Estadística Lingüística.

Dr. Adolfo Sarabia Santander

I

En su conferencia inaugural de curso ante la Universidad de Oxford, el 10 de noviembre de 1969, T. W. Reid afirmaba: "En el estudio serio del lenguaje como totalidad ha tenido lugar en los últimos treinta años una innovación fundamental, con algunos aspectos realmente revolucionarios, debida al impulso de ciertos investigadores, inicialmente formados como filólogos, cuyas reflexiones, en su mayor parte independientes sobre la naturaleza del lenguaje, han convergido hacia la formación de una nueva disciplina lingüística. Los más influyentes entre ellos (aunque todos tuvieron sus precursores) fueron sin duda Ferdinand de Saussure con su dicotomía entre lingüística sincrónica (o descriptiva) y lingüística diacrónica (o histórica), y su concepción del lenguaje como un sistema sincrónico, N. S. Trubetzkoi con su teoría sobre el fonema en términos de oposiciones funcionales y aspectos sonoros distintivos; y (especialmente en América) L. Bloomfield con su programa para el análisis descriptivo de las lenguas basado en la psicología conductista. De las enseñanzas de estos investigadores han surgido las numerosas y variadas escuelas de pensamiento lingüístico que son generalmente descritas con el término de estructuralistas" (1).

---

(1) T. W. REID, "Historical Philology and Linguistic Science". Oxford, 1960, pp. 6 y 7.

Ciertamente que podemos describir como "numerosas y variadas" a las escuelas de pensamiento lingüístico de nuestros días, producto de un creciente interés por las materias lingüísticas que, rebasando el estricto campo de los investigadores, tiene amplios ecos en diversos sectores de la sociedad humana.

Una de las facetas más modernas y fascinantes de la lingüística es la constituida por la que fue en ocasiones llamada *Lingüística Matemática* y otras *Estadística Lingüística.*

Antes de nada, concretemos. Es cierto que la expresión "lingüística matemática" puede tener dos sentidos perfectamente diferenciados: de un lado, el de "Lingüística cuantitativa", orientada básicamente a los aspectos numéricos del lenguaje incluidos en la "estadística lingüística" (la estadística lingüística sería, pues, parte de la lingüística matemática), y de otro lado, "lingüística matemática" puede tener el sentido de "lingüística algebraica o algorítmica" y viene a significar que en el curso de la investigación lingüística se utilizan símbolos que pueden aparecer reunidos por medio de operaciones y fórmulas "matemáticas", es decir, que este segundo sentido es completamente distinto del cuantitativo o estadístico: las gramáticas transformativas constituyen un ejemplo de lingüística matemática o algebraica en este sentido.

Quede, pues, claro que al referirnos en adelante a "lingüística matemática" lo hacemos en el sentido no cuantitativo, y al referirnos a "estadística lingüística" nos referimos a los aspectos cuantitativos y numéricos del lenguaje, es decir, no consideramos a la estadística lingüística como parte de un todo llamado estadística matemática, sino como un todo con unos fines y unos métodos que admiten la posibilidad de ser definidos con independencia.

Pese a que nos hemos referidos a la estadística lingüís-

tica como ciencia nueva, no olvidemos que no hay nada nuevo bajo el sol y que los primeros intentos de captar numéricamente la realidad y las relaciones íntimas entre fenómenos de todo tipo no son de ayer.

Entre los babilonios en los comienzos del segundo milenio antes de Cristo y entre los griegos del siglo IV a. de C. existía la fuerte convicción de que las formas podían ser susceptibles de expresión numérica y, en algunos casos, como la armonía musical o la *armonía estética,* se afirmaba que su misma esencia residía en aquella típica teoría de los números que constituyó la base de la escuela pitagórica que floreció principalmente en las colonias griegas del sur de Italia entre los años 600 y 300 a. de C. (2). Ahora bien, aquellos primeros intentos de los griegos para descubrir la esencia de la forma fueron planteados en términos geométricos y no algebraicos y aunque hoy en día podemos ver que la parte de la geometría conocida como topología hubiera podido ayudarles en sus investigaciones, la realidad es que en el estado de sus conocimientos de geometría ésta constituía un callejón sin salida en muchos de los problemas planteados. El siglo IV a. de C. había de proporcionar un planteamiento radicalmente distinto, basado en el silogismo aristotélico que barrería casi totalmente la intensa preocupación de la escuela pitagórica por aquello que yace bajo las formas externas. Esta es la época también en que caen en el olvido las ideas sobre el heliocentrismo del sistema solar, de la esfericidad de la Tierra y de la naturaleza atómica de la materia que

---

(2) Vide: STAPLETON, H. E., "The Hand with its fingers, as the primitive basis of geometry, arithmetic and algebra", Actes du 8ème. Congrés International Hist. Scr., París, 1956, p. 1.103.

STAPLETON, H. E., "Ancient and modern aspecs of Pythagoreanism" *Osiris,* n.º 30, 1958, pp. 12-13.

LEVY, I., "Recherches sur la légende de Pythagore", París, 1926.

VAN DER WAERDEN, "Science Awakening", New York, 1963.

GUTHRIE, EK. C., "A history of Greek Philosophy", Ginebra, 1958.

propugnaba el contemporáneo de Pitágoras, Demócrito, desde su ciudad de Abdera.

Ahora bien, como las ideas básicas de la teoría pitagórica se ayoyaban en el convencimiento casi mágico de la íntima unidad e interrelación de todo lo creado, los intentos de explicación numérica del universo pervivieron, un poco a la manera de francotiradores de la cultura, y dieron lugar a algunas curiosas obras como las de Enel, Paracelso, Von Rosenroth y otros (3). Sobre el principio subyacente a la mayoría de estos autores, recordemos que "l'unité serait la loi dominante de l'Univers, unité de la force, unité de la matière, et l'ensemble de ces unités serait Dieu, l'infiniment unique et l'uniquement infini. L'essence de notre univers serait la Loi, ensemble des règles qui régissent les actions et les réactions des modalités vitales... La vie étant unique et ses seules modalités variables, *tout n'est donc* que rapports et tout n'est qu'harmonie" (4).

Una interpretación similar de las íntimas relaciones existentes entre todo lo creado y de la armonía universal puede verse en Hume, sobre todo en sus "Dialogues Concerning Natural Religion" especialmente en su Parte VI, y en Diderot, "Pensées sur l'interprétation de la Nature". Y sin restar nada a la genialidad de Zipf, con su principio del mínimo esfuerzo, podemos recordar que fue Maupertuis quien estableció en el siglo XVIII el principio de *ac-*

---

(3) PARACELSI, A. P., "Opera Chemica et Philosophica", Ginebra, 1958.

ENEL, "Essai d'astrologie cabbalistique", Toulon (sin fecha).

ANONIMO (An Oxonian), "Ellucidations of the marvellous", Londres, 1835.

PAGLIARO, A., "Il segno vivente", Nápoles, 1952.

