I. LAS LEYES DE ESTOUP-ZIPF Y EL VOCABULARIO LOPE DE RUEDA

1. Pretendemos en este pequeño trabajo aplicar algunas de las llamadas "leyes de (Estoup-) Zipf[1] a un material nuevo para probar s validez. Por el momento no nos interesan aquí las modificaciones introducidas a estas leyes, entre otros, por Guiraud[2], Mandelbrot[3] y Devooght[4].

2. Nuestro material está constituido por el vocabulario[5] total de las obras de Lope de Rueda[6], cuya paternidad no ha sido puesta en duda, esto es:

Comedias: "Comedia llamada Armelina"
"Comedia llamada Eufemia"
"Comedia llamada de Los Engañados"
"Comedia llamada Discordia y Questión de Amor".
Coloquios: "Colloquio de Camila"
"Colloquio de Tymbria"

[1]Vid. ZIPF, G. K., *Relative frequency as a determinant of phonetic change, HSPh'*, XL 1929, 1-95; *Selected studies of the principle of relative frequency in language,* Cambridge-Mass., 1932; *Human behaviour and the princip of least effort. An introduction to human ecology,* Cambridge-Mass., 1949.

[2]Vid. GUIRAUD, P. *Les Caractères statistiques du Vocabulaire,* Paris, . U. F., 1954.

[3]Vid. MANDELBROT, B., *Structure formelle des textes et Communication,* Word 10, 1954, pp. 1-27.

[4]Vid. DEVOOGHT, *Sur la loi de Zipf-Mandelbrot.* Bulletin de la Classe des Sciences de l'Academie Royale de Belgique. 5e. Série, t. 43 (1957, 4, 244-251).

[5]Usamos aquí la distinción de Guiraud entre *vocabulario* de un texto, o sea, el inventario particular de sus elementos lexicales, y *léxico* de una lengua, el inventario general de donde se ha obtenido el vocabulario.

[6]Hemos empleado la edición de EMILIO COTARELO: Madrid, R. A. E., 1908, 2 tomos.

"Colloquio llamado Prendas de Amor"
Colloquio en Verso".
Pasos: "El Deleytoso"
"Registro de Representantes" (Pasos IV, V y VI).

2.1. Mediante el uso de computadoras hemos obtenido todos los elementos lexicales que aparecen empleados en estas obras y su respectiva frecuencia. Llamamos a estas unidades *formas-textuales*. Posteriormente hemos formado conjuntos con todas aquellas formas-textuales que tienen la más estrecha afinidad para el hablante: {canto, -as, -a...}; {canto, -os, -ito}; {bello, -a, -os, -as}; {yo, me, mí}. Cada uno de estos conjuntos está representado por una *forma básica* (CANTAR, CANTO, BELLO, YO), que puede o no ser una forma-textual y que sirve para determinar el conjunto respectivo[7].

3. Zipf nos dice que existe una relación entre el *rango* de un elemento lexical y su *frecuencia* (rango es el número ordinal que resulta al poner las unidades en orden decreciente de frecuencia). Esta relación puede ser expresada mediante la fórmula $R \times F = C$ (constante) [8].

Así, por ejemplo, en el Ulysses de Joyce

	Rango	*Frecuencia*	
La palabra Nº	10	es usada 2.653 veces	26.530
	100	265	26.500
	1.000	26	26.000
	10.000	2	20.000
	29.000	1	29.000

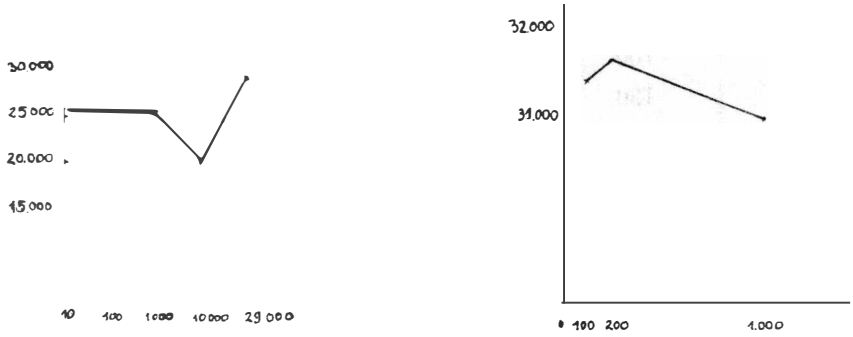
o bien, en el French Word Book de V. A. C. Hemmon

