TRABAJO METODOLÓGICO

UNA METODOLOGÍA PARA EL ANÁLISIS ESTADÍSTICO DE DATOS TEXTUALES: EL PROGRAMA ALCESTE

por
Javier Gil Flores, Eduardo García Jiménez
y Gregorio Rodríguez Gómez
Dpto. D.O.E. y M.I.D.E.
Universidad de Sevilla

RESUMEN

Las palabras empleadas dentro de un texto son consecuencia de los lugares semánticos desde los cuales el emisor emite su discurso. En este trabajo, utilizando datos textuales recogidos en el curso de una investigación educativa, mostramos cómo el estudio estadístico de la distribución del léxico empleado en un texto permite detectar la estructuración de los significados presentes en el mismo. Para ello seguimos la metodología ALCESTE, basada en técnicas como la clasificación jerárquica descendente o el cálculo de χ^2.

ABSTRACT

The words used in a text are product of the semantic position where speaker place's one self when talk. In this article, using textual data from educational research, we try to show how the statistical study of the lexical occurances distribution in a text allows detecting the structure of meanings within the text. So that we use the ALCESTE methodology, based on techniques such as hierarchical classification or χ^2.

Analizar estadísticamente textos puede resultar contradictorio, si ponemos frente a frente el carácter verbal de la información escrita y la naturaleza numérica propia del tipo de datos a los que se aplican las técnicas estadísticas. En cierto modo, no nos sorprende la actitud de quienes descartan cualquier posibilidad de análisis

cuantitativo con el material textual, argumentando que difícilmente los números podrían reflejar los sentidos polisémicos, los sutiles significados que se esconden tras un determinado uso del lenguaje y el entramado de relaciones que aveces únicamente la intuición y perspicacia del analista cualitativo, apoyadas en un buen conocimiento del contexto en que fueron producidos los datos, pueden llegar a desentrañar.

La oposición entre el cálculo numérico, ajeno al sentido de las palabras, y la sutileza y multiplicidad de niveles de significación asociados al discurso parecen evidentes. Sin embargo, entre las características del discurso puede ser considerada la repetición de unidades elementales (generalmente las palabras), característica susceptible de tratamiento cuantitativo que pueden complementar los hallazgos a los que se llegue por otros procedimientos. Concretamente, en este artículo nos introduciremos en la metodología ALCESTE para el análisis estadístico de datos textuales, ilustrando su desarrollo a partir de los datos recogidos en una investigación realizada en el campo educativo.

El *análisis estadístico de textos* tiene su origen en los análisis cuantitativos realizados sobre obras literarias, que iban dirigidos al recuento de palabras, el estudio de la distribución del vocabulario, la comparación del léxico empleado por distintos autores o por un mismo autor en diferentes períodos creativos. Las investigaciones realizadas por Yule (1944), Zipf (1946), Guiraud (1960), Muller (1968), entre otros, y el posterior desarrollo y popularización de la informática se encuentran en la base de los métodos de la denominada *estadística textual*, que han acabado aplicándose al estudio de los datos textuales en muy diversos ámbitos: historia, literatura, sociología, educación, etc.

La aparición del programa GENERAL INQUIRER (Stone y otros, 1966) marcó el inicio del análisis automático de textos. El procedimiento desarrollado por este programa se basaba en la búsqueda y recuento de palabras y frases previamente identificadas por el analista mediante la definición de un diccionario confeccionado para el análisis. Mochmann (1983) describe programas posteriores basados en estos mismos principios: EVA, SPENCE, COFTA, COTAG, TEXPACK. En otros programas, en lugar de partir de un diccionario previo, se han intentado extraer los temas presentes en un texto sometiendo a tratamientos estadísticos las frecuencias de cada palabra en cada una de las unidades consideradas. Es el caso del programa WORDS (Iker, 1975). La idea base sigue siendo el recuento de unidades, para realizar cálculos estadísticos a partir de su recuento.

Los métodos de la escuela francesa de análisis de datos, desarrollados a partir de las aportaciones de Benzécri (1973), resultan especialmente adecuados para el análisis de grandes matrices de datos como las originadas al examinar la distribución de unidades elementales dentro de un texto. Enmarcados en esta línea, se han difundido programas específicamente diseñados para el análisis de datos textuales, tales como SPAD.T (Lebart, Morineau y Bécue, 1989) o LEXICO1 (Salem, 1990).

Se inspira igualmente en los métodos de análisis de datos de la escuela francesa el procedimiento de análisis que ofrece el programa ALCESTE (Analyse Lexicale

par Contexte d'un Ensemble de Segments de Texte), creado por Max Reinert (1986)[1]. Su finalidad es *analizar la estructura de la distribución del vocabulario en un corpus textual* (respuestas abiertas en cuestionarios, entrevistas, diarios, obras literarias, etc.). A diferencia del enfoque lexicométrico propuesto por Lebart y Salem (1988), no se trata de comparar textos diferentes o diferentes subcorpus resultantes de una partición inicial del texto, sino de trabajar con el conjunto de datos considerados como un todo unitario.

El supuesto de partida en la metodología ALCESTE es que el emisor, durante su elocución, toma como referencia determinados lugares semánticos, compuestos de significados socialmente construidos, desde los cuales elabora su discurso. Tales posicionamientos implican el uso de un determinado vocabulario que, por tanto, no constituye sino la huella textual de un mundo referencial para el sujeto; de alguna forma, el léxico empleado es un reflejo del mundo semántico donde el emisor se sitúa para hablar. El análisis estadístico, si bien resulta limitado para explicitar con detalle el sentido de un texto, permite elaborar una «cartografía» de los mundos léxicos elegidos por el emisor para expresarse y, por tanto, de los sistemas de referencia desde los cuales construye su forma de ver la realidad (Reinert, 1991; 1992b).