AYER, A. J., "Philosophy and Lenguage", Oxford, 1960.

(4) HAVEN, M., "Les sept libres de L'Archidoxe de Paracelse", París, 1909, pág. 5.

ción mínima, mereciendo las alabanzas de Hume en carta al abate Le Blanc (5).

Y puesto que ha aparecido el nombre de Zipf tal vez sea buen momento para regresar a nuestra estadística lingüística pero sin olvidar ya esta nueva perspectiva de la unidad inmanente de todas las cosas y de las íntimas relaciones que las unen, relaciones que suponemos que podemos poner más en claro por medio de manipulaciones numéricas.

II

Tradicionalmente el recuento de los elementos lingüísticos ha tenido un lugar más bien secundario y marginal dentro de la lingüística y ha estado dirigido principalmente a ciertos aspectos aplicados, tales como el estudio de frecuencia de letras o palabras con la finalidad de crear o mejorar sistemas estenográficos o mecanográficos (6). Las investigaciones sobre el vocabulario que han producido frutos más inmediatos y prometedores desde el punto de vista lingüístico son aquellas que se han basado en una aplicación de las teorías de la estadística al mundo del lenguaje.

Los fenómenos lingüísticos presentan una serie de características que los convierten en elementos ideales para el tratamiento estadístico. Si es cierto que al actualizar en habla nuestra lengua lo hacemos de acuerdo con toda una serie de necesidades personales, temáticas o estilísticas, no es menos cierto también que al reunir en un corpus suficientemente amplio los resultados de esa actuali-

(5) "Letters of David Hume", Oxford, 1932, t. I, p. 227.
(6) O. W. MELIN, "Stenografiens Historia", Estocolmo, 1929.

zación y estudiarlos a través de métodos estadísticos, se nos presentan una serie de resultados sorprendentes que en algunos casos podemos comenzar a considerar como "leyes" internas del lenguaje y que podemos esperar que nos aclaren algún día aspectos de ese mismo lenguaje que hoy nos son desconocidos.

Tropezamos, empero, con un problema de métodos que reúne no pocas dificultades: el tratamiento estadístico del lenguaje supone unos conocimientos de matemáticas que generalmente el lingüista no posee. No es de extrañar, pues, que dos de los principales autores sobre el tema sean un lingüista interesado por las matemáticas, P. Guiraud (7), y un matemático interesado por la lingüística, G. Herdan (8).

Ahora bien, el verdadero precursor (tal vez debiéramos decir fundador) de la estadística lingüística fue G. K. Zipf (9). Aunque ya la Escuela de Praga propugnaba en el decenio anterior a la segunda guerra mundial la importancia de la consideración estadística de los fenómenos lingüísticos (en especial en la noción de rendimiento funcional), fue Zipf quien, a partir de 1929, identificó y planteó varias leyes y tendencias sumamente interesantes y que posteriormente serían corregidas y formuladas de manera más precisa (en especial por Herdan).

He aquí algunos de los hallazgos de Zipf:

I) Si colocamos todas las palabras de un texto por

---

(7) GUIRAUD, P., "Les caractères statistiques du vocabulaire", París, 1954. Id. "Problèmes et méthodes de la statistique linguistique", París, 1960. Id "Statistique et analyse linguistique", París, 1966.

(8) HERDAN, G., "Language as choice and chance", Groningen, 1956. Id. "Quantitative linguistics", London, 1965.

(9) ZIPF, G. K., "Relative frequency as a determinant of phonetic change", Harvard Studies in Classical Philology, XL, 1929, pp. 1-95. Id. "Human behavior and the principle of least effort", Cambridge (Mass.), 1949.

orden de frecuencia (frecuencia = número de veces que aparece una palabra en dicho texto) a partir de la frecuencia mayor, y llamamos *rango* al número de orden que a cada palabra le corresponde en nuestra lista, es evidente que la relación existente entre el rango y la frecuencia es una relación inversa (a mayor rango corresponde menor frecuencia, y al rango menor, la primera palabra de la lista, corresponde la mayor frecuencia). Ahora bien, lo que no es tan evidente es que dicha relación no es solamente inversa, sino inversamente *proporcional,* es decir, que el producto del rango por la frecuencia es aproximadamente constante. Así, en el capítulo XI de *Kim* de Rudyard Kipling encontramos las cifras siguientes:

rango r)	vocablo	frecuencia (f)	f. r.
23	all	51	1.173
26	on	45	1.170
32	they	37	1.184
39	man	30	1.170
50	what	22	1.100

Este primer planteamiento de la ley de Zipf fue corregido por Mandelbrot y posteriormene por Herdan, quien ha presentado una nueva fórmula de la ley que corrige las aparentes irregularidades que se observan en Zipf en las frecuencias más altas y más bajas.

La simetría de la relación entre frecuencia y rango, parece poder explicarse como expresión de una relación de equilibrio entre dos fuerzas contrarias determinadas por dos claras tendencias existentes en el uso de la lengua: por una parte, el locutor tiende a facilitar su propia elocución (a disminuir su esfuerzo) por medio del empleo de palabras de amplio espectro semántico y de la sustitución de palabras concretas por otras más amplias y vagas, y por otra parte el interlocutor desea disminuir su

esfuerzo de comprensión exigiendo del locutor un máximo de palabras precisas y un máximo de matización semántica. Entre los dos extremos: "la palabra comodín" y "una palabra para cada concepto" se establece un equilibrio que es el que pretende reflejar la ley de Zipf. En opinión de Malmberg (10) "le même principe s'aplique du reste également au plan de l'expression, puisqu'on voit dans la chaîne sonore effectivement prononcée un resultat de deux forces contraires: la tendence à effectuer le minimum d'effort et le besoin de se faire comprende. Ici les deux extrêmes sont d'une part l'onde sonore indifferenciée dont la valeur d'information est presque équivalente à zéro, d'autre part l'onde sonore fortement cisaillée avec tous ses élements distinctifs nettement proposés. La langue concrète est toujours un compromis entre ces deux tendences, puisque normalment il s'opère entre les differences de sons de la langue autant d'assimilations ou suppressions de differences sonores qu'il est posible sans compromettre l'intelligibilité dans une situation donnée".

II) Otra de las interesantes relaciones existentes entre las palabras de un texto apuntadas por Zipf consiste en el hecho de que el número de palabras que tienen la misma frecuencia multiplicado por el cuadrado de la frecuencia tiende a ser una constante. Si frecuencia = f, y al número de palabras de esa frecuencia lo llamamos *a*, tendremos que:

$$a \cdot f^2 = C \ (= \text{constante})$$

Guiraud (11) presenta los siguientes resultados tomados del conjunto de vocabulario del "Ulises" de James Joyce:

---

(10) MALMBERG, B., "Les nouvelles tendences de la linguistique", Paris, 1966, p. 287.

