.

GARCÍA HOZ, Víctor (1976) *El vocabulario general de orientación científica y sus estratos*, Madrid: C.S.I.C.

IKEDA, Hiroshi (1976) *Tōkēteki hōhō. I. Kiso (Métodos estadísticos. I. Fundamento),* Tokio: Shinyosha.

ISHIKAWA, Sinichirō, Tadahiko MAEDA & Makoto YAMAZAKI (2010) *Gengo kenkyū no tameno tōkē nyūmon (Introducción a la estadística para estudios lingüísticos)*, Tokio: Kuroshio Syuppan.

ITŌ, Masamityu (2002) *Kēryō gengogaku nyūmon (Introducción a la lingüística cuantitativa)*, Tokio: Taisyūkan shoten.

[11] An example would be the treatment of the *Diccionari català-valencià-balear*, whose more than 154,000 words are marked, among others, with the following tags: "dialectal", "ancient", loanword" or "vulgar", in addition to their inclusion in more than sixty different semantic fields. Their statistical study would provide interesting information about their lexical composition.

[12] Currently the program, called LYNEAL (*Letras y Números en Análisis Lingüísticos*) works in two places, in Madrid and Tokyo:

http://shimoda.lllf.uam.es/ueda/lyneal/
https://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/

JOLLIFFE, Ian T. (2002) *Principal component analysis*, 2nd edition, New York: Springer.

JUSTICIA, Fernando (1995) *El desarrollo del vocabulario. Diccionario de frecuencias*, Granada: Universidad de Granada.

JUILLAND, Alphonse & CHANG-RODRÍGUEZ, Eugenio (1964) *Frequency dictionary of Spanish words*, The Hague: Mouton.

MORALES, Amparo (1989) "Reseña de Hiroto UEDA, *Frecuencia y dispersión del vocabulario español*", *Lingüística (Asociación de Lingüística y Filológica de la América Latina)*, 1, 282-289.

MINO, Tairai (2001) *Tōkē kaiseki no tameno senkē daisū (Álgebra lineal para análisis estadísticos)*, Tokio: Kyōrityu shuppan.

PEÑA, Daniel (2002) *Análisis de datos multivariantes*, Madrid: McGraw Hill.

SCHOLFIELD, Phil (1995) *Quantifying language. A researcher's and teacher's guide to gathering language data and reducing to figures*, Clevedon: Multilingual Matters.

UEDA, Hiroto (1987, 2007) *Frecuencia y dispersión del vocabulario español*, Instituto de Estudios Lingüísticos: Universidad de Estudios Extranjeros de Tokio.
<https://lecture.ecc.u-tokyo.ac.jp/~cuedákenkyúgoífrec-disp/frec-disp-0.pdf>

UEDA, Hiroto (1993) "Notas sobre lexicometría del español", *Lingüística (Asociación de Lingüística y Filológica de la América Latina)*, 5, 147-154.

UEDA, Hiroto (1999-2007) *Análisis de datos cuantitativos para estudios lingüísticos.*
<https://lecture.ecc.u-tokyo.ac.jp/~cuedágengó4-numeros/doc/numeros-es.pdf>

WOODS, Anthony, Paul FLETCHER & Arthur HUGHES (1986) *Statistics in language studies*, Cambridge: Cambridge University Press.

ADDENDA-1. DISPERSION[13]

Standard Deviation, which is used as an indicator of variation, has the property of increasing according to the scale of the data. For this reason, a constant indicator of variation independent of the data scale has been sought. Consequently, the Coefficient of Variation (CV) is calculated with the Standard Deviation (DT) divided by the Mean (m).

CV = SD / m

Since the Coefficient of Variation (CV) is not normalized, that is, it does not fluctuate between 0 and 1, a standardized variation indicator, which we call "Regular Standard Deviation" (RSD), is formulated, which is calculated by dividing the Standard Deviation (SD) by the maximum value of the same Standard Deviation (SD.max.):

RSD = SD / SD.max.