El estudio de la distribución del vocabulario a lo largo de los enunciados elementales en que puede ser segmentado el texto, permite reconstruir los sistemas de referencia desde los cuales se emitió el discurso. Operativamente, el estudio de la distribución del vocabulario se realiza a partir de una *tabla de presencia/ausencia* en la que las columnas corresponden a los vocablos empleados y las filas a los enunciados diferenciables en el texto. La presencia o ausencia de un vocablo en un enunciado se traduce en los valores 1 ó 0 respectivamente para la celda intersección de la fila y la columna consideradas. El estudio estadístico se dirige a clasificar los enunciados (filas), agrupando aquéllos que más se aproximan entre sí de acuerdo con el vocabulario empleado (secuencia de unos y ceros para la fila). Las *clases* resultantes se distinguen por el empleo específico de cierto tipo de vocablos, que nos remiten a los sistemas de significados desde loss que se situó el sujeto para hablar. Del mismo modo, determinados enunciados pueden ser destacados como enunciados característicos de cada clase, ilustrando el discurso producido desde tales sistemas de significados.

Se trata, por tanto, de un *enfoque puramente formal*, basado en la distribución del vocabulario, que permite acceder a los mundos semánticos de referencia desde los cuales se construyó el discurso que analizamos.

LA METODOLOGÍA ALCESTE

La metodología de análisis emanada del planteamiento que acabamos de expresar puede ser desarrollada de forma automática gracias al programa ALCESTE,

1 La actual versión data de 1992.

disponible para Macintosh con procesador 68030 ó 68040. Este programa permite trabajar con textos de hasta 1 Mb (aproximadamente 20.000 líneas de 70 caracteres) vaciados en matrices de datos de hasta 4.000 líneas por 1.400 columnas.

El análisis estadístico siguiendo la metodología ALCESTE consta de tres etapas que implícitamente han sido señaladas en el apartado anterior: la construcción de la matriz de datos, la clasificación de las unidades de contexto (enunciados) y la descripción de las clases. El examen de cada una de ellas será ilustrado con la aplicación de esta metodología a los discursos producidos por seis grupos de profesores reunidos para conversar acerca de la reforma educativa. Este corpus de datos fue generado en el marco de una investigación reciente (Gil Flores, 1992). Sin entrar a pormenorizar los detalles de ésta, nos limitaremos a utilizar los datos para ejemplificar el proceso de análisis estadístico de textos de acuerdo con la metodología ALCESTE.

Nos centraremos en el análisis de los textos tomando como unidad el vocablo, aunque también es posible trabajar con unidades constituidas por parejas (secuencias de dos palabras) o segmentos (secuencias de varias palabras) que se repiten en el texto por encima de un determinado nivel de frecuencias.

Construcción de la matriz de datos

La primera operación realizada, previa a la aplicación de las técnicas estadísticas, va dirigida a construir la matriz de datos sobre la que se realizará el análisis. Ello comporta diferenciar los elementos que aparecerán en filas y en columnas. Las filas de esta matriz corresponden a las denominadas *unidades de contexto* —fragmentos de texto resultantes de la segmentación del corpus—. Las palabras presentes en una misma unidad de contexto estarían aludiendo a significados relacionados en el discurso, y por tanto a objetos de un mismo sistema semántico de referencia.

Para fragmentar el corpus textual en unidades de contexto, comenzamos por distinguir una serie de unidades impuestas por la propia naturaleza del texto. En nuestro caso, las *unidades de contexto iniciales* son los seis textos producidos tras la transcripción de los discursos generados por otros tantos grupos de profesores. Tales unidades son identificadas introduciendo al comienzo de cada texto determinados caracteres de identificación y las palabras explicativas que consideremos oportunas. Unas y otras irán precedidas del símbolo * (ver Tabla 1). La segmentación del texto se realiza de forma automática, respetando las unidades de contexto iniciales, es decir, se considera que éstas constituyen una primera división a partir de la cual es necesario proseguir la fragmentación del texto hasta descender a unidades elementales. La segmentación automática procede separando frases en función de los signos de puntuación y de un criterio de longitud máxima de las unidades resultantes. A partir de las *unidades de contexto elementales* (u.c.e.), surgidas de esta operación, son construidas las unidades de contexto que constituirán las filas de la matriz de datos. Como explicaremos más adelante, estas unidades

tendrán un tamaño máximo, en número de «formas analizadas», impuesto por el analista.

Las 8 primeras u.c.e. diferenciadas en el texto aparecen en la Tabla 1. Al mostrarnos los resultados del proceso automático de reducción a unidades, el programa suprime los acentos y mayúsculas.

