LENGUA NATURAL Y LENGUAJE LEXICOGRAFICO

Manuel Ollero Toribio
Miguel A. Pineda Pérez

This paper deals with the analysis of the language used in the witing of dictionaries. We try to prove that the lexicographic norm responds to a specific linguistic system from which it can be deduced certain general characteristics. The matter is completed with a statistical survey of the corpus we have studied.

Introducción

Aunque la lexicografía ha sido siempre una disciplina que ha motivado el interés y la atención de los filólogos, en el momento actual adquiere una especial trascendencia debido a la aplicación de los nuevos métodos informáticos al desarrollo de los diccionarios. Por una parte los nuevos diccionarios se confeccionan siguiendo técnicas más rigurosas que son posibles gracias a la utilización de instrumentos informáticos que «vigilan» la coherencia en la redacción de los artículos. Por otra, los diccionarios resultantes en formato computacional (*machine readeable form*), tienen la posibilidad de ser utilizados como instrumentos para sistemas de procesamiento automático del lenguaje natural (Traducción Automática, Indización Automática de Textos, Interfaces en Lenguaje Natural, Sistemas de Análisis del Habla, etc.).

En este segundo aspecto el diccionario es un objeto que será sometido a análisis para obtener una base de datos lexicográfica. El proceso de análisis implica, necesariamente, la intervención de un *parser*, es decir, de un sistema de análisis, compuesto generalmente de una gramática formalizada y de un lexicón.

En principio este tipo de parser puede ser el mismo utilizado para el análisis de un

texto. Sin embargo, la experiencia de los lexicógrafos computacionales demuestra que la lengua utilizada en el discurso lexicográfico es un subconjunto de la lengua natural, por lo cual el *parser lexicográfico* tiene que ser específico.

Si conseguimos delimitar este subconjunto mediante la definición de una nueva gramática y un nuevo lexicón, podremos obtener un parser más adecuado para el análisis del lenguaje lexicográfico.

Una posible forma de definir tanto la gramática como el lexicón es realizar un estudio contrastivo entre el texto lexicográfico y el texto de la lengua natural.

Haciendo abstracción de la información enciclopédica contenida en la generalidad de los diccionarios, la primera caracterización evidente del texto lexicográfico es su clara función metalingüística restringida a una sola parte del signo, es decir, a la infraestructura de la significación absoluta, puesto que en un diccionario, –salvo en los denominados gramaticales– no se explica el comportamiento sintáctico de las unidades léxicas ni las reglas de composición de los enunciados. Por este motivo será preferible hablar de metalenguaje lexicográfico o de la metalengua lexicográfica, tal como hace R. Werner[1].

Esta función contenida en el discurso lexicográfico es un hecho del que los lexicógrafos han sido conscientes desde antiguo, y en la bibliografía actual se hace frecuente referencia a esta cuestión en aspectos concretos, como los de la paráfrasis lexicográfica, la definición sinonímica, el círculo vicioso en la definición, los estratos léxicos, etc.[2].

Por otra parte, sabemos que todas las funciones del lenguaje sólo son analizables lingüísticamente cuando utilizan procedimientos formalizables, pertinentes o convencionales, de tal manera que podremos decir que un diccionario comporta un metalenguaje lexicográfico si es posible demostrar que existe una técnica de discurso propia y específica de él, distinta de la técnica de discurso del lenguaje habitual o normal, y que puede ser formalizable mediante una gramática particular.

Intentar demostrar esta última afirmación, o ensayar un procedimiento mediante el cual se pueda llevar a cabo la demostración, es el propósito de este trabajo.

Nos limitaremos aquí al estudio estadístico-contrastivo de las formas usadas en el discurso normal y el discurso lexicográfico, para saber si la distribución estadística de los lexemas sustantivos y verbales, así como la de algunos elementos de relación (preposiciones y conjunciones) presenta una dispersión suficientemente importante para poder afirmar que se trata de dos técnicas de discurso diferenciadas.

Método

De la lengua española estándar, por el momento, sólo existe un diccionario de frecuencias: el *Frequency Dictionary of Spanish Words* (DF) de A. Juilland y E. Chang Rodríguez,

---

[1] R. Werner, *La Lexicografía*, Madrid, 1982.

[2] J. Fernández Sevilla, *Problemas de Lexicografía Actual*, Bogotá, 1974.

que es el instrumento que usaremos como patrón estadístico para establecer cuáles son los rangos de las palabras más frecuentes en la norma estándar de nuestra lengua.