Rango	*Frecuencia*	
100	314	31.400
200	158	31.600
1.000	31	31.000

[7] Esta distinción entre *formas-básicas* y *formas-textuales* es, en principio, la misma que hace Guiraud entre *mots-formes* y *unités de lexique*. También se ha acudido, entre otros, a los términos *parolalemma* y *parole-flesse*.

[8] Tomamos la ejemplificación de Guiraud, op. cit.

Veamos estos mismos datos representados gráficamente.



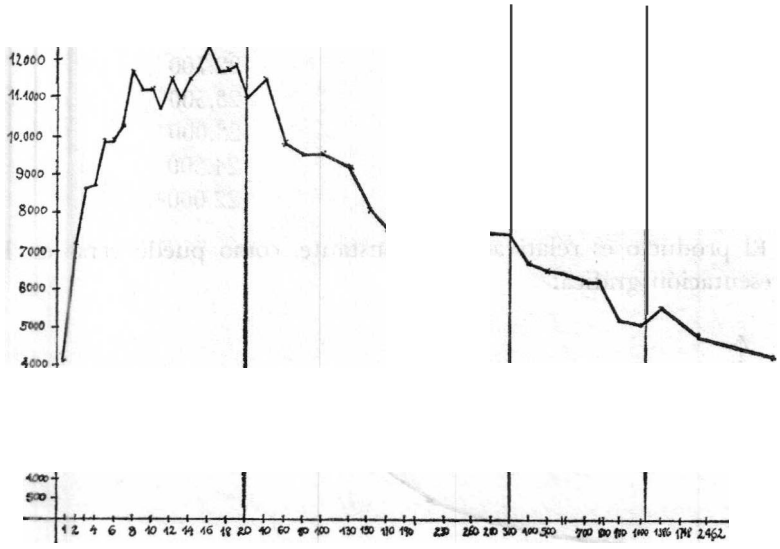
En los gráficos, el Producto forma, simplificando fuertemente, una especie de paralela de la abscisa. Veamos qué sucede si aplicamos la misma fórmula a nuestro material. Doy aquí una muestra representativa:

3.1.	*Rango*	*Frecuencia*	*Producto*
	1	4.181	4.181
	2	3.556	7.112
	3	2.889	8.667
	4	2.176	8.704
	5	1.988	9.940
	6	1.650	9.900
	7	1.472	10.304
	8	1.469	11.752
	9	1.250	11.250
	10	1.126	11.260
	11	982	10.802
	12	964	11.568
	13	867	11.271
	14	822	11.508
	15	781	11.715
	16	777	12.432
	17	694	11.798
	18	658	11.844
	19	630	11.970
	20	551	11.020
	40	288	11.520

60	164	9.840
80	121	9.680
100	96	9.600
130	72	9.360
150	54	8.100
170	45	7.650
190	40	7.600
200	38	7.600
230	34	7.820
260	30	7.800
280	27	7.560
300	25	7.500
320	22	7.040
340	20	6.800
360	19	6.840
380	18	6.840
400	17	6.800
430	15	6.450
460	14	6.440
500	13	6.500
540	12	6.480
590	11	6.490
640	10	6.400
700	9	6.300
790	8	6.320
890	6	5.340
1.040	5	5.200
1.386	4	5.544
1.748	3	5.244
2.462	2	4.924
4.338	1	4.338

La simple lectura de los productos nos muestra que el resultado no es en ningún caso una constante, ni siquiera aproximadamente. Se

trata de algo muy distinto, como se aprecia más claramente en la representación gráfica en la que hemos eliminado algunos datos, que no ofrecen ya mayor interés.