(11) "Les caractères statistiques du vocabulaire", cap. IX.

a			f		$a.f^2$
900	palabras	aparecen	5	veces	22.400
770	"	"	6	"	22.700
480	"	"	7	"	23.500
370	"	"	8	"	23.600
300	"	"	9	"	24.300

El mismo Guiraud ha formulado otra ley interesante con respecto a la relación entre la frecuencia y la longitud de las palabras. Según él, las palabras largas son las menos frecuentes y las palabras cortas las más frecuentes, lo que lleva a plantear la ecuación:

$$f \cdot e = C \text{ (constante)}$$

donde *f* es la frecuencia y *e* la energía exigida por la señal (en el caso de la lengua, el número de fonemas en cada palabra).

La observación de éstas y otras así llamadas "leyes del lenguaje" han provocado toda una serie de estudios que constituyen cada uno de por sí otras tantas finalidades de la estadística lingüística. Para ofrecer una idea de estas posibilidades, presentamos a continuación una breve muestra de los principales fines de la estadística lingüística.

III

En el mismo ámbito de la finalidad general de la estadística lingüística, que pudiera ser definida como un intento de la aplicación de los métodos estadísticos a los problemas lingüísticos con el fin de conseguir unos resul-

tados lo más objetivos, numéricos y comparables posibles, podemos distinguir diversas provincias que corresponden a los intereess concretos de la investigación emprendida por cada lingüista. Creo que como muestra de la infinita gama de posibilidades, podemos ofrecer tres campos de gran interés en sí mismos, y en los que se han obtenido resultados que pueden ser calificados de positivos, a saber:

a) FINES PEDAGOGICOS

Dentro del mismo orden de ideas con que concluíamos el párrafo anterior, se pueden citar otras tendencias no menos interesantes. Así por ejemplo es posible preparar, a partir de una muestra importante de una lengua, una lista de palabras ordenadas en orden descendente de frecuencias y observar unos resultados aproximados a los siguientes: suponiendo una muestra de 50.000 palabras:

las	15	primeras	palabras	cubren el 25 %	del texto
"	100	"	"	" 60 %	"
"	1.000	"	"	" 85 %	"
"	4.000	"	"	" 97,5 %	"

de manera que para las 46.000 palabras restantes no queda más que el 2,5 % del texto.

Las implicaciones pedagógicas, tanto para el aprendizaje de la lengua materna, como de las lenguas extranjeras, son evidentes, y estas mismas implicaciones llevaron a Thorndike y a Lorge a la elaboración de su "The Teacher's Word Book of 30.000 words" (1.ª edición 1931, última 1968, con considerables diferencias). El número de palabras utilizadas para obtener las frecuencias de los 30.000 voca-

blos ha superado a los 15 millones (12) y el trabajo todavía continúa en la Escuela Univesitaria de Formación del Profesorado de la Universidad de Columbia. Pese a que la finalidad fundamental de la lista fue el proporcionar con toda seguridad los vocabularios básicos que debieran contener los libros de enseñanza en sus diversos grados, es evidente que este memorable trabajo está lleno de posibilidades para el lingüista. Otros estudios estadísticos han sido elaborados para diversas lenguas con miras fundamensalmente pedagógicas, como el *French Word Book* de G. E. Vanderbeke (1929-1935) para el francés; para el danés, por A. Nosgard en 1934-1937; para el noruego por E. Haugen en 1942. E. C. Hills publicó en 1931 *Spanish Words of High Frequency,* y en la actualidad está a punto de ser publicado en París el *Diccionario del Español Fundamental,* sin olvidar el *Vocabulario Usual* de García Hoz (13).

Otros trabajos no menos ambiciosos han seguido y siguen siendo realizados incesantemente. En 1967, se inició en la Universidad de Essex, y todavía continúa su desarrollo, un programa para crear nuevos métodos de aprendizaje de lenguas extranjeras, basado por una parte en presupuestos psicológicos del aprendizaje y por otra en las aportaciones de la estadística lingüística, y del que es principal responsable M. H. Alford (14). Parte el programa del principio de que el aprendizaje es un proceso que

---

(12) Labor de gigantes pese a haber sido realizada en equipo y con generosas ayudas, que nos hace admirar aún más a F. W. Kading, quien, a partir de 1897, procesó solo y con paciencia investigadora verdaderamente germánica once millones de palabras para establecer el orden de frecuencias de la lengua alemana.

(13) GARCIA HOZ, Víctor, *Vocabulario usual, común y fundamental (determinación y análisis de sus factores),* Madrid, 1953.

(14) Del Language Center de la Universidad de Essex; en la actualidad en el Literary Linguistic Computing Center de la Universidad de Cambridge.

se prolonga durante toda la vida como respuesta a determinados estímulos; el proceso se compone de un aprendizaje inicial y de una actualización del aprendizaje anterior. Sin los efectos de la actualización o repaso, los efectos del olvido superarán a los del aprendizaje.

Se hace necesario reunir una gran cantidad de información sobre la lengua estudiada con el fin de constituir grupos significativos y comprensivos que reflejen los diversos niveles de dicha lengua (lengua hablada, diversos géneros literarios, literatura científica, etc.). Con el fin de establecer la capacidad del nativo y las finalidades del estudiante, los detalles textuales tienen que ser cuantificados de acuerdo con los límites de la experiencia de cada persona. Los datos, pues, deben ser reorganizados de diversas formas, de manera que se facilite el aprendizaje del estudiante; estas posibles formas son tan complejas y variadas con relación a cada posible estudiante, su nivel de aprendizaje y sus intereses futuros, que sólo un computador es capaz de realizarlas, proporcionando a cada estudiante una enseñanza totalmente personalizada. Las manipulaciones del computador pueden ser utilizadas, por otra parte para una gran variedad de posibilidades, tales como:

a) Problemas de aprendizaje (palabras de altas y bajas frecuencias).

b) Organización didáctica (tamaño del vocabulario).

c) Estudios sociales (presencia, ausencia o gran frecuencia de determinadas palabras).

d) Identificación de autores (frecuencias idiosincráticas).

Por otra parte, el computador proporciona las siguientes ventajas para la elaboración de los cursos de aprendizaje:

a) *Identificación.*

Las palabras del texto pueden ser identificadas como:

1. Formas gráficas, tal y como aparecen en el texto.
2. Igualadas con las formas del diccionario.
3. Formas iguales con significados diferentes.
4. Combinaciones de morfemas.

b) *Recuento.*

Este puede ser llevado a cabo con cualquier forma de identificación y en cualesquiera circunstancias; a partir de él pueden realizarse todos los cálculos necesarios o convenientes sobre porcentajes, promedios y distribuciones.

c) *Comparación.*

Los diversos textos pueden ser comparados, a cualquier nivel de identificación, para establecer los elementos comunes o diferentes en cada uno.

d) *Amplificación.*

Basados en la identificación previa, se pueden obtener otros diversos tipos de información, tales como:

1. La forma de diccionario correspondiente a la palabra identificada.
2. La segmentación morfémica de la palabra identificada y de su forma en el diccionario.
3. La frecuencia de ocurrencia de la forma identificada, en diversos textos y de acuerdo con diversos principios predeterminados.
4. El porcentaje de cualquier palabra en un texto.

e) *Mezcla.*

Los recuentos, listas léxicas y demás datos pueden ser reunidos en la memoria del computador para su utilización en posteriores trabajos de más extensión.

f) *Ordenación.*

Las palabras pueden ser ordenadas por el computador de diversas maneras, como:

1. Alfabéticamente de izquierda a derecha.
2. Alfabéticamente de derecha a izquierda.
3. Alfabéticamente a partir de un morfema dado.
4. Por orden de frecuencias ascendentes o descendentes.
5. A partir de un número especificado de ocurrencias.