En la segunda parte de este diccionario (pp. 383-500) encontramos, desde el rango 1 hasta el 5024, los lexemas con sus respectivas frecuencias. De entre ellos hemos seleccionado las cien primeras palabras, y de entre éstas los sustantivos y verbos. La elección de estas cien primeras palabras se justifica porque suponen el diez por ciento del total de formas que componen el diccionario.

Estos mismos sustantivos y verbos son los que determinan la otra parte de nuestra muestra, que hace referencia al discurso lexicográfico. Así, hemos tomado la primera acepción de todos y cada uno de ellos en tres de los diccionarios más prestigiosos de nuestra lengua: el *Diccionario de la Real Academia Española* (DRAE), el *Diccionario Ideológico* (DI) de Julio Casares y el *Diccionario de Uso del Español* (DUE) de María Moliner.

El conjunto de enunciados que componen estas acepciones será la muestra que aceptaremos como representativa del discurso lexicográfico. Garantizar la representatividad de la muestra es la causa que nos ha obligado a renunciar a incluir en esta parte del corpus las acepciones de las otras categorías incluidas también en el rango cien. Y es que la definición que en un diccionario se da de una preposición, conjunción, artículo o pronombre constituye un enunciado cuya función metalingüística, como decíamos en el apartado anterior, se diferencia notablemente de la que denominábamos lexicográfica, y su inclusión en el corpus produciría una importante desviación no deseable, ya que en la totalidad de un diccionario la definición de estos elementos supone una parte ínfima.

En cuanto a los adjetivos, no han sido considerados porque entre las palabras del rango elegido no aparecen calificativos sino demostrativos, posesivos, numerales, etc., cuyas definiciones presentan los problemas antes mencionados. Siguiendo los principios lexicométricos, la muestra lexicográfica hubo de ser segmentada así como computadas las frecuencias de cada forma. Dado que esta tarea es ardua si se realiza manualmente, hemos recurrido a las técnicas de análisis estadístico automatizado.

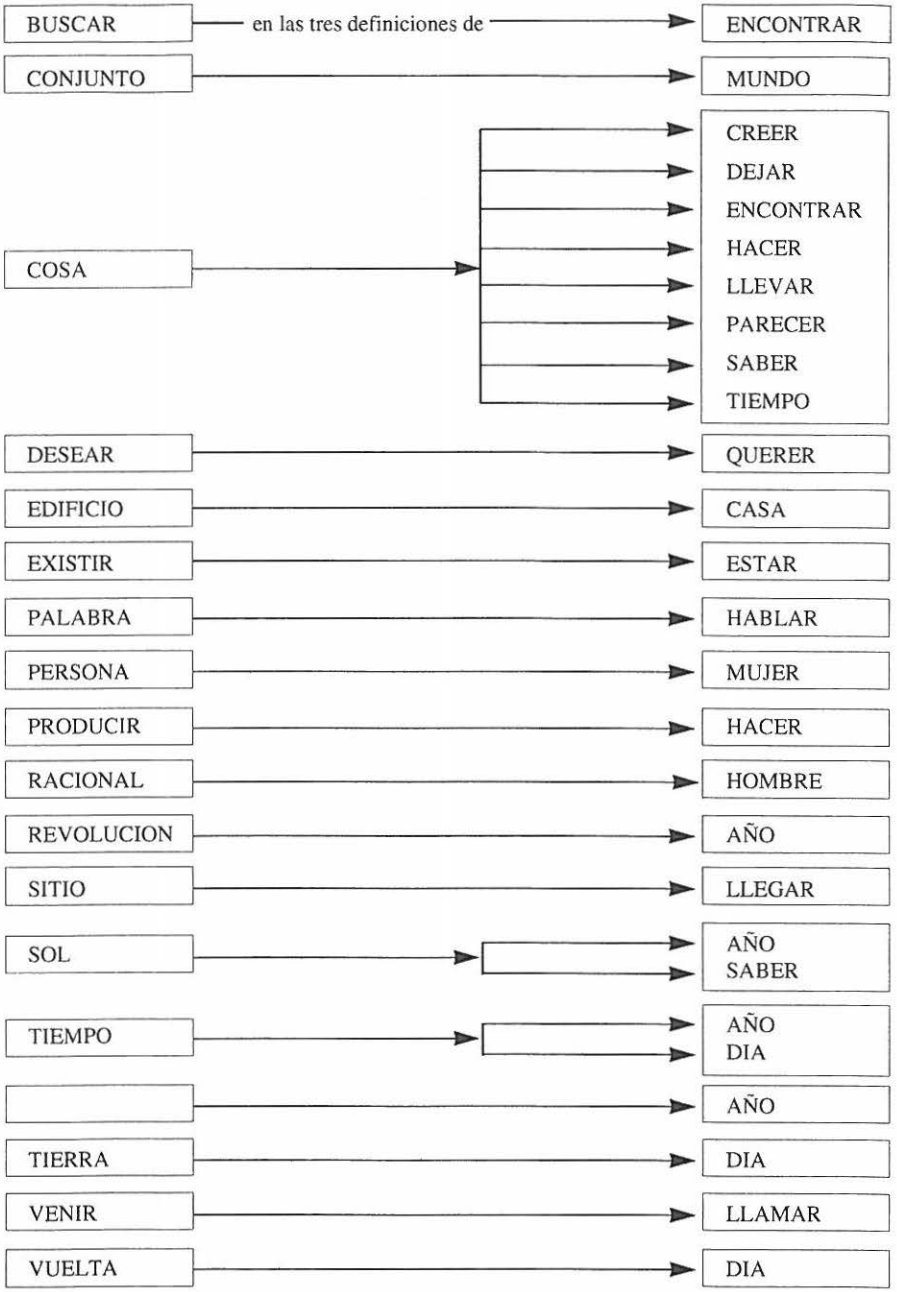
Después de introducir el texto en ficheros de ordenador, se redactaron una serie de programas en lenguaje informático que permitieron obtener los datos correspondientes a las frecuencias en los tres diccionarios.