Como vemos, el producto comienza en 4.181 y sigue subiendo abruptamente hasta alcanzar en la forma básica de rango 16 el punto más alto (12.432), luego comienza el descenso, que termina en 4.338, es decir, como había empezado, poco más o menos. No se trata aquí de una paralela a la abscisa como en los casos del Ulysses y del French Word Book, sino de una estructura de ascenso y descenso. La fórmula de Zipf no traduce con fidelidad la situación que encontramos en las formas-básicas de Lope de Rueda.

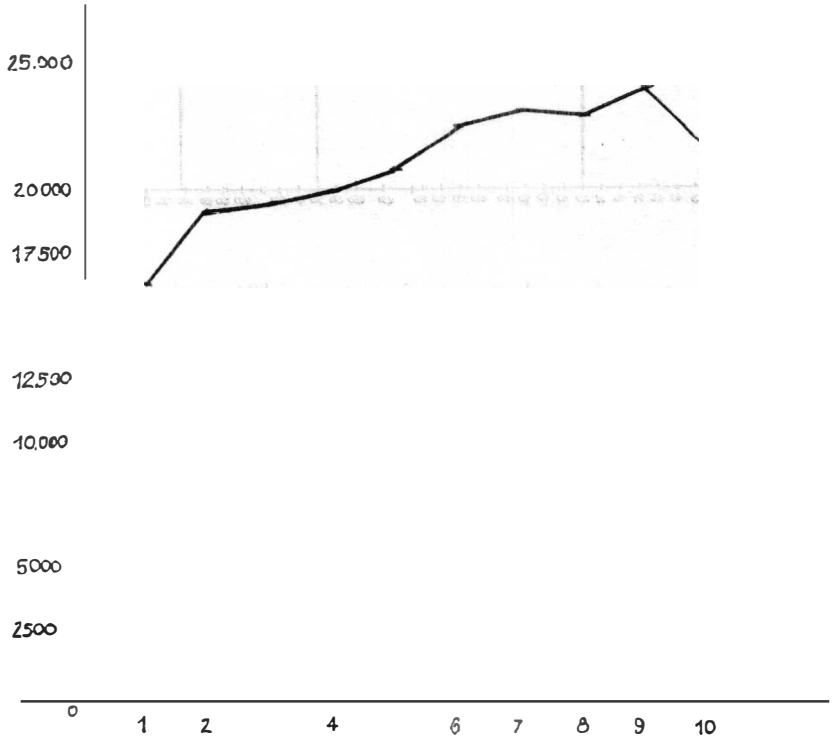
Resulta interesante señalar que el máximo del producto coincide aproximadamente con la mitad de las formas-textuales del vocabulario, ya que éstas suman en números redondos 30.000 al alcanzar la forma-básica de rango 16 y el total son 69.000.

4. Una ley de Zipf que tiene estrecha relación con la anterior es la siguiente: el producto del número de las palabras que son usadas con una misma frecuencia por el cuadrado de dicha frecuencia es constante, esto es $a \times b^2 = C$.

En el Ulysses de Joyce encontramos

a	*b*	
16.432	1	16.432
4.776	2	19.000
2.194	3	19.600
1.400	4	20.239
900	5	22.400
770	6	22.700
480	7	23.500
370	8	23.000
300	9	24.300
220	10	22.000[9]