Utilizando estas posibilidades el computador nos permitirá programar cursos de aprendizaje adecuados a las necesidades de un grupo reducido de personas, o incluso de un solo individuo. Por ejemplo, si un grupo de estudiantes de Física desea aprender el ruso por motivos profesionales, no debemos olvidar el hecho general de que un número relativamente reducido de palabras de alta frecuencia es responsable de la formación de la mayor parte del texto. El aprendizaje habrá de concentrarse, pues, en esas palabras de alta frecuencia; en el caso de la lengua rusa aplicada a la física los porcentajes son los siguientes:

Palabras diferentes	Texto que cubren
50	34 %
100	45 %
1.000	86 %
2.750	96 %

Es decir, que una persona que pueda reconocer las 1.000 palabras más frecuentes, sólo encontrará una o dos palabras nuevas en cada frase y las probabilidades de poder deducir su significado por el contexto son muy elevadas. Con el conocimiento de unas 3.000 palabras solamente presentará problemas una palabra por párrafo, y la posibilidad de deducción de su significado por el contexto es mucho más alta; a partir de este momento el aprendizaje se basará fundamentalmente en la ampliación de lecturas sobre el mismo tema por parte del propio estudiante. No podrá leer ciertamente a Tolstoi, pero sí podrá leer cualquier texto de Física, que es lo que le interesaba al comienzo de su instrucción. Huelga decir lo interesante que sería que nuestros Institutos de Ciencias de la Educación basaran la programación del aprendizaje de lenguas extranjeras basándose en esta metodología.

b) ATRIBUCION DE AUTORES

Han sido muy numerosos los trabajos que en los últimos años han utilizado la estadística lingüística para resolver problemas de cronología e identificación de autores, y numerosos también los diversos métodos elaborados para resolver estos problemas específicos. Una extensa relación de los trabajos de Ellegard, Morton, Wallace y otros investigadores puede encontrarse en la sección "Problems of Chronology and Disputed Authorship" del libro de R. W. Bailey "An Annotated Bibliography of Statistical Liguistics", Ann Arbor, 1968.

Como ejemplo posterior a dicho libro vale la pena citar el estudio de A. Mackinnon y R. Webster, de la Universidad McGill de Montreal, sobre Kierkegaard. Como es sabido, las obras de Kierkegaard pueden ser distribuidas en dos grandes grupos: las obras "religiosas", firmadas todas

con su propio nombre, y las obras "estéticas", firmadas con pseudónimos. En principio, la finalidad general de todas sus obras era la de llevar al lector a una existencia auténtica, o, en otras palabras, a la madurez espiritual y psicológica. En este aspecto, el papel de sus obras religiosas es obvio, pero el de sus obras estéticas requiere mayor matización. Estas últimas obras reflejan su convicción de que sus contemporáneos vivían en medio de unas categorías estéticas, o, como mucho, estético-éticas y que, por consiguiente, su labor como autor religioso consistía en encontrar a sus contemporáneos en su propio terreno. Desde el punto de vista de Kierkegaard las ilusiones de sus contemporáneos solamente podían ser eliminadas a través de una serie de obras que, tomadas una por una, presentaran diversos puntos de vista, y, tomadas en conjunto, ofrecieran una auténtica progresión dialéctica desde un concepto del Universo a otro. De aquí el uso de sus variados pseudónimos que, como él dice, son "creaciones poéticas" o, más bien, personalidades literarias cuyos estilos distintos de vida quedan perfectamente expresados en sus diversas obras. Por ejemplo, A, el pseudónimo del volumen primero de *"Una cosa u otra"* expresa perfectamente el punto de vista estético, no sólo en lo que dice, sino de la manera como lo dice, mientras B, el pseudónimo del volumen segundo nos ofrece la encarnación perfecta de la actitud ética. Cada uno tiene una consistencia perfectamente propia, interna, lo cual resulta también cierto de cada uno de sus otros pseudónimos. Cada uno de ellos refleja una personalidad idealmente consistente con sus puntos de vista distintos y propios, incluso con evidentes contradicciones entre los diversos pseudónimos. Este planteamiento puede llevar a dos claros problemas: uno, tomar todas las ideas expuestas en los pseudónimos como ideas de Kierkegaard, hecho contra el que él mismo llegó a prevenir, y otro, tomar al pie de la letra las palabras de Kierkegaard "en los pseudónimos no hay una sola palabra

que sea mía", y rechazar de lleno los pseudónimos con lo que daríamos de mano a algunos de los escritos más interesantes de Kierkegaard. El problema que se han planteado, pues, McKinnon y Webster es el de conseguir un método que les permita establecer las relaciones entre Kierkegaard y sus propios pseudónimos. A través de una labor de estadística matemática han podido establecer: 1.º que, tanto individual como colectivamente, las selecciones tomadas de los escritos firmados por Kierkegaard son significativamente diferentes de los escritos firmados con pseudónimo, y 2.º una relación jerarquizada de los pseudónimos en comparación con el Kierkegaard reconocido. El índice de vocabulario, por ejemplo, en el promedio de las obras pseudónimas es de 0,845, mientras que en las firmadas por Kierkegaard es del 0,819, es decir, que los imaginarios autores pseudónimos tienen un vocabulario mucho más rico que Kierkegaard cuando éste escribe como tal Kierkegaard.

Los autores propugnan que su trabajo, utilizado para establecer las diferencias y jerarquías entre imaginarios "autores distintos", pese a conocer con certeza que todos provienen de la misma mano, puede ser utilizado con ventaja para resolver otros diversos problemas que pueden presentarse dentro de la problemática general de la identificación de autor.