Tabla 1
FRAGMENTACIÓN DE LAS PRIMERAS LÍNEAS DEL TEXTO EN UNIDADES DE CONTEXTO ELEMENTALES

```
* mayores
1 yo estoy harta de asistir a asambleas, encuentros, seminarios, y
1 entonces resulta que en todos ocurre lo mismo.
2 yo asistia el otro dia, en psicologia, a un seminario sobre metodos
2 audiovisuales.
3 y entonces resulta que se hacia la presentacion de lo que se esta
3 haciendo en la universidad.
4 nos hablaban de camara de video, retroproyector, cifras, porcentajes
4 obtenidos en la investigacion.
5 entonces yo estoy bastante esceptica, porque veo que lo que se esta
5 investigando se puede aplicar poco en la escuela.
6 yo no se si tu conoces el poligono sur.
7 decimos la reforma, pero bueno hay que empezar por ver la problematica
7 que tenemos en cada zona,
8 y despues ver como adaptar el curriculum al poligono sur.
(...)
```

La palabra «mayores» se introdujo como palabra explicativa al dividir el corpus global de datos en unidades de contexto iniciales, indicando así que esté texto correspondía al grupo de profesores que por tener una edad superior a los cuarenta y cinco años contaban con una dilatada experiencia profesional. Las palabras explicativas son características de todas las unidades de contenido elementales incluidas en una misma unidad de contexto inicial. Además, permiten definir clases de unidades de contexto a priori con objeto de realizar descripciones parciales.

Las columnas de la matriz de datos corresponden a los elementos del vocabulario. Utilizando el ordenador, podemos distinguir sin dificultad las *formas simples*, o secuencia de caracteres separados por un espacio o signo de puntuación. Sin embargo, de acuerdo con el enfoque asumido, interesa el modo en que se organiza la estructura semántica del texto; de ahí que sea preferible conservar aquellas formas que poseen un significado pleno y prescindir de las que soportan una carga semántica de segundo orden. Por ello, una vez delimitadas las formas simples, el ordenador procede a identificar las palabras funcionales (proposiciones, artículos, conjun-

ciones, pronombres y verbos auxiliares) con ayuda de un diccionario[2]. Estas formas son consideradas *ilustrativas* y servirán únicamente para describir las clases que obtengamos tras el análisis. Las palabras utilizadas en la descripción de cada unidad de contexto inicial (en nuestro caso, «mayores» era una palabra de este tipo) son también formas ilustrativas.

A continuación, las formas simples no identificadas como funcionales, es decir las que poseen un sentido pleno (sustantivos, adjetivos, verbos y ciertos adverbios) han de ser agrupadas, de forma que desaparezcan las diferencias no relevantes de significado, agrupando bajo una misma palabra las distintas formas provocadas por las marcas de género y número o por las desinencias de conjugación. Se trata, en definitiva, de un proceso de «lematización», siguiendo el término empleado habitualmente para nombrar la operación consistente en reducir varias palabras con un mismo significado a una sola forma común.

Los posicionamientos a favor o en contra de la denominada lematización constituyen uno de los elementos que han caracterizado la reflexión metodológica en el ámbito de la estadística textual. En la metodología que presentamos en este trabajo, lo importante son los mundos semánticos a los que nos remiten las palabras empleadas en un texto, de ahí que no tenga un excesivo interés, por ejemplo, conservar por separado las formas *profesor* y *profesora* que aluden a un mismo significado. Podemos afirmar que aquí la lematización previa al análisis resulta coherente con los planteamientos de partida. El problema de realizar una lematización del texto se encuentra en la dificultad que ello supone desde el punto de vista del tratamiento automático. Mientras las formas simples son inmediatamente identificables por un ordenador dado su carácter físico (secuencia de caracteres no delimitadores comprendidos entre dos caracteres delimitadores), la localización de las formas que denotan significados similares es más compleja. Algunos autores proponen mecanismos que permiten la lematización automática, definiendo las reglas de equivalencia entre las formas gráficas y los «lemas», y pueden lograr la separación automática del texto en unidades utilizando diccionarios de raíces y de sufijos. No obstante, en su defensa de la lematización, presentada en el prólogo de la obra de Lafon (1984), Muller reconoce que no existe consenso entre los lematizadores sobre las reglas que deben seguirse, y que cualquier intento de lematización automática resulta necesariamente parcial.

La vía seguida en la metodología ALCESTE para llevar a cabo la lematización automática del texto consiste en la utilización de un diccionario de raíces (caso de los verbos irregulares), con ayuda del cual pueden ser identificadas las formas verbales que perteneciendo a un mismo verbo presentan distinta raíz, o bien un diccionario de sufijos, que permite detectar palabras diferenciadas únicamente en las marcas de género, número o en las desinencias verbales. El resultado final del proceso de lematización es un repertorio de *formas reducidas*, del cual presentamos

2 La versión actual de ALCESTE incorpora diccionarios adaptados a la lengua castellana.

un extracto correspondiente al texto que analizamos (Tabla 2). Así, por ejemplo, las formas iniciales «abierto» y «abiertos» han sido reagrupadas bajo la forma reducida «abierto+».

Tabla 2
REPERTORIO DE FORMAS REDUCIDAS CORRESPONDIENTE A LOS DISCURSOS SOBRE LA REFORMA

F. reducida	F. inicial
abajo	abajo
abierto+	abierto
abierto+	abiertos
abogado+	abogado
abogado+	abogados
absoluto	absoluto
absurdo	absurdo
aca	aca
acaba+	acaba
acaba+	acabas
acceder	acceder
actividad+	actividad
actividad+	actividades

Frente a las formas ilustrativas, no consideradas para los análisis, las formas reducidas serán las utilizadas para realizar la clasificación de las unidades de contexto y reciben la denominación de *formas analizables*.

Una vez segmentado el corpus en u.c.e. e identificadas las formas reducidas, construimos la matriz de datos, en la que aparece recogida la presencia/ausencia de las formas analizables en determinadas unidades de contexto. Estas unidades de contexto están integradas por u.c.e., abarcando siempre un número entero de ellas.