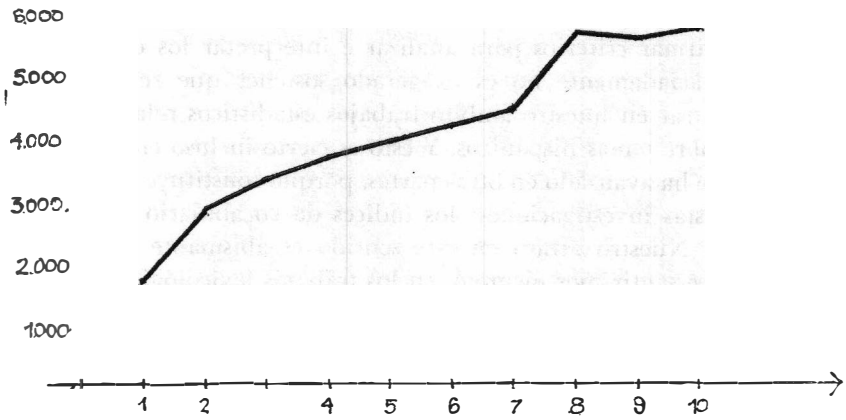
El producto es relativamente constante, como puede verse en la representación gráfica:



[9]Ejemplo cit. por GIULIO LEPSCHY, *La linguistica strutturale* (Torino, Einaudi, 1966), p. 193. Los productos exactos son 19.104, 19.746, 22.400, 27.720, 23.520, 23.680, 24,300, 22.000.

4.1. Apliquemos esta segunda ley al vocabulario de las obras de Lope de Rueda:

a	*b*	
1.876	1	1.876
714	2	2.856
362	3	3.258
230	4	3.680
152	5	3.800
115	6	4.140
91	7	4.459
90	8	5.760
68	9	5.508
56	10	5.600



Hay un claro ascenso que se estabiliza entre los 5 y 6.000, esta situación se mantiene con altibajos luego de la frecuencia 10 (mínimo 2.880 y máximo 8.125 hasta la frecuencia 25). No hay que olvidar que, si bien es cierto que lo más frecuente es que el valor de *a*

descienda, mientras que el de *b* sube, encontramos ya antes de la frecuencia 25 una serie: *a* 15- 13- 9- 11- 5- 13- 7- que rompe toda posible constancia del producto. En las frecuencias más altas es éste el caso más general y así tenemos, por ejemplo:

a	*b*	
1	80	6.400
2	82	13.448
3	83	20.667
1	85	7.225

5. El análisis precedente nos lleva a las siguientes conclusiones:

— Ninguna de estas dos leyes de Zipf funciona sin objeciones serias con las *formas-básicas* del vocabulario de Lope de Rueda.
— Fundamentalmente escapa a la primitiva formulación de estas leyes la gradación ascendente inicial, muy marcada en nuestros dos ejemplos.
— Un juicio definitivo sobre la validez y utilidad de estas leyes sólo será posible luego de su confrontación con los más distintos vocabularios.

II. MISCELANEA NUMERICA A PROPOSITO DE LOPE DE RUEDA

0. En la Lingüística Estadística como en la mayoría de las ciencias uno de los métodos más valiosos es la comparación, que permite crear o confirmar criterios para analizar e interpretar los datos recogidos. Desgraciadamente no es exagerado sostener que resulta muy difícil encontrar en nuestro ámbito trabajos estadísticos relativamente completos sobre temas hispánicos. Y esto es cierto incluso en el campo donde más se ha avanzado en otras partes, porque constituye un primer estadio de estas investigaciones: los índices de vocabulario y las concordancias[10]. Nuestro atraso en este sentido es abismante y naturalmente se hace sentir, por ejemplo, en los trabajos lexicológicos diacrónicos o sincrónicos, en especial de épocas pretéritas. Esto significa que por el momento no podemos pensar en comparar, sino en presentar elementos que puedan ser comparados.

Esta es nuestra poco halagüeña situación actual. Para remediarla se precisa de un trabajo lingüístico-matemático, que puede ser realiza-

[10] En nuestro *El léxico de Lope de Rueda. Clasificaciones conceptual y estadística* (Bonn, Universidad de Bonn, 1968, 415 pp.), entregamos una muestra de "index verborum".

do por lingüistas, cuyos intereses y capacidades por lo general están muy alejados de las manipulaciones matemáticas, o bien por matemáticos, los que raramente demuestran comprensión por los tan poco exactos problemas lingüísticos. Como vemos, un hibridismo de este tipo es difícil de encontrar y de allí la escasez de este tipo de estudios y la necesidad de asesoramientos recíprocos.