Para establecer el contraste entre el discurso normal y el lexicográfico con mayor riqueza de matices dividimos este último en cuatro estratos:

1. El conjunto de acepciones de los tres diccionarios, que aparecerá con la etiqueta D's.
2. Las acepciones que aparecen en el DRAE.
3. Las acepciones que aparecen en el DI.
4. Las acepciones que aparecen en el DUE.

De cada uno de estos estratos se obtiene, mediante un primer programa, la lista de formas con referencia a la entrada del diccionario en que aparece, y las veces que lo hace en esa acepción. Así, como ejemplo, el adjetivo *abstracta* aparece en las definiciones de «COSA» en el DRAE y en el DI, y en ambas una vez.

El listado de formas definitorias con referencia al término definido en cada uno de los diccionarios reveló que son pocas las palabras que aparecen en las tres definiciones de los mismos términos, en los tres diccionarios. Las coincidencias son las siguientes:



Posteriormente se obtiene la frecuencia de cada forma y la ordenación decreciente de las mismas (rangos).

Este estudio se realiza para el conjunto de los tres diccionarios y, de forma independiente, para cada uno de ellos. Al mismo tiempo se han considerado como posibles variables que pueden tener importancia posteriormente, el que la frecuencia de cada forma se dé en la definición de un sustantivo o de un verbo.

El total de formas que compone el corpus lexicográfico es de 1111, 483 en la definición de sustantivos y 628 en la definición de verbos. Teniendo en cuenta que son 12 los sustantivos definidos y 22 los verbos, el promedio de formas en la definición de estas categorías es de 13,42 formas en los sustantivos y 9,51 en los verbos.

En el DRAE el total de formas es de 359, 194 en la definición de sustantivos y 165 en la de verbos, con un promedio de 16,16 en los sustantivos y 7,5 en los verbos.

En el DI el total de formas es de 300, 136 en los sustantivos y 164 en los verbos, con un promedio de 11,33 en los sustantivos y 7,45 en los verbos.

En el DUE el total de formas es de 452, 153 en los sustantivos y 299 en los verbos, con un promedio de 12,75 en la definición de sustantivos y 13,59 en la definición de verbos.

De estos datos se deduce que el DUE es el diccionario que más formas emplea para la definición de los 34 lexemas considerados.

En cuanto a las definiciones de sustantivos es el DRAE el que más formas emplea y, en cambio, es del DUE el que más emplea en las definiciones de verbos.

En lo que respecta a la diversidad de formas usadas en el discurso lexicográfico también se hacen notar diferencias considerables que ponen de relieve distintos grados de riqueza léxica.

Así, en el conjunto de los tres diccionarios son 367 formas las representadas, 197 en las definiciones de sustantivos y 211 en las de los verbos.