Para comprobar la consistencia de los resultados a los que lleguemos en el análisis, se realizan dos clasificaciones paralelas: consideramos matrices de datos diferentes, construidas a partir de unidades de contexto de distinto tamaño, medido en número mínimo de formas analizables. En este caso, hemos construido dos matrices tomando unidades de contexto con un número mínimo de 16 y 14 formas analizables respectivamente. Tales matrices de datos tendrán necesariamente un porcentaje alto de ceros, toda vez que el número de formas presentes en una unidad de contexto es muy inferior al número total de formas analizadas. La fila correspondiente a una unidad de contexto tendrá un número pequeño de unos y el resto serán ceros.

Clasificación de las unidades de contexto

La clasificación de las unidades de contexto se realiza siguiendo un algoritmo de clasificación jerárquica descendente creado específicamente para tratar este tipo de datos —matrices de pocos efectivos y grandes dimensiones— (Reinert, 1985). Puesto que las unidades tomadas aisladamente tienen una baja probabilidad de contener formas analizadas comunes, se considera preferible utilizar un algoritmo descendente en lugar de uno ascendente. Describiremos en líneas generales este algoritmo.

Siendo I el conjunto de unidades de contexto (filas de la matriz) y J el conjunto de formas reducidas (columnas), llamamos k_{ij} al valor presente (1 ó 0) en la intersección de la fila i y la columna j de la tabla IxJ.

El problema que se plantea no es sino encontrar una partición $[I_1, I_2]$ de I que dé lugar a dos clases bien diferenciadas, maximizando el valor chi-cuadrado de la tabla de contingencia resultante al condensar en dos —una por cada clase de la partición— las filas de la tabla IxJ. Es decir, la tabla que aparece en la Figura 1.

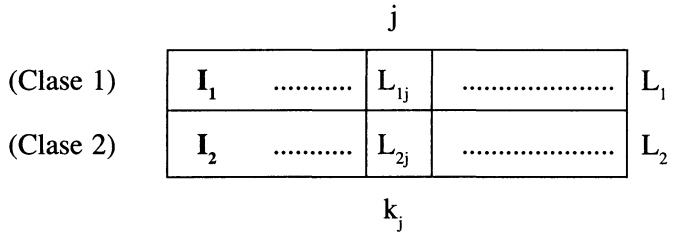


Figura 1
TABLA DE CONTINGENCIA CONDENSADA TRAS UNA PARTICIÓN EN DOS CLASES

En esta tabla, el valor L_{1j}, correspondiente a la celdilla intersección de la partición I_1 con la forma analizada j, es el número de unidades de contexto de la clase I_1 en que la forma j está presente.

$$L_{1j} = \sum_{i \in I_1} k_{ij} \; ; \quad L_1 = \sum_{j=1}^{J} L_{1j} \; ; \quad K = L_1 + L_2$$

$$L_{2j} = \sum_{i \in I_2} k_{ij} \; ; \quad L_2 = \sum_{j=1}^{J} L_{2j} \; ; \quad K_j = L_{1j} + L_{2j}$$

El valor de chi-cuadrado obtenido mediante la fórmula

$$\chi^2 = L_1 L_2 \sum_{j=1}^{J} (L_{1j}/L_1 - L_{2j}/L_2)^2 / k_j$$

permite llevar a cabo un contraste entre los perfiles de ambas clases. Bastará, por tanto, encontrar la partición $[I_1, I_2]$ que maximiza el valor de χ^2. La técnica empleada para ello no garantiza la obtención del máximo valor de χ^2 pero sí una buena aproximación a éste. El algoritmo seguido opera del siguiente modo:

— Se calcula el primer eje factorial de la nube de puntos N(I) (análisis factorial de correspondencias de la tabla IxJ) en el espacio R^J, dotado de la métrica de chi-cuadrado.

— Se busca el hiperplano perpendicular al primer eje factorial, que separa la N(I) en dos subnubes de puntos $N(I_1)$ y $N(I_2)$ de forma que la inercia interclases sea máxima. Este valor máximo resulta ser casi igual que el valor de χ^2 asociado a la tabla de contingencia condensada.

— Puesto que los centros de gravedad de las dos nubes $N(I_1)$ y $N(I_2)$ construidas de este modo no se sitúan exactamente sobre el primer eje factorial, la inercia interclases puede ser aumentada mediante un proceso iterativo de intercambio de puntos entre las dos nubes, que nos permite aproximarnos aún más al valor de χ^2 buscado. Este proceso consiste en comprobar si el cambio de clase de cada punto i I aumenta o disminuye la inercia interclases, cambiándolo de clase en caso positivo.

Una vez encontrada la partición $[I_1, I_2]$, se inicia de nuevo el proceso tratando de dividir en dos la mayor de las clases de unidades de contexto resultantes. De este modo, se origina una sucesión de análisis que culminan al ser alcanzado un número de clases (terminales) previamente fijado por el analista. Los primeros pasos de este proceso podrían ser representados por la Figura 2.

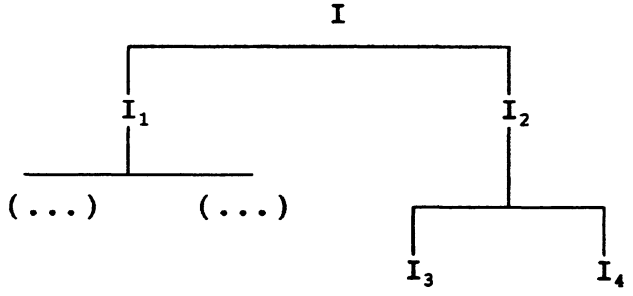


Figura 2
PROCESO ESQUEMÁTICO DE LA CLASIFICACIÓN JERÁRQUICA DESCENDENTE

Desarrollado este procedimiento de clasificación, se plantea el problema de decidir qué clases consideraremos como resultado final de la clasificación; problema que se resuelve en la metodología ALCESTE, garantizando la estabilidad de los resultados mediante la comparación de dos clasificaciones obtenidas sobre el mismo corpus textual. Ambas clasificaciones se llevan a cabo considerando unidades

de contexto de distinto tamaño. En este caso, construíamos dos matrices de datos para unidades de contexto de 14 y 16 formas analizadas respectivamente. Las dos clasificaciones obtenidas para diez elementos terminales eran las mostradas en la Figura 3.