En todo caso, nos parece ésta una dirección de la investigación lingüística cuya importancia está fuera de toda duda y que merece un poco más de atención que la que ha recibido hasta ahora.

Entregamos aquí una pequeña contribución en cuatro notas que describen numéricamente algunos aspectos del vocabulario[11] de Lope de Rueda: 1. Relaciones entre "formas-básicas", frecuencias y "formas-text ales"; 2. "formas-básicas" de las mayores frecuencias; 3. Relación entre el número de "formas-básicas" y la escala de frecuencias; 4. Dos posibles índices numéricos para caracterizar el léxico de Lope de Rueda.

1. Encontramos en la reducida obra de Lope de Rueda 4.385 "formas-básicas" (F-B) y 69.145 "formas-textuales" (F-T), lo que da un media de 15,76 F-T por cada F-B.

Aun cuando la inmensa mayoría de las F-B (el 90,2%) se agrupan en las frecuencias que van de 1 a 17 documentaciones, el número de F-T que les corresponden no alcanza ni siquiera a constituir el 18% del total. En general, tenemos el siguiente cuadro:

3.958	F-B con frecuenci asque van de	1 a	17	documentaciones:	11.999 F-T
349		18 a	121		14.088 "
26		122 a	186		3.926 "
22		187 a	356		5.763 "
2		357 a	982		11.612 "
10		1.126 a	4.181		21.757 "
4.385		1 a	4.181		69.145 "

Reduzcamos este cuadro a porcentajes:

F-B	*Frecuencias*		*F-T*
90,2 %	de las F-B con frecuenci asque van de	1 a 17	document. ti enen el 17,3%
7,9 %		18 a 121	20,3%
0,59%		122 a 186	5,6%
0,5 %		187 a 356	8,3%
0,45%		357 a 982	16,7%
0,22%		1.126 a 4.181	31,4%

y acumulativamente, partiendo de las F-B más frecuentes:

[11] Usamos "vocabulario" en el sentido fijado en la nota 5.

F-B		*F-T*	
10	más frecuentes	(0,22%) : 21.757	(31,4%)
30		(0,67%) : 33.369	(48,1%)
52		(1,17%) : 39.132	(56,4%)
78		(1,77%) : 43.058	(62 %)
427		(9,67%) : 57.146	(82,3%)

y el resto, esto es

3.958	F-B	(90,22%) : 11.999	(17,3%)

Estos cuadros nos indican a grosso modo que 1 de cada 3 F-T que aparecen en la obra de Lope de Rueda pertenece a alguna de las 10 F-B de mayor frecuencia y 1 de cada dos, al grupo de 30. Por el contrario, las 3.958 F-B cuyas frecuencias van de 1 a 17 tienen una probabilidad en extremo reducida: 1 por cada 5 F-T, con lo que naturalmente aumenta su cantidad de información.

2. ¿Cuáles son estas diez F-B que por sí solas constituyen un tercio de todas las F-T?

Como era de suponer, aparecen en este conjunto los elementos lexicales más descoloridos semánticamente: el artículo definido, preposiciones, conjunciones, verbos auxiliares, pronombres, algún adverbio, en todo caso ni sustantivos ni adjetivos:

1.	EL	4.181	F-T
2.	QUE	3.556[12]	
3.	DE	2.889	
4.	Y	2.176	
5.	YO	1.988	
6.	SER	1.650	
7.	A	1.472	
8.	NO	1.469	
9.	EN	1.250	
10.	HABER	1.126	

Con las 20 F-B siguientes completamos casi el 50% del total de las F-T. Dentro de este segundo grupo se destacan los verbos, algunos de uso a veces perifrástico o auxiliar: *ir, estar, tener, hacer* y otros como *decir, ver, dar, querer,* que ya nos proporcionan una primera pauta acerca de las preferencias semánticas y temáticas del autor; encontramos asimismo pronombres posesivos: *mío, suyo;* pronombres personales: *tú, él;* preposiciones, conjunciones, el artículo indefinido y el primer substantivo: *señor.*

[12]Este QUE es el conjunto de todos los *que* inacentuados.