Otro ejemplo interesante de intento de identificación de autor lo constituye el realizado por P. Köster, de la Universidad de Victoria, Canadá, y que describimos brevemente a continuación. El trabajo surgió de un largo y sostenido interés en Swift y sus asociados en la maquinaria propagandística Tory durante el reinado de la reina Ana y más inmediatamente del deseo de averiguar el autor o autores de uno de los mejores productos de dicho grupo político, a saber, *The Story of the St. Albans Ghost.* Este trabajo resulta doblemente interesante, primero, por la ha-

bilidad, cuidado y perfección con que está realizado, contando con los medios y la ayuda del Centro de Computación de la Universidad de Victoria, y, segundo, porque a pesar de todo, los resultados han sido parcialmente negativos, es decir, a pesar de que se ha llegado a resultados conclusivos en cuanto a determinar que la obra en cuestión *no pertenece* a determinados autores, no se ha podido llegar a una conclusión definitiva en cuanto a quién fue el autor de la misma. Esto viene a confirmar que, pese a que la Estadística Lingüística ha obtenido algunos resultados excelentes, no constituye todavía en algunos aspectos un instrumento infalible y que siguen siendo necesarios la investigación, el intercambio de ideas y los esfuerzos consiguientes a toda rama de una ciencia que en realidad aún se halla en sus principios.

Wöster se basa en la metodología de su trabajo en la obra ya clásica de Milic *A Quantitative Approach to the Style of Jonathan Swift.* Milic parte del postulado de que "la consistencia sintáctica es una característica de toda la obra de Swift, independientemente de la materia que trate". Utilizando un esquema gramatical basado en las "word-classes" de Fries, asigna un número de dos dígitos a cada palabra de una muestra, de acuerdo con su función gramatical; así pues, codifica los sustantivos como 01, los verbos como 02, los verbos auxiliares como 21, y así sucesivamente hasta un total de veinticuatro clases gramaticales. Marcando estas palabras codificadas en las primeras setenta y dos columnas de tarjetas IBM, reúne aproximadamente cien tarjetas para cada muestra. Después de realizar diferentes pruebas con diez muestras de las obras de Swift, y con ocho muestras de control, dos de Addison, dos de Johnson, dos de Gibbon y dos de Macaulay, Milic llega a la conclusión de que tres de las pruebas constituyen los discriminantes con más fiabilidad y en ellos basa lo que denomina su "Swift Profile". Los

tres discriminantes del perfil están constituidos por elevadas frecuencias en el uso de *Verbals* (VB), *Introductory Connectives* (IC) y *Different three-word Patterns* (D). En otros aspectos, Swift no se diferencia significativamente de los cuatro autores tomados como control de la prueba pero en los tres "discriminantes del perfil" obtiene unas puntuaciones significativamente más elevadas a través de diversos géneros literarios y a lo largo de un elevado número de años. Como resultado de su investigación, Milic concluye que, "si se deseara probar la atribución de una obra determinada a Swift, el problema podría quedar rápidamente resuelto haciendo referencia a los puntos del perfil".

Tomando, pues, el libro de Milic como base metodológica para su propio trabajo, Köster emprendió la tarea de establecer la paternidad de *The Story of the St. Albans Ghost*. Tras de realizar todas las laboriosas operaciones pertinentes de comparación con los tres perfiles de Swift establecidos por Milic, el resultado fue que solamente el perfil D se aproximaba a la cifra característica de Swift, por consiguiente, al fallar en dos de los tres perfiles del estilo de Swift, es muy probable que *The Story of the St. Albans Ghost* no fuera escrita por él. Los posibles autores de *The Story of the St. Albans Ghost* que tradicionalmente se han venido apuntando como probables han sido Swift (15), Swift en colaboración con Arbuthnot (16) y, finalmente, William Wagstaffe (17).

---

(15) Propuesto por T. ROSCOE en sus *The Warks of Jonathan Swift,* Londres, 1850, t. I, p. 529.

(16) Walter Scott sugirió esta colaboración "judging from the style" en su edición de *The Works of Swift,* 2.ª edición, Londres, 1883, t. V, p. 414.

(17) Mantenido por V. A. DEARING en su artículo *"Jonathan Swift or William Wagstaffe?",* publicado en el *Harvard Library Bulletin,* t. VIII, 1953, pp. 121-30; sin olvidar al anónimo compilador de *The Miscellaneous Works of Dr. Wagstaffe,* Londres, 1726, que incluyó entre ellas la *Story,* pero sin discutir ni justificar su inclusión.

Una vez eliminada la posibilidad de que Swift hubiera escrito por sí solo *The Story of the St. Albans Ghost* restaba probar la posibilidad de los otros dos casos. Resulta difícil resumir en breves líneas un proceso que resultó extremadamente laborioso. Se establecieron los perfiles pertinentes de las siguientes obras: de Arbuthnot "Art of Political Lying" y "John Bull" (ambas escritas en 1712, el mismo año que *The Story of the St. Albans Ghost)* y "A sermon Preached... at the Mercat Cross of Edinbugh" (escrita en 1706, cuatro años antes de que Arbuthnot conociera a Swift); de Wagstaffe: *"A letter to Dr. Freind, showing the Danger"* (1722), *"A Comment upon the History of Tom Thumb"* (1711) y *"The State and Conditions of our Taxes"* (1714). Aparte de estos dos probables autores, se tomó como control a una autora del mismo grupo Tory, que, con toda seguridad, no está relacionada con *"The Story of the St. Albans Ghost"*, Mrs. Manley (18). Las *Memoirs* de Mrs. Manley se parecen tanto a *The Story of the Albant Ghost* como a *John Bull* en el aspecto de que constituyen una narración con un segundo nivel de significado político y propagandístico.

Abreviando, las conclusiones de comparar los resultados de los diversos perfiles de todas estas obras, aunque permiten llegar a ciertas afirmaciones, dejan en el aire el problema de la paternidad de *The Story of the St. Albans Ghost.* Arbuthnot queda eliminado como posible autor, pero, por otra parte, sus perfiles quedan prácticamente confundidos con los de Swift (es decir, que los perfiles de Swift no son tan definitivos como creía Milic), lo que

(18) Mrs. Manley participó activamente en las campañas propagandísticas (y difamatorias) de los Tories, siendo primer ministro Harley. Sus *Memoirs of Europe,* fueron escritas durante el invierno de 1709-1710, y publicadas en mayo de 1710, época en la que Swift se encontraba en Irlanda, lo que elimina prácticamente la posibilidad de que actuara de corrector de estilo, como hizo con otras muchas obras de ficción política Tory.

viene a dar la razón a un crítico literario tradicional que, en 1912, afirmaba: "...Arbuthnot tanto en el ritmo de su prosa como en otras características es prácticamente inseparable de Swift" (19).