En el DRAE aparecen 191 formas distintas, 106 en las definiciones de los sustantivos y 108 en las de los verbos.

En el DI aparecen 167 formas distintas, 93 en las definiciones de los sustantivos y 92 en las de los verbos.

Y en el DUE, 217 formas distintas, 111 en las definiciones de los sustantivos y 124 en las de los verbos.

Por tanto es el DUE el que demuestra una riqueza de léxico mayor, seguido del DRAE y el DI.

En cuanto a las palabras más frecuentemente usadas en el conjunto de los tres diccionarios son:

S-V	S	V
1° **de**	1° **de**	1° **cosa**
2° **que**	2° **que**	2° **o**
3° **o**	3° **la**	3° **una**

En el DRAE:

S-V	S	V
1° **o**	1° **de**	1° **o**
2° **de**	2° **la**	2° **cosa**
3° **una**	3° **que**	3° **una**

En el DI:

S-V	S	V
1° **de**	1° **de**	1° **cosa**
2° **una**	2° **que**	2° **una**
3° **cosa**	3° **o**	3° **o**

En el DUE:

S-V	S	V
1° **que**	1° **que**	1° **de**
2° **de**	2° **en**	2° **que**
3° **la**	3° **de**	3° **cosa**

Aunque estos datos serán objeto de un análisis más cuidadoso en el apartado siguiente, en una primera impresión se constata que el primer lexema en los tres diccionarios es «cosa», y que la conjunción «o» aparece siempre en lugar destacado.

Análisis Contrastivo

Como decíamos al principio nuestro interés se centra en el análisis del discurso normal y el lexicográfico para saber si existen diferencias que puedan permitirnos afirmar que el lenguaje lexicográfico es distinto del lenguaje normal.

Para ello emplearemos pruebas estadísticas que nos darán el grado de probabilidad con que puede aceptarse la veracidad de esta afirmación. En concreto usaremos el test de correlación de rangos de Spearman y el test de concordancia de Kendall[3].

El primero da una medida de la asociación que existe entre dos series ordenadas de datos según su frecuencia. Esta medida, o coeficiente de correlación, se calcula mediante la expresión:

$$1 - \frac{6 \sum d^2}{N (N^2-1)}$$

en donde *N* es el número de datos en ambas ordenaciones y *d* la diferencia de rangos.

El test de concordancia de Kendall, a diferencia del anterior, permite comparar más de dos ordenaciones, y ha sido empleado para comparar el grado de concordancia entre los tres diccionarios. Para conocer el valor de este coeficiente se empleará la expresión:

$$\frac{12 \cdot S}{K^2 \cdot N(N^2-1)}$$

donde *K* es el número de ordenaciones, *N* el número de datos en cada ordenación y *S* la suma de cuadrados de las diferencias entre las sumas parciales de los rangos y el valor medio de todos ellos.

Estas pruebas serán aplicadas en los siguientes casos:

1º los SUSTANTIVOS usados en las definiciones, comparando

- a) el diccionario de frecuencias (DF) con:
  - a.1.) el conjunto de los tres diccionarios (D's)
    - a.1.1.) en las definiciones de sustantivos y verbos
    - a.1.2.) en las definiciones de sustantivos
    - a.1.3.) en las definiciones de verbos
  - a.2.) el DRAE
    - a.2.1.) en las definiciones de sustantivos y verbos

---

[3] Ch. Muller: *Estadística Lingüística*, Madrid, 1978.

a.2.2.) en las definiciones de sustantivos
a.2.3.) en las definiciones de verbos

a.3.) el DI

a.3.1.) en las definiciones de sustantivos y verbos
a.3.2.) en las definiciones de sustantivos
a.3.3.) en las definiciones de verbos

a.4.) el DUE

a.4.1.) en las definiciones de sustantivos y verbos
a.4.2.) en las definiciones de sustantivos
a.4.3.) en las definiciones de verbos

b) el DRAE con:

b.1.) el DI

b.1.1.) en las definiciones de sustantivos y verbos
b.1.2.) en las definiciones de sustantivos
b.1.3.) en las definicioens de verbos

b.2.) el DUE

b.2.1.) en las definiciones de sustantivos y verbos
b.2.2.) en las definiciones de sustantivos
b.2.3.) en las definiciones de verbos

c) el DI

c.1.) con el DUE

c.1.1.) en las definiciones de sustantivos y verbos
c.1.2.) en las definiciones de sustantivos
c.1.3.) en las definiciones de verbos

d) las definiciones de sustantivos y las definiciones de verbos:

d.1.) en los tres diccionarios
d.2.) en el DRAE
d.3.) en el DI
d.4.) en el DUE

e) en los tres diccionarios al mismo tiempo (test de Kendall)

e.1.) en las definiciones de sustantivos y verbos
e.2.) en las definiciones de sustantivos
e.3.) en las definiciones de verbos

2° los VERBOS usados en las definiciones, en la misma secuencia de comparaciones anteriores: a,b,c,d, y e.