11.	POR	982	F-T
12.	SEÑOR	964	
13.	TÚ	867	
14.	MÍO	822	
15.	DECIR	781	
16.	QUÉ	777	
17.	UN	694	
18.	ÉL	658	
19.	CON	630	
20.	HACER	551	
21.	TENER	544	
22.	SI	507	
23.	SE	506[13]	
24.	ESTAR	463	
25.	IR	448	
26.	SUYO	430	
27.	PUES	422	
28.	QUERER	401	
29.	DAR	397	
30.	VER	375	

En las 22 F-B de frecuencia inmediatamente inferior a las ya conocidas figuran más elementos "plenos": 4 verbos: *saber, venir, poder, dejar;* 4 substantivos: *dios, merced, hijo, casa;* los demostrativos *este, ese, aquel;* posesivos, conjunciones y preposiciones.

31.	COMO	357
32.	SABER	350
33.	ESTE	348
34.	VOSOTROS	325
35.	TUYO	323
36.	MÁS	317
37.	VUESTRO	314
38.	PARA	289
39.	ESE	288
40.	DIOS	288
41.	VENIR	278
42.	PODER	272
43.	MERCED	246

[13] Al igual que los *que* pertenecen estos *se* a distintas clases: impersonales, reflexivos, etc.

44.	YA	242
45.	HIJO	235
46.	TODO	225
47.	OTRO	225
48.	TAN	222
49.	AQUEL	203
50.	CASA	199
51.	DEJAR	194
52.	SÍ	191

Estas 52 F-B constituyen el 56,4% de todas las F-T del vocabulario de Lope de Rueda. Se advierte un claro predominio de las formas con menor cantidad de información: artículos, preposiciones, conjunciones, pronombres, verbos "vacíos" frente a los verbos "plenos" y a los substantivos. Hasta aquí no ha aparecido ningún adjetivo calificativo o de cualidad. Ordenando de otro modo estas F-B podremos tener una mejor visión de conjunto:

EL	DE	Y	YO	NO	MIO	ESTE	QUÉ	TODO	SEÑOR
	EN	PUES	TÚ	SÍ	SUYO	ESE		OTRO	DIOS
	POR	YA	ÉL	MÁS	TUYO	AQUEL			MERCED
	CO	COMO	VOSOTROS		VUESTRO				HIJO
	PARA	TAN							CASA
UN	A	SI							

SER[14] HABER DECIR HACER TENER ESTAR IR QUERER DAR VER SABER VENIR PODER DEJAR

3. Pasemos ahora a las F-B del otro extremo de la escala. No nos interesa aquí cada una de ellas consideradas en sí misma, sino agrupadas en función de su frecuencia. El problema es aquí: ¿Cuántas F-B tienen frecuencia 1, 2, 3, 4, . . . ? Al hacer una lista con estos dos elementos nos llama inmediatamente la atención una suerte de relación inversa que existe dentro de ciertos límites: a mayor número de F-B, frecuencia mínima y a medida que ésta aumenta decrece la concentración de F-B.

[14]Dejamos fuera de esta ordenacióna los *que* y a los *se*.

F-B	*Frecuencia*
1.876	1
714	2
362	3
230	4
152	5
115	6
91	7
90	8
68	9
56	10
46	11
34	12
32	13
31	14

Más allá de la frecuencia 14 empieza a vacilar esta regularidad:

19	15
23	16
19	17
19	18
15	19
15	20
13	21

que ya se pierde definitivamente a partir de esta zona de frecuencias, aunque se conserva la tendencia a la disminución del número de F-B:

9	22
11	23
5	24
13	25
7	26

Sabido es que a medida que aumenta la probabilidad de un elemento de un código disminuye su cantidad de información, pero al mismo tiempo crece su aplicabilidad al hacerse más amplio su contenido semántico ("Cosa" es una "argolla", un "automóvil", un "manómetro" o una "piedra", su frecuencia será por ende muchísimo mayor). Entran en juego dos tendencias, una que busca la precisión y la otra, el menor esfuerzo. Ambas se repelen, por ello es general la relación inversa entre la frecuencia y el número de F-B, aun cuando en particu-

lar pueda funcionar no rígidamente como sucede en este caso. No resulta fácil encontrar la expresión matemática de esta relación. En otra oportunidad nos referiremos a las soluciones propuestas.