Los resultados del establecimiento de los perfiles de Wagstaffe, también o eliminan como autor de *The Story of the St. Albans Ghost,* aunque sí mantienen aproximaciones grandes con los estilos de Swift y Arbuthnot. Y algo similar ocurre con la autora tomada como control: por muy diferenciada que pueda parecer Mrs. Manley cuando leemos sus escandalosas historias sobre los Whigs, utiliza unos perfiles gramaticales muy parecidos a sus colegas Tories más conocidos y está tan apartada como ellos de los perfiles de *The Story of the St. Albans Ghost.*

¿Qué vienen a mostrar, pues, estos resultados? En primer lugar, el autor de *The Story of the St. Albans Ghost* sigue permaneciendo en el anónimo, y, en segundo lugar, aunque Köster parece haber encontrado ciertas constantes en el estilo Tory de la época de la reina Ana, a la vez ha eliminado la seguridad con que Milic presentaba los tres perfiles discriminadores del estilo Swift. La obvia consecuencia es que para casos de autores de la misma época, con temas similares, géneros similares, y frecuencias y disponibilidad de léxico equivalentes, se hace preciso continuar investigando nuevos métodos estadísticos que permitan apreciar diferencias de matiz en el estilo que, por el momento y pese al optimismo de Milic, la práctica parece demostrar que aún no se han conseguido.

Blackith publicó en 1963, un interesante experimento sobre el verso elegíaco latino. Partía del principio de que gran parte de la poesía en lengua latina escrita en la Edad Media y en épocas más recientes consiste en intentos de

---

(19) G. SAINTSBURY, *History of English Prose Rhythm,* Londres, 1912, p. 252.

escribir elegías en un estilo lo más aproximado posible al de la época de Augusto. El trabajo hubo de consistir en principio en el establecimiento por medio del análisis de componentes básicos, de las principales características del verso elegíaco en la época de Augusto: estos componentes resultaron ser:

1. Frecuencia de elisión (sílabas escritas pero no pronunciadas por razones de eufonía o por exigencias mismas del metro).
2. Número medio de sílabas por palabra.
3. Varianza (20) de la distribución de sílabas por palabra.
4. La entropía (21) de la distribución de palabras de diferente número de sílabas.

El tercer vector demostró ser el que mejor permite al analista determinar la fecha de un poema, independientemente de todos los esfuerzos de un autor para imitar el estilo (22).

Blackith ha publicado también un texto dentro de la

---

(20) *Varianza:* Término estadístico normalmente representado por *Var* o σ, definido como la media de los cuadrados de las desviaciones con relación a la media del conjunto. La varianza es un parámetro de dispersión al que a veces se alude con el nombre de *fluctuación.*

(21) La definición de *entropía* puede variar considerablemente según los contextos; para nuestro caso valdrá la siguiente. En tanto los componentes de un sistema se hallen separados en acción o posición, podemos hablar de un "estado ordenado" del sistema; en tanto en cuanto se hallen, mezclados, barajados, o confusos, podemos hablar de la "entropía" del sistema. Entropía, es, pues, la medida del grado de mezcla o confusión de los componentes Es considerada como concomitante del progreso del tiempo, y, por consiguiente, del devenir.
Vide: HERDAN, G., "The Advanced Theory of Lenguage as Choice and Chance", Berlín, 1966, pp. 257-264; 404-430.

(22) Pueden verse por extenso los resultados del trabajo en: BLACKITH, R. E., "A Multivariate analysis of Latin elegiaic verse", *Language and Speech,* n.º 6, 1963, pp. 196-205.

misma línea del estudio de multivariantes, con crecientes posibilidades de aplicación práctica (23).

c) ANALISIS ESTILISTICO

Uno de los aspectos de la filología tradicional que no podía tardar en suscitar el interés de la lingüística estadística es el del análisis estilístico de obras y autores, por sus propias características tan esencialmente subjetivas. El problema, en líneas muy generales, podría plantearse así: si el estilo de una obra o de un autor es tan esencialmente personal e íntimo, tan inseparablemente unido a una identidad profunda, si "el estilo es el hombre", y, por otra parte, el acercamiento a dicho estilo por parte del crítico, está necesariamente teñido de subjetividad, ¿sería posible realizar acercamientos y análisis cuantitativos del estilo que garantizaran unos resultados concretos, comparables y objetivos?

Las respuestas a esta pregunta varían, aun entre los especializados, desde la postura optimista y largamente razonada de Herdan que dedica más de ciento cuarenta páginas de su obra "The Advanced Theory of Language as Choice and Chance" a la descripción y posibilidades de lo que él llama "estiloestadística", hasta la postura mucho más llena de reservas de Tallentire, del "Literary and Linguistic Computing Centre" de la Universidad de Cambridge, especialmente tal como queda expresada en su obra "Mathematical Modelling in Stylistics: its extent and general Limitations".

Pienso que, sin excesivas ambiciones, ni excesivas reservas, podemos considerar que, mientras no se pierda de

---

(23) BLACKITH, R. E., "Multivariate Morphometrics", Londres, 1971.

vista que las operaciones estadísticas sobre modelos matemáticos de obras literarias son una herramienta auxiliar y no una finalidad en sí mismas, podemos confiar en que repetidas y variadas experiencias vengan a dilucidar y a cuantificar algunos aspectos del estilo que les haga más fácilmente comprensibles y comparables.

Como en apartados anteriores, creo que la mejor ilustración de las finalidades de la estadística lingüística en el dominio del análisis del estilo podrá consistir en la breve descripción de algunas investigaciones realizadas o en curso de realización sobre este tema.

En "The Calculus of the Linguistic Observations", Herdan dedica algunas páginas al estudio de lo que él denomina "aliteración por repetición de la letra inicial de verso". La repetición de un fonema al comienzo de un verso es un procedimiento que, bajo formas diversas ha constituido parte de la técnica poética de diversas literaturas. Este tipo de aliteración aparece en la poesía escocesa; los bardos lo utilizaban como procedimiento mnemotécnico, y en la Biblia encontramos las Lamentaciones redactadas de tal manera que cada grupo de tres versos comienza por la misma letra.

Herdan, al observar en las Geórgicas de Virgilio la frecuente repetición de determinados fonemas al comienzo de verso se planteó la cuestión de si tal procedimiento no formaría también parte de la técnica poética latina. Apuntó también el interés que existiría en extender la investigación no sólo a Virgilio, sino también a otros autores latinos, e intentar averiguar si en la utilización de iniciales de verso no se podría encontrar una nueva característica del estilo de un autor.

Inspirada en estos postulados, Suzanne Govaerts, del "Laboratoire d'Analyse Statistique des Langues Anciennes"

de la Universidad de Lieja, se planteó (24) la posibilidad de que un estudio de este tipo pudiera abrir perspectivas nuevas y llevar a conclusiones interesantes sobre la sensibilización de un autor hacia ciertos sonidos y, a través de este hecho, iluminar su personalidad. Podemos preguntarnos en efecto, si la presencia y la repetición de determinado fonema en un lugar preferente como es la inicial de un verso no refleja una tendencia consciente o *inconsciente* del autor, y si cada autor no presenta en este campo una preferencia que podamos considerar como peculiar de su estilo.