3° las PREPOSICIONES, comparaciones a,b,c,d y e.

4° las CONJUNCIONES, igualmente, comparaciones a,b,c,d y e.

LOS SUSTANTIVOS

Los sustantivos estudiados son aquellos que aparecen en los tres diccionarios con frecuencias más elevadas:

	RANGOS				
	DF	D's	DRAE	DI	DUE
COSA	1	1	1	1	1
PERSONA	2	2	2	2	5.5
LUGAR	3	3	3	9	8.5
TIEMPO	4	4	4	3	3
NOMBRE	5	5.5	7	5.5	3
SUJETO	6	5.5	7	5.5	3
FACULTAD	7	7.5	7	5.5	8.5
PENSAMIENTO	8	7.5	7	5.5	8.5
ACCION	9	9.5	7	9	8.5
ASPECTO	10	9.5	10	9	8.5

En el cuadro 1 se recogen los valores de los distintos coeficientes calculados. Así, la primera casilla del cuadro (0.99) corresponde a la comparación a.1.1, la siguiente (0.68) a la a.1.2, etc. Dado que los niveles de significación vienen especificados en todos los cuadros, la interpretación de los coeficientes puede ser fácil. Por ejemplo, en la comparación a.1.1 podemos decir que el coeficiente obtenido, que es superior a 0.75, nos permite afirmar con menos del 1 por ciento de error (más del 99 por ciento de certeza) que existe una fuerte concordancia entre el diccionario de frecuencias y el conjunto de los tres diccionarios en cuanto a los sustantivos usados en las definiciones de sustantivos o verbos indistintamente. Sin embargo esta concordancia sólo puede aceptarse con un 95 por ciento de certeza en el caso de la comparación entre el diccionario de frecuencias y el conjunto de los tres diccionarios en cuanto a los sustantivos usados en las definiciones de sustantivos, puesto que el coeficiente obtenido (0.68) está por debajo de 0.75. Por último, el coeficiente obtenido en la comparación de los sustantivos usados en las definiciones de verbos del DI y el diccionario de frecuencias (0.49) nos permite afirmar que no existe concordancia alguna entre ambos diccionarios.

Cuadro 2

Coeficientes de concordancias de los VERBOS usados en las definiciones.

Niveles de significación: al 95% : 0.83
al 99% : 0.94

Verbos	D's			DRAE			DI			DUE		
	S-V	S	V	S-V	S	V	S-V	S	V	S-V	S	V
DF	0.60	0.54	0.72	0.54	0.10	0.67	0.40	0.40	0.40	0.06	0.2	0.09
DRAE							0.51	0.20	0.17	0.08	0.23	0.24
DI										0.03	0.44	0.04
S:V	0.20			0.96			1.00			-0.19		

DRAE : DI : DUE
S-V : 0.03
S : 0.35
V : 0.42

Niveles de significación: al 95% : 0.65
al 99% : 0.77

Las Preposiciones

Las preposiciones estudiadas con los mismos criterios que lo fueron sustantivos y verbos son las siguientes:

	RANGOS				
	DF	D's	DRAE	DI	DUE
DE	1	1	1	1	1
A	2	2	2	2	2
EN	3	3	3	5	3
POR	4	5	5	6	6
CON	5	6	6	4	4
PARA	6	4	4	3	5
SIN	7	9.5	9	9	9.5
SOBRE	8	9.5	9	9	9.5
ENTRE	9	9.5	9	9	9.5
HASTA	10	9.5	9	9	9.5
DESDE	11	7	9	9	7

Tal como puede comprobarse por los coeficientes obtenidos (cuadro 3) cualquiera que sea la comparación establecida no existen nunca valores que nos permitan hablar de diferencias significativas entre la norma estándar del español y la norma lexicográfica.

Cuadro 3

Coeficientes de concordancias de los PREPOSICIONES usados en las definiciones.