4. Por otra parte se ha intentado repetidamente reducir las relaciones que se establecen en el vocabulario de una obra, de un autor, de una época a una constante numérica que permita identificar o caracterizar al corpus dado[15]. Así, por ejemplo, YULE propuso $K = \frac{S_2}{S_1^2}$ esto es, la suma de los cuadrados de las diferencias existentes entre la frecuencia efectiva y la frecuencia media (S_2) dividida por la suma de dichas diferencias (S_1) elevada al cuadrado[16]. HERDAN prefiere el índice $v\ m = \frac{\sigma}{\sqrt{NM}}$ en el que es la raíz cuadrada del cuadrado de la suma de las diferencias entre frecuencia efectiva y frecuencia media partida por el número de unidades consideradas o N. M es la media[17]. MANDELBROT utiliza $a = \frac{\log N}{\log V}$ donde N es el número de nuestras F-T y V, el de las F-B[18]. Por último, para GUIRAUD el mejor índice y el más prácticamente utilizable es $\frac{V}{\sqrt{N}}$ [19].

[15]Para lo que sigue véase el capítulo *"L'équation d'Estoup-Zipf et les caractères statistiques du vocabulaire"* en GUIRAUD, op. cit.

[16]Vid. YULE, G., *A statistical study of vocabulary*, Cambridge, 1944.

[17]Vid. HERDAN, G., *Language as chance and choice*, Groningen, 1956.

[18]Vid. GUIRAUD, op. cit., pp. 86 y ss.

[19]V y N representan los mismos valores que en *a*.

Siete meses después de terminada la primera parte de este artículo nos ha llegado *La relation Rang-Fréquence et la structure stilistique de la langue parlée* de R. MICHÉA (*BSLP*, LXXII, 1, 1968, pp. 9-13) que se refiere al mismo tema. Michéa se propone estudiar el valor de *fr* en la lengua escrita y en la lengua hablada y demostrar que fr = constante no es una ley. En *L'elaboration du français fondamental* (GOUGENHEIM, MICHÉA, RIVENC, SAUVAGEOT, Paris, Didier, 1964) y en el *Grunddeutsch* de J. ALAN PFEFFER (University of Pittsburgh, 1964 encuentra la misma estructura de ascenso y descenso que veíamos en Lope de Rueda (Cf. I. 3.1). El fenómeno lo explica por la tendencia de la lengua escrita a evitar las repeticiones, lo que no sucede en absoluto en la lengua hablada, con lo que en este caso se elevan las frecuencias iniciales, produciéndose de esta manera la configuración señalada (el carácter coloquial de la obra de Lope de Rueda parece reforzar esta hipótesis). En todo caso, para Michéa "En résumé, la formule fr = constante n'est pas vérifiée dans la langue parlée".

Tenemos serias dudas acerca de la utilidad general de tales índices, ya que, al parecer, sólo son válidos dentro de determinadas zonas de frecuencias y sólo en el caso de textos "normales". De todos modos, y pensando que probablemente puedan servir de base para futuras comparaciones, hemos calculado, según las fórmulas de Mandelbrot y de Guiraud, dos tipos de índices que caracterizarían el vocabulario de Lope de Rueda:

$$a = \frac{\log N}{\log V} \qquad \frac{\log 69.145}{\log 4.385} \quad 1{,}33$$

$$\frac{V}{N} \quad \frac{4.385}{69.145} \quad 16{,}7$$

Instituto Pedagógico de Valparaíso
Universidad de Chile

LEOPOLDO SÁEZ G.