The way to look for the formula of SD.max. is given below. It starts from the formula of the Standard Deviation (SD), which is a square root of the Variance:

$SD = \{[(x_1 - m)^2 + (x_2 - m)^2 + ... + (x_n - m)^2] / n\}^{1/2}$ m: mean; n: number of data

In a data set with an extreme case of deviation, for example {10, 0, 0, 0, 0}, the maximum value of Standard Deviation (SD.max.) is presented. To generalize the problem, k is used instead of a concrete figure: {k, 0, 0, ..., 0}. Then, only the first term is $(k - m)^2$, and all others are $(0 - m)^2 = m^2$, and therefore the maximum value of standard deviation (SD.max.) is:

$SD.max. = \{([(k - m)^2 + (n - 1) m^2]\}^{1/2}$

where, k is equal to the sum of the data, since the remainder are null. As the sum is equal to the mean (m) multiplied by the number (n) of data (Sum = n m ← m = Sum / n), k is equal to n m:

k = Sum = n m

[13] Partial reproduction of Ueda (1999-2017: 2.6).

Thus,

SD.max = $\{[(n\ m - m)^2 + m^2\ (n - 1)]\ /\ n\}^{1/2}$ ← k = n m

= $\{[(m\ (n - 1))^2 + m^2\ (n - 1)]\ /\ n\}^{1/2}$ ← m outside

= $\{[m^2\ (n - 1)^2 + m^2\ (n - 1)]\ /\ n\}^{1/2}$ ← m^2 is common

= $m^2\ (n - 1)\ [(n - 1) +1]\ /\ n\}^{1/2}$ ← m2 (n - 1) is common

= $\{(m^2\ (n – 1)\ n\ /\ n\}^{1/2}$ ← (n – 1) +1 = n

= $[(m^2\ (n – 1)]^{\ 1/2}$ ← n / n = 1

= $m\ (n – 1)^{1/2}$ ← $(m^2)^{1/2} = m$

Therefore, the Regular Standard Deviation (RSD) is:

RSD = SD / SD.max. = SD / $[m\ (n - 1)^{1/2}]$

The difference between Coefficient of Variation (CV) and Regular Standard Deviation (RSD) is that in the second one the value of $(n - 1)^{1/2}$ is found in the denominator. When it comes to data whose number (n) is high, the RSD becomes small. It is recommended to use RSD not vertically with n individuals, but horizontally with p variables, whose p-number is usually small.

The Dispersion (Disp) is the complement of the Regular Standard Deviation (RSD) with respect to 1, which also fluctuates between 0 and 1:

Disp = 1 - RSD = 1 - SD / $[m\ (n - 1)^{1/2}]$

ADDENDA-2. PRINCIPAL COMPONENT ANALYSIS[14]

Multiplying the values of variables by a vector of weights so that the variance of all variables of the data is maximal and, at the same time, the correlation between all the variables is 0, the variables thus multiplied take a new synthetic meaning. The same weights can also be applied to the data themselves to see their positions within the new synthetic variables. For example, the new variable that shows a high correlation with the Mathematics variable and the Science variable within the test scores is considered as an indicator of Exact Science weight. K. Pearson calls the method Principal Component Analysis.

The new variable that multiplies by a weight so that its variance covers its maximum value facilitates the interpretation of the variation of the data in the best possible way. And another variable that follows in its variance, which presents the second best explanation of the data. The two variables, which show zero correlation (0), offer their own non-overlapping explanations. The number of such new variables is the same as that of the variables of the object data. However, successive variables present less and less variance, which reduces their explanatory capacity, so that it is sufficient to analyze the first new variables.