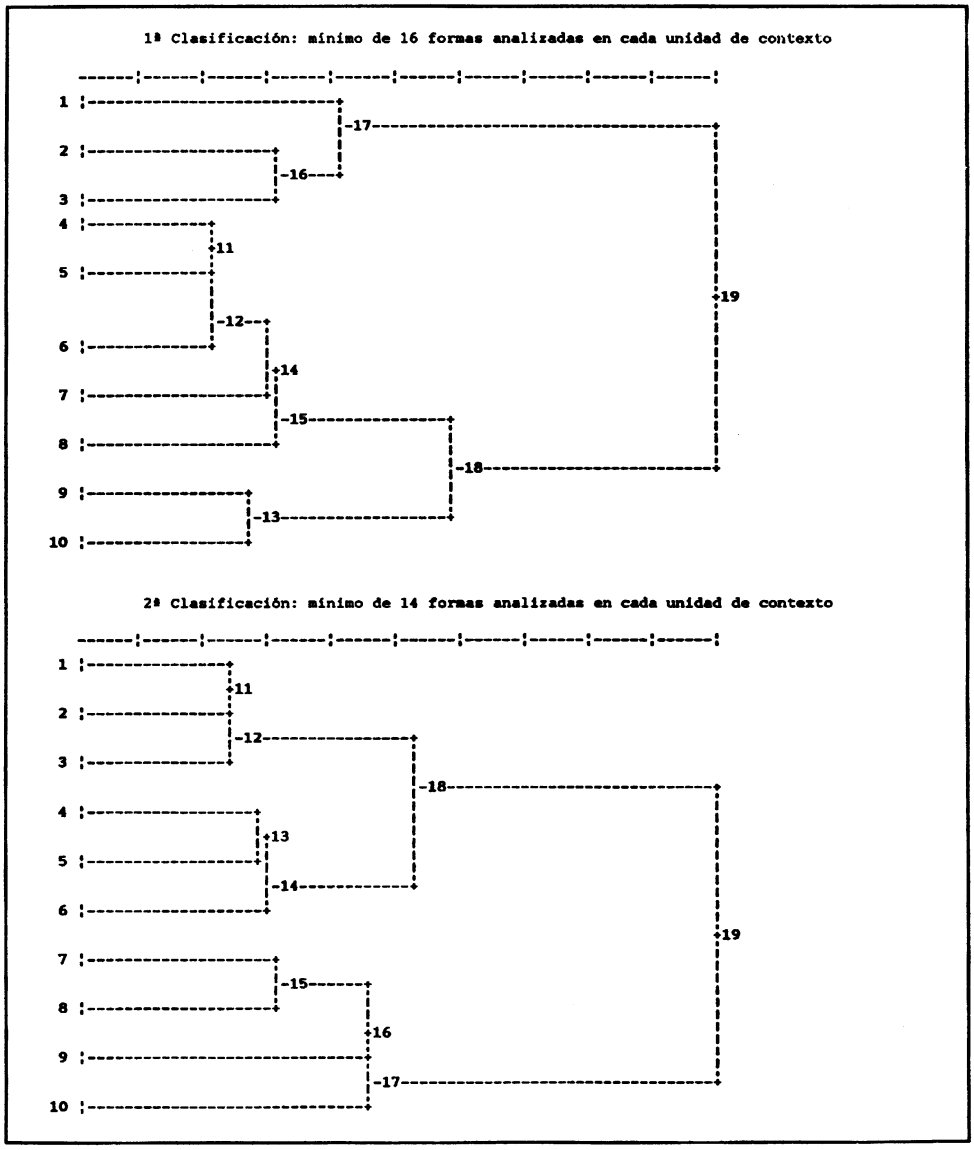


Figura 3
DENDOGRAMAS PARA LAS DOS CLASIFICACIONES

Puesto que hemos pedido al programa ALCESTE que nos ofrezca 10 clases terminales, aparecen otras 9 no terminales, numeradas desde el 11 al 19. Es decir, a pesar de que el proceso de clasificación es descendente, las clases aparecen numeradas en sentido ascendente. La clase 19, que comprende todas todas las unidades objeto de clasificación, se divide en otras dos, las números 17 y 18, y éstas a su vez en otras dos. El proceso de subdivisión jerárquica continúa hasta alcanzar los 10 elementos terminales.

La comparación de las clasificaciones para determinar las clases de mayor estabilidad se realiza en función del número de u.c.e. incluidas en cada clase (recuérdese que las unidades de contexto se componen de unidades de contexto elementales). El procedimiento seguido consiste en comparar cada una de las clases de la primera clasificación con todas las clases de la segunda (sean terminales o no), utilizando como medida de asociación el valor de chi-cuadrado. A partir de estos cálculos, se retiene toda pareja de clases (I_L, I_H) en la que el valor de chi-cuadrado es mayor que el encontrado para cualquier otra pareja en la que toman parte alguna de las clases I_L o I_H, lo cual significa que ambas se asocian más entre sí que a cualquier otra clase de la jerarquía.