De acuerdo con esta finalidad, S. Govaerts realiza un estudio comparativo de Lucrecio y Virgilio cuyas obras, escritas en el mismo metro, se diferencian claramente por el tema. Lucrecio es ante todo un filósofo; defensor apasionado de una doctrina, su empeño principal consiste en acumular pruebas, en convencer a su interlocutor; las preocupaciones de finura estilística pasan para él a segundo plano; Virgilio, por el contrario, inmerso en una época en la que la poesía latina ha sufrido una gran revolución bajo la influencia de la literatura alejandrina, representa a la poesía culta: quiere realizar una auténtica labor de artista y se muestra tan preocupado por la perfección de la forma como de la idea expresada.

La tesis de Govaerts se plantea en las siguientes palabras: "s'il y a chez les Latins une recherche dans le domaine phonétique, la comparaison entre deux poétes aussi différents doit faire ressortir plus nettement les traits par lesquels cette recherche se manifeste, soit que ces traits relèvent d'une préocupation personnelle propre à chaque poète, soit que, communs à plusierurs auteurs, ils

(24) S. GOVAERTS, "Les Initiales de Vers chez Lucrèce et Virgile", *Statistique et Analyse Linguistique,* P.U.F., París, 1966, p. 40.

traduisent une particularité d'un genre littéraire donné" (25).

Su estudio se limita a los primeros libros respectivamente de *De Rerum Natura* y de la *Eneida.* La metodología, puesta a punto por Herdan parte del principio de que la técnica del poeta en la aliteración del fonema inicial del verso se define esencialmente por una preferencia hacia un intervalo dado entre las diferentes apariciones de un fonema en dicha posición inicial. Diremos, pues, que para dos ocurrencias sucesivas de un fonema en posición inicial, el intervalo es cero, lo que se representa con la notación xx, en la que x simboliza el fonema de que se trate. Si las dos ocurrencias están separadas por un verso, el intervalo es 1, y lo representamos x 1 x, y así sucesivamente.

Con el fin de saber si la longitud de los intervalos para un fonema dado es debida al azar, basta con tomar como punto de comparación la distribución aleatoria. Esta se obtiene calculando, a partir de la probabilidad de un fonema dado en el texto estudiado, la probabilidad de encontrar para este fonema intervalos de diferente valor. La teoría de las probabilidades compuestas nos permite plantear la fórmula general:

$$Pr = p \times (1 - p)r$$

Es decir, que la probabilidad de un intervalo de longitud r es igual a la probabilidad del fonema multiplicada por su no probabilidad elevada a la potencia *r.*

Calcula Govaerts según esta fórmula las probabilidades de los intervalos de diferente valor para los fonemas que constituyen más del 5 % de las iniciales. La representación gráfica de los resultados permite comparar la distri-

(25) *Ibid.*, p. 42.

bución aleatoria y la distribución real. Al final de sus cálculos expone la autora sus conclusiones que brevemente resumimos. En el caso de Lucrecio, cinco fonemas (cuatro letras) destacan inmediatamente, tanto por su frecuencia como por su desviación con respecto a la curva aleatoria: N, Q, C, A, E. Se puede observar una clara preferencia por el intervalo cero en todos los fonemas, excepto en uno (E) en el que el intervalo 2 es el más frecuente. La magnitud de las desviaciones con respecto a la curva teórica permite rechazar la intervención del azar, y, por consiguiente hay que atribuir esta peculiaridad a una preocupación consciente o inconsciente del escritor. Ahora bien, cabe plantearse el dilema de si esta preocupación se aplica puramente al dominio fonético, es decir, si los sonidos han sido recogidos o repartidos en función de sí mismos, o bien si (como resultó ser el caso de Lucrecio) los hechos fonéticos estaban ligados a la semántica y, por encima de ésta, a la materia tratada por el autor y a la estructura misma de su pensamiento.

La repetición de N en inicial de verso se justifica en la mayor parte de los casos por la presencia de palabras como: *non, nec, neque;* es decir, que está ligada a la expresión de la negación, que ocupa en la obra de Lucrecio un lugar muy importante; de temperamento combativo, Lucrecio se alza violentamente contra las ideas de sus contemporáneos, sus prejuicios y sus supersticiones; no los expone en su obra más que para demostrar su vanidad, para negarlos con vehemencia. Para subrayar la idea de negación y darle más peso, Lucrecio coloca frecuentemente las palabras que la expresan al principio del verso; incluso repite la negación al principio de dos versos sucesivos. Es decir, que los hechos observados a propósito de N no son puramente fonéticos, sino que están fundamentalmente condicionados por la personalidad profunda del autor y su actitud general frente a la vida. De la observación de los otros fonemas, se llega a resultados simila-

res; las aliteraciones iniciales de Lucrecio no son función del valor intrínseco sonoro de un fonema, sino más bien de la significación de la palabra de la que ese fonema es inicial, y, por consiguiente, función en gran medida del tema tratado.

En el caso de Virgilio, por el contrario, los resultados adquieren un carácter completamente diferente. Al examinar las diversas gráficas resulta evidente que la distribución real de la mayor parte de los fonemas es considerablemente inferior a lo que cabría esperar de las curvas de distribución aleatoria. Es decir, que parece poderse afirmar que, en la *Eneida,* Virgilio actúa como si evitara la sucesión inmediata, o en intervalos breves, de versos que comienzan por el mismo fonema. Virgilio parece sensible a los sonidos en sí mismos independientemente del significado de las palabras de que forman parte y evita la repetición a intervalos cortos de sonidos idénticos. Este hecho resulta aún más interesante si lo comparamos con la situación en las *Geórgicas,* donde Virgilio se inclina marcadamente por los intervalos de valor cero, uno y dos, es decir, por la repetición del fonema inicial de los versos en intervalos breves. Por consiguiente, desde este punto de vista, la Eneida se distingue de las Geórgicas por una composición más cuidada, más selecta, cuidadosa de evitar el menor choque a la estética personal en el dominio de la fonética. Esta comparación vendría a confirmar la hipótesis de Herdan según la cual el estudio de las iniciales de verso podría revelar una evolución en el estilo de Virgilio. También este estudio podría proporcionar elementos para un mejor conocimiento de la llamada "memoria inmediata": ¿en qué momento la memoria del poeta pierde conciencia de una repetición? ¿Es constante esta memoria para todos los fonemas? ¿Existen fonemas privilegiados por motivos de su propia sonoridad o por motivos de la búsqueda de un efecto transitorio?

Como se puede ver, la estiloestadística no sólo persigue fines concretos, sino que de la persecución de esos fines brota un abanico de nuevas metas de investigación con unas posibilidades que sólo el futuro nos dirá hasta dónde pueden llegar.