Niveles de significación: al 95% : 0.53
al 99% : 0.73

Preposiciones	D's			DRAE			DI			DUE		
	S-V	S	V	S-V	S	V	S-V	S	V	S-V	S	V
DF	0.86	0.82	0.79	0.92	0.82	0.90	0.88	0.80	0.71	0.86	0.92	0.94
DRAE							0.96	0.95	0.88	0.95	0.94	0.91
DI										0.94	0.99	0.80
S:V	0.79			0.86			0.77			0.82		

DRAE : DI : DUE
S-V : 26.7
S : 22.5
V : 23.1

Niveles de significación: al 95% : 18.37
al 99% : 23.21

De la misma forma tampoco se comprueban diferencias entre el uso que de las preposiciones se hace en las definiciones de sustantivos y verbos.

Por último, tampoco se descubren diferencias entre las tres normas lexicográficas.

Las Conjunciones

Dentro de esta categoría hemos considerado, siguiendo la misma pauta, las siguientes conjunciones:

	RANGOS				
	DF	D's	DRAE	DI	DUE
Y/E	1	3	3	3	4
QUE	2	2	2	2	1
COMO	3	4	6.5	6.5	3
PERO	4	7	6.5	6.5	7
O	5	1	1	1	2
SI	6	7	6.5	6.5	7
PORQUE	7	7	6.5	6.5	7
PUES	8	7	6.5	6.5	7
NI	9	7	6.5	6.5	7

Al igual que en el caso de las preposiciones, el uso de las conjunciones no delata ninguna diferencia entre la norma estándar y la lexicográfica. Tampoco entre las tres normas lexicográficas, más aún, el uso es igual en los tres diccionarios cuando lo que se define es un sustantivo, o, en el DRAE, entre las definiciones de sustantivos y verbos.

A pesar de que esta afirmación es válida para el conjunto de las conjunciones hay que hacer notar que en la norma lexicográfica hay un uso preponderante de la conjunción «o», que, obsérvese en los rangos anteriores, es siempre la más frecuente en el DRAE y el DI, y la segunda más frecuente en el DUE.

Cuadro 4

Coeficientes de concordancias de los CONJUNCIONES usados en las definiciones.

Niveles de significación: al 95% : 0.60
al 99% : 0.78

Conjunciones	D's			DRAE			DI			DUE		
	S-V	S	V	S-V	S	V	S-V	S	V	S-V	S	V
DF	0.70	0.66	0.69	0.61	0.66	0.66	0.61	0.66	0.61	0.72	0.66	0.70
DRAE							1.00	1.00	0.98	0.86	1.00	0.88
DI										0.86	1.00	0.86
S:V	0.89			1.00			0.98			0.88		

DRAE : DI : DUE
S-V : 0.69
S : 0.82
V : 0.69

Niveles de significación: al 95% : 0.58
al 99% : 0.67

Conclusiones

De este estudio, teniendo en cuenta lo reducido del corpus y que por tanto las conclusiones a las que podemos llegar son necesariamente provisionales, se deduce que la categoría de verbo es lo único que diferencia la norma estándar de la norma lexicográfica; y considerada en cualquiera de sus dos aspectos: como elemento definidor –que forma parte de una definición– o como elemento definido en un diccionario.

También el uso de los sustantivos hace pensar que es legítima la diferencia entre lo que hemos llamado norma estándar y norma lexicográfica cuando aquellos son elementos definidores de otras categorías, y sobre todo cuando la categoría definida es un verbo.

En cuanto a los elementos de relación, preposiciones y conjunciones, no permiten, por sí mismos, en cuanto elementos definidores, establecer ninguna diferencia entre el discurso normal y el lexicográfico.

Bibliografía

Alvar Ezquerra, M.: «Lexicografía» en *Introducción a la Lingüística actual*, Madrid, 1983

Catach, M.: *Orthographe et Lexicographie*, Paris, 1971

Dubois, J.: *Introduction à la Lexicographie: le Dictionaire*, Paris, 1971

Fernández Sevilla, J.: *Problemas de Lexicografía Actual*, Bogotá, 1974

Hervey, S.: *Axiomatic Semantics*, Edinburgh, 1979

Muller, CH.: *La Statistique Linguistique*, Paris, 1968

Principes et Méthodes de Statistique Lexicale, Paris, 1977

Perrot, J.: «Le Lexique» en *Le Langage* (A. Martinet, ed.), Paris, 1968

Werner, R.: *La Lexicografía*, Madrid, 1982