La comparación entre dos clases I_L e I_H, correspondientes a la primera y segunda clasificación respectivamente se haría a partir de la tabla de contingencia presentada en la Figura 4, donde los efectivos no indicados son calculables mediante diferencias.

	I_H	$I - I_H$	
I_L	n_{12}	-	n_1
$I - I_L$	-	-	-
	n_2	-	n

Figura 4

TABLA DE CONTINGENCIA PARA EL CÁLCULO DEL VALOR χ^2 DE ASOCIACIÓN ENTRE CLASES

En esta tabla, n_1 es el número de u.c.e. incluidas en la clase I_L tras la primera clasificación, n_2 el número de u.c.e. incluidas en la clase I_H correspondiente a la segunda clasificación, n_{12} el número de u.c.e. presentes simultáneamente en la clase I_L de la primera clasificación y la clase I_H de la segunda, y n es el total de u.c.e. clasificadas.

En el caso del texto que analizamos, el valor de chi-cuadrado para las parejas de clases en las que éste era máximo aparece en la Tabla 3. La presencia en esta tabla, por ejemplo, de la pareja 1 <-> 9 significa que la clase 1 del primer análisis (mínimo de 16 formas analizadas en cada unidad de contexto), que cuenta con 403 u.c.e. se corresponde con la clase 9 del segundo análisis (mínimo de 14 formas analizadas), en la que quedaron incluidas un total de 595 u.c.e. El total de u.c.e. presentes en ambas clases asciende a 265, lo que implica un χ^2 de asociación de valor 1024, con un grado de libertad.

Tabla 3
MÁXIMA CORRESPONDENCIA ENTRE CLASES PARA LAS DOS CLASIFICACIONES

1ª Cla. <->	2ª Cla.	*	frec1	frec2	frec12	chi2	*
1 <->	9	*	403	595	265	1024	*
4 <->	6	*	16	55	15	1104	*
8 <->	10	*	70	31	17	560	*
13 <->	11	*	1215	1048	741	1262	*
14 <->	14	*	1279	1273	895	1459	*
16 <->	15	*	1365	1354	948	1371	*
17 <->	16	*	1779	1967	1525	2006	*
18 <->	18	*	2587	2368	2123	1980	*

Consideramos una *clase estable* aquélla formada por las u.c.e. presentes simultáneamente en las dos clases que forman parte de una pareja de máxima asociación. Para determinar las clases estables resultantes del proceso de análisis descrito, elegimos entre todas las parejas de clases retenidas aquéllas que pertenecen, al menos en una de las clasificaciones, a la misma partición. De este modo conseguimos que las unidades seleccionadas pertenezcan a una y sólo a una de las clases estables resultantes. Entre las particiones posibles se encuentran, por ejemplo, la partición en dos clases constituida por las parejas 17<->16 y 18<->18, la formada por tres clases resultantes de las intersecciones 1<->9, 16<->15 y 18<->18 ó la partición en cuatro clases 1<->9, 13<->11, 14<->14 y 16<->15. Siguiendo el doble criterio de retener el mayor número de clases sin bajar de una frecuencia mínima de 100 u.c.e. clasificadas en cada una de ellas, ALCESTE incorpora un algoritmo que permite seleccionar de forma automática la partición óptima. En este caso, la partición propuesta es una formada por 4 clases (Tabla 4).

Tabla 4
PARTICIÓN ÓPTIMA EN CLASES ESTABLES

1ª clase : intersección	13	y	11	; número de u.c.e. : 741
2ª clase : intersección	14	y	14	; número de u.c.e. : 895
3ª clase : intersección	16	y	15	; número de u.c.e. : 948
4ª clase : intersección	1	y	9	; número de u.c.e. : 265

La suma de los efectivos de las cuatro clases alcanza la cifra de 2.849 u.c.e. Este número de unidades contenidas en las clases puede ser expresado en relación a las 4.410 presentes en el corpus. De este modo, se obtiene un porcentaje del 64.6% de u.c.e. bien clasificadas.

Interpretación de las clases

Las clases pueden ser descritas a partir de las formas reducidas (bien se trate de formas analizadas o de formas ilustrativas) más características presentes en las u.c.e. incluidas en ellas. El procedimiento consiste en calcular un coeficiente de asociación de una forma a una clase: chi-cuadrado calculado a partir de una tabla de contingencia construida para cada forma, cruzando la presencia o ausencia de la palabra analizada en las u.c.e. y la pertenencia o no de las u.c.e. a la clase en cuestión. La tabla presentada en la Figura 5 serviría de base para el cálculo de la asociación de la forma F_a a la clase I_L.

	F_a presente	F_a ausente	
I_L	n_{12}	–	n_1
$I - I_L$	–	–	–
	n_2	–	n

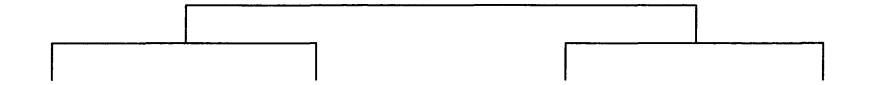
Figura 5
TABLA DE CONTINGENCIA PARA EL CÁLCULO DEL VALOR χ^2 DE ASOCIACIÓN DE UNA FORMA A UNA CLASE

En esta tabla de contingencia, n_1 es el número de u.c.e. incluidas en la clase, n_2 el número de u.c.e. en las que está presente la palabra considerada, n_{12} el número de u.c.e. de la clase que cuentan con la presencia de la palabra, y n el número total de u.c.e. clasificadas. El valor n_{12} es comparado al valor teórico n1.n2/n al calcular chi-cuadrado para la tabla anterior. Al valor obtenido se añade el signo de la diferencia n_{12}-(n1.n2/n) con el fin de caracterizar la clase por la presencia o, por el contrario, la ausencia de la palabra en cuestión.