Otro trabajo de gran interés sobre análisis estilístico en curso de realización es el que está llevando a cabo J. Leighton del Departamento de Alemán de la Universidad de Bristol. Mientras el trabajo de Suzanne Govaerts a que nos acabamos de referir es de relativa simplicidad desde el punto de vista metodológico, el de Leighton implica una metodología mucho más compleja cuya resolución requiere la utilización de un computador, en este caso el Elliott 503 del Centro de Computación de la Universidad de Bristol. Aunque la labor aún prosigue, Leighton expone su metodología y resultados de las pruebas previas en su trabajo "Sonnets and Computers: an experiment in stylistic analysis..." (26). El proyecto consiste en un estudio del soneto alemán en el siglo XVII, con especial atención a los sonetos escritos por circunstancias especiales, los que Leighton llama "sonetos ocasionales", y refiriéndose no sólo a los autores conocidos, sino también a los sonetos ocasionales escritos por ciudadanos alemanes de cierto nivel cultural, pero sin especiales ambiciones literarias; con ello se confía en que surgirá una especie de geografía literaria con énfasis especial en ciudades como Nurenberg, Breslau, Hamburgo, Danzig, Leipzig y Dresden. Una de las finalidades secundarias, dentro del estudio general, consiste en intentar determinar si ciertas ciudades o ciertos grupos de poetas se distinguen unos de otros por una tendencia hacia técnicas estilísticas específicas. Pronto comenzaron a acumularse los problemas: existen tantas

---

(26) Joseph LEIGHTON, "Sonnets and Computers: an experiment in stylistic analysis using an Elliott 503 computer", en *The Computer in literary and linguistic research,* C.V.P., Cambridge, 1971, p. 149.——

posibilidades de analizar un soneto, que cualquier decisión subjetiva, con tal de que sea consistente a lo largo de la investigación, está justificada; pero es mucho más difícil archivar la información de cada soneto en particular de tal manera que, cuando llegue el momento de comparar los varios millares de sonetos que se han de uilizar en el estudio, sea posible separar los datos interesantes y estrictamente comparables, de los datos menos interesantes. En este punto es cuando se hizo evidente que sería necesario el uso de un computador para la realización del trabajo. Otro problema básico lo constituyó el de la utilización de la estilística comparativa; tradicionalmente, la descripción del estilo ha consistido en racionalizar impresiones subjetivas, basadas en ejemplos cuidadosamente seleccionados; aunque esta actitud pueda ser considerada adecuada en sí misma como acercamiento a los problemas del estilo, parecía poco satisfactoria para enfrentarse con la discusión de varios millares de sonetos; la inevitable decisión, pues, hubo de ser la de utilizar datos cuantitativos para la descripción del análisis del estilo. Ahora bien, había llegado el momento de decidir qué datos eran los que habían de ser cuantificados.

Para llegar a ello, Leighton observa que la rigidez de la estructura de su rima impone ciertos límites en la clase de manifestación verbal posible dentro del soneto; de ello deduce que puede esperarse que el estilo y los manerismos de un poeta se revelen claramente a través de la manera en que intenta adaptar su expresión a los requerimientos de la forma del soneto y que, en este aspecto, la estructura de las frases puede constituir el indicador más importante del estilo personal. La descripción de la estructura de la oración no es, ciertamente un problema que haya sido resuelto sin ambigüedades, pero en el caso del soneto no era sólo la estructura de la oración lo importante, sino su relación con la estructura de cada verso.

Fue J. Sinclair (27) quien sugirió la posibilidad de llevar a cabo un análisis relativamente simple basado en la distinción entre oraciones principales y subordinadas. En resumen, la técnica de Sinclair puede quedar ilustrada con la siguiente estrofa del poema "First Sight" de Philip Larkin:

a Lambs that learn to walk in snow
b When their bleating clouds the air,
(-a)a- Meet a vast unwelcome, know
(-a) Nothing but a sunless glare.

donde *a* representa una oración principal, *b* una subordinada, a- y b- indican que una oración queda interrumpida por el final de verso o por otra oración y (-a) y (-b) indican la conclusión de una oración interrumpida.

Basándose en este simple esquema Leighton elaboró otro mucho más complicado que se hacía necesario para poder incluir las técnicas retóricas del siglo XVII, y que está compuesto por diecinueve caracteres distintos, y de los que cito algunos como ejemplo:

a oración principal
b oración principal interrumpida
i oración subordinada
n final de oración
p anacoluto
s final de línea y oración.

Aplicando su sistema al famoso soneto de Donne "Death be not proud..." éste quedaría codificado para el computador de la siguiente manera:

---

(27) J. SINCLAIR, "Taking o Poem to Pieces", en *Essays on Style and Language,* Londres, pp. 68-81.

ajr kir jlmr kfir bir cr ar ess ar dr

ar daqs mar aas.

Para probar su método y como prólogo para el trabajo definitivo, Leighton realizó un experimento inicial de comparación de 20 sonetos de Paul Fleming con 31 sonetos de Andreas Gryphius. Aunque la elección de estos sonetos fue básicamente arbitraria, estuvo sin embargo, condicionada por tres factores importantes: primero, todos los sonetos están escritos en versos alejandrinos, lo que les hace más fácilmente comparables; segundo, los sonetos de Fleming están ordenados cronológicamente, y, en tercer lugar, los sonetos de Gryphius constituyen primeras versiones que fueron posteriormente muy revisadas por el autor, con lo que proporcionan una excelente base para ulteriores comparaciones. Anoto a continuación algunos de los resultados a que llega Leighton:

Frecuencias medias de algunas variables

	Fleming	*Gryphius*
Oraciones principales por período	1,65	2,62
Oraciones subordinadas	0,88	3,57
Oraciones principales por poema	13,45	7,25
Períodos por poema	8,15	2,77
Encabalgamientos	1,95	2,80
Punto al final del 2.º cuarteto	80,00	32,26

Estos resultados, que ya de por sí evidencian claras diferencias estilísticas entre los dos autores, serán mucho más significativos cuando existan frecuencias "normales" con las que puedan ser comparadas. Cuando los valores de los resultados correspondientes a los sonetos de Fleming fueron representados gráficamente, los últimos sone-

tos cronológicamente mostraban mucha menos variación de la media que los primeros ¿es esto una indicación de mayor madurez de estilo y de más experimentación con el lenguaje en sus primeros tiempos? Por otra parte, Gryphius muestra mucha menos desviación de la media en todos sus sonetos que Fleming.

En resumen, Leighton cree haber encontrado un método de análisis estadístico que le permitirá realizar un análisis estilístico del soneto alemán en el siglo XVII que puede ofrecernos interesantes aspectos y resultados y que, a la vez, puede ser aplicado a otros autores, otras épocas y otras lenguas.

Es decir, que, en cuanto a la finalidad de la estadística lingüística como investigación del estilo, podemos afirmar: primero, que se están realizando numerosos trabajos dentro de este ámbito; segundo, que los resultados de algunos de ellos pueden ser calificados de esperanzadores; y tercero, que, sin perder de vista que la estadística no es más que una ciencia auxiliar de la lingüística, podemos confiar en que, con metodologías adecuadas, pueda llegar a ayudarnos eficazmente a la objetivización del estudio del estilo y a una cuantificación que haga mucho más fácil y directa la comparación entre las diversas obras de un autor, entre diversos autores y entre diversas épocas y géneros literarios.