To calculate, the standard score (Xnp) are prepared with the vector of Vertical Means (Mp) and another of Standard Deviations (Sp):

Xnp = (Dnp - Mp) / Sp

This matrix Xnp is multiplied by the incognito vector Wp to formulate the vertical vector Zn:

[1] Zn = Xnp Wp

We look for the Variance (V) of this composite vector:

[2] V = (Zn^T Zn) / N ← Zn^T is transposed vector of Zn
= $(Xnp\ Wp)^T$ (Xnp Wp) / N ← [1]
= Wp^T Xnp^T Xnp Wp / N ← $(A\ B)^T = B^T A$
= Wp^T (Xnp^T Xnp / N) Wp ← N is scalar, movable

[14] Partial reproduction of Ueda (1999-2017: 5.3.2.).

$= Wp^T$ Rpp Wp ← Rpp = Xnp^T Xnp / N

The condition of Wp is stipulated, whose sum of products is 1. Without this condition, there is a unlimited number of Wp:

[3] Wp^T Wp = 1

With this restriction [3], to find the Maximum of Variance (V) [2], we calculate the differential of F, with the Lagrange multiplier (L), which must be 0, where the inclination is zero:

F = Wp^T Rpp Wp - L (Wp^T Wp-1)

The differential (Df) of F by Wp is:

[4a] Df (F, Wp) = 2 Rpp Wp - 2 L Wp = 0

[4b] Rpp Wp = L Wp ← move 2 L Wp to the right side, divide both sides by 2

This form [4b] is called the Characteristic Equation, from which both the eigenvalue L and the eigenvector Wp are derived. The eigenvalue corresponds to the variance as follows:

V = Wp^T Rpp Wp ← [2]
= Wp^T L Wp ← [4b]
= L Wp^T Wp ← L is scalar, movable
= L ← [3]

There are as many eigenvalues and eigenvectors as variables of the object data, and these vectors are called Components, which are ordered by the magnitude of their corresponding eigenvalue or the Variance.

The lower left table corresponds to the scores of individuals d1 to d7, of variables M: Mathematics, S: Science and L, Latin. The PCA.d table corresponds to the points of the three components, # 1, # 2, # 3. PCA.v is the eigenvector matrix and, finally, PCA.e corresponds to the eigenvalues, i.e. to the Variances of the components:

D	M	S	L
d1	45	48	66
d2	56	59	54
d3	58	51	78
d4	77	72	20
d5	43	44	32
d6	58	34	90
d7	50	53	100

PCA.d	#1	#2	#3
d1	-.823	-.544	.325
d2	.635	-.149	.369
d3	-.176	.588	.007
d4	3.171	.218	-.239
d5	-.510	-1.668	-.270
d6	-1.383	.789	-1.025
d7	-.916	.766	.834

PCA.v	#1	#2	#3
E	.569	.616	-.545
L	.635	.093	.767
S	-.523	.782	.338

PCA.e	#1	#2	#3
E.value	2.026	.672	.303

Program[15]

```
Sub PRINCIPAL_COMPONENT() 'Principal component Analysis (H. Ueda)
 Dim Znp, Rpp, EG, Ep, Epp, Snp, Wp
 Znp = stdV(Inp): Rpp = d(x(t(Znp), Znp), n) 'Standard score: Correlation matrix
 EG = eigen(Rpp): Ep = EG(0): Epp = EG(1) 'Eigenvalue: Eigenvector
 ReDim Onp(3, nC(Ep)) 'Output
 Onp = copyR(Onp, 0, Ep, 0): Onp = copyR(Onp, 1, Ep, 1) 'Eigenvalue
 Wp = d(Ep, nR(Rpp)): Onp = copyR(Onp, 2, Wp, 1): Onp(2, 0) = "Ratio"
 Onp = copyR(Onp, 3, accumH(Wp), 1): Onp(3, 0) = "Ac.Ratio" 'Accumulated ratio
 Call OP(Onp, 0, 0) 'Output: Eigenvalue
 Call OP(Epp, NextR, 0) 'Output: Eigenvector
```

[15] For mathematical derivation and programming, Mino (2001: 155-157) has been consulted. The program in Excel.VBA is the web page:
<https://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo>.