Los vocablos presentes que resultan específicos para cada una de las cuatro clases diferenciadas en el corpus sobre la reforma educativa, seleccionados teniendo como criterio un valor de chi-cuadrado superior a 20 —criterio fijado por el analista con el fin de reducir el número de términos específicos— y un signo positivo en la diferencia n_{12}-(n1.n2/n) quedan recogidos en la Figura 6.

Interpretar las palabras características de cada clase se reduciría, según Reinert (1992b), a destacar las palabras semánticamente próximas, que connotan un mismo concepto, para reunirlas en conjuntos de conceptos que permitan una interpretación global de cada una de las clases.

Las u.c.e. características de cada una de las clases, pueden ser también útiles instrumentos de cara a realizar la descripción de éstas. De nuevo, se toma como criterio el valor de chi-cuadrado, aplicado en esta ocasión a una tabla de contingencia donde se cruzan la pertenencia o no de las palabras al conjunto de formas características de una clase I_L (conjunto A) y la presencia o ausencia entre las formas de la u.c.e. considerada (u_i). Si tomamos una u.c.e. denotada por u_i, la tabla



CLASE 1	CLASE 2	CLASE 3	CLASE 4
actividad+	acuerdo+	aula+	apoyo+
alumno+	administracion	año+	año+
calidad	aqui	casa+	ciclo+
cambiar	cursillo	caso+	curso+
cambio	den	cinco	departamento
cierto+	dicho	clase+	dijo
contenido+	estoy	colegio+	dos
creo	general+	conozco	escribir
cuestion	habeis	crio+	especialista+
ello	hecho+	cuarenta	especial
enseñanza	igual+	cuarto+	etapa
enseñar	ilusion	cuatro	gimnasia
escuela	informacion	cuidador+	inicial
formacion	me	dice+	instituto
forma+	mi	dieciocho	integracion
fracaso+	mis	ejemplo+	junto+
futuro+	necesaria	guarderia+	leer
grande+	parece	inspector+	llegan
habria	perfeccionamiento	llega+	medio+
importante+	polígono	llevan	niño+
ley	poquito	madre+	obligatoria
logicamente	privada	material+	pasan
mentalidad	proyecto	meter	preescolar
muy	quiero	mire	primero+
negativo+	reforma	niño+	primer+
palasi	sexenios	os	profesor+
pienso	sido	padre+	programacion
profesional+	siento	pueblo+	quince
profesorado	sur	ratio	segunda
punto+	tengo	siete	segundo
realidad+	turno+	te	seis
realmente	vamos	ti	sexto
resultado+	ver	tienes	siguiente
ruptura	yo	treinta	superior+
seria		tres	tendrian
sistema		tu	viene+
sociedad		usted	
solamente		vas	
tipo		veinticinco	
valor+			
villar			
vista			

Figura 6
VOCABULARIO ESPECÍFICO DE LAS CLASES

	Formas de u_i	Formas de $(I-u_i)$	
A	n_{12}	-	n_1
A'	-	-	-
	n_2	-	n

Figura 7
TABLA DE CONTINGENCIA PARA EL CÁLCULO DEL VALOR χ^2 DE ASOCIACIÓN DE UNA U.C.E. A UNA CLASE

que permite el cálculo de la asociación entre esta unidad y la clase I_L sería la que aparece en la Figura 7.

En la tabla de la Figura 7, donde n representa el número total de ocurrencias en el corpus, n_1 es la frecuencia total del conjunto de palabras características de una determinada clase, n_2 la frecuencia total de las palabras contenidas en la unidad de contexto u_i, y n_{12} la frecuencia total de aparición de las palabras características de la clase (pertenecientes al conjunto A) que están presentes en el enunciado u_i. I representa el conjunto de u.c.e. presentes en el corpus global, y con la notación A' hemos expresado el conjunto complementario de A, es decir, el conjunto de palabras no específicas de la clase I_L. Como en tablas anteriores, los efectivos no consignados se obtienen por diferencia entre los valores de la tabla, dado que ésta posee sólo un grado de libertad.

Ordenadas según valores crecientes de χ^2, las u.c.e. más representativas de cada clase son las que recogemos en la Tabla 5. A la izquierda de cada unidad aparece el número de orden que le corresponde dentro del corpus y el valor chi-cuadrado, para un grado de libertad, de asociación a la clase.

A la luz del vocabulario y las u.c.e. características de cada clase, es posible llevar a cabo una interpretación sobre el modo en que los profesores opinan acerca de la reforma educativa. Fundamentalmente, podemos distinguir cuatro espacios de significado desde los cuales se emite el discurso, que se corresponden con otras tantas clases identificadas en el análisis.

En el caso de la primera clase, identificamos distintos «campos léxicos» relacionados con una conceptualización de la enseñanza de que son objeto los alumnos. Así, aparecen alusiones generales en formas tales como *sistema, sociedad, enseñanza, enseñar, escuela, profesorado*; referencias al cambio que en ella se propugna: *cambiar, cambio, ley, ruptura*; o a los aspectos sobre los que éste incidiría: *calidad, fracaso, formación, futuro, realidad, resultado, negativo*. La segunda clase incluye una serie de enunciados que se caracterizan por haber tenido su origen en lugares semánticos relacionados con el perfeccionamiento del profesorado: *cursillo, informacion, perfeccionamiento, proyecto, sexenios*. Es un discurso construido desde la primera persona: *me, mi, mis, yo, estoy, quiero, siento, tengo, vamos*. Desde este campo léxico se han construido enunciados característicos de esta clase, entre los cuales, el que alcanza un valor de chi-cuadrado más alto (χ^2=25) refuerza con

CLASE NÚMERO : 1

2069	29	Y vamos a seguir, a pesar de todo, digo, que a pesar de la ley Villar Palasí, y a pesar de esto, la escuela va a seguir adelante.
1736	24	Y qué se entiende por fracaso, un chico que no da la talla, en los contenidos, un chico que tiene realmente problemas psicoconductuales.
1680	21	solamente que la ley, la teoría está bien hecha, porque además está bastante elaborada.
1730	21	ése es muy difícil de que a la larga se pueda sentir integrado en la marcha, en el mecanismo de la responsabilidad cotidiana de la formación.
1897	21	el alumno capta no solamente los contenidos sino que capta la formación del maestro, en qué valores, cómo se toma la responsabilidad.

CLASE NÚMERO : 2

1462	25	es que si yo tengo que hacer este cursillo, y tengo que sacarlo de mi tiempo personal, y encima me los tengo que pagar yo si son privados.
237	21	pero yo estoy diciendo lo que siento; yo estoy poniendo aquí, vamos, el corazón en la mesa.
68	17	bueno yo estoy diciendo de nuestra zona. yo digo que la reforma la seguiremos intentando aplicar como sea.
89	17	pero, ¿qué reforma?, ¿de qué reforma hablamos?, porque yo no sé de qué reforma hablamos después del tiempo que estamos.
486	17	yo lo he visto, y lo tengo en fotocopias. si yo sé que iba a ser la reforma me hubiera traído documentación. llama la atención.

CLASE NÚMERO : 3

437	19	pero si no quieren pues por antigüedad en el centro, pues se van a tercero, se van a cuarto y se van a quinto. el problema es gordísimo.
1044	19	ya el descontento que tú tienes, porque ves que los padres también intervienen, y te dicen lo que tú tienes que hacer.
2278	19	que no, que tenemos que tener unas condiciones. te falta material en clase, no tienes material, no tienes mobiliario.
3815	19	lo que más conozco es eso, que nos van a meter a niños de tres años y, según dicen nos los van a mezclar con cuatro sin quitarnos.
1089	18	después, el tema que antes has tocado tú de que sí, que entran niños escolarizados de tres años, pero esos niños con qué edad terminan.

CLASE NÚMERO : 4

465	67	luego viene la ESO, la secundaria obligatoria, primer ciclo de secundaria obligatoria, dos años, segundo ciclo otros dos años.
1824	63	en los institutos a los de EGB, los de EGB de la segunda etapa a los de ciclo inicial y ciclo medio.
257	54	porque el verdadero escollo que nos encontramos en la segunda etapa es cuando llegan los niños y no saben leer ni escribir.
133	46	que fue en el setenta y cinco setenta y seis, el primer curso, me entraron un curso de primero,
501	46	es que no lo sé. quién va a dar el primer ciclo, quién el segundo, los licenciados, los profesores de la segunda etapa.

Tabla 5
UNIDADES DE CONTEXTO ELEMENTALES MÁS CARACTERÍSTICAS DE CADA CLASE

claridad la interpretación que hacemos de ella. El vocabulario específico de la clase tercera denota un discurso emitido teniendo como referencia el contexto donde tiene lugar la enseñanza. Así, aparecen vocablos que expresan lugares: *aula, casa, clase, colegio, guardería, pueblo, alli*; cantidades: *cuarenta, cuatro, dieciocho, ratio, siete, treinta, tres, veinticinco*; o evocan condiciones en que se desenvuelve la docencia: *cuidador, material.* En este caso, el emisor trata de poner en situación a la persona que escucha utilizando la segunda persona: *te, ti, os, tienes, tu, usted, vas.* Por último, los vocablos característicos de la cuarta clase remiten a un contexto de tipo institucional, que tiene que ver con la estructuración del sistema educativo: *ciclo+, curso+, etapa, inicial, medio+, obligatoria, preescolar, primero, primer, segunda, segundo, sexto, superior+*, o con la organización de la enseñanza: *departamento, especial, integracion, programacion.* En cualquier caso, los enunciados característicos recogidos en la Tabla 5 ejemplifican las peculiaridades del discurso producido desde estas posiciones.

Si examinamos el árbol resultante de la clasificación (Figura 3), podemos estructurar estas cuatro clases en otras dos, de modo que la primera de ellas quedaría constitudida por las clases 1 y 2, mientras que la otra abarcaría las clases 3 y 4. Ambas podrían ser identificadas como campos léxicos que hacen referencia respectivamente a la enseñanza y al contexto en que ésta se produce.

CONCLUSIÓN

La metodología presentada permite destacar determinadas características formales de los textos que pueden ser interpretadas desde un enfoque estadístico. El análisis de la distribución del vocabulario en un corpus nos ha permitido detectar grupos de palabras que no son sino la huella textual de los escenarios semánticos desde los que los profesores opinan sobre la reforma. El análisis realizado, de carácter objetivo, no tiene en cuenta el sentido de lo que los sujetos han expresado, ni precisa conocer la intención o la situación en la que fueron producidos los discursos. No consiste sino en realizar otra lectura del texto, otra decodificación, violando la linealidad para dar una visión sintética (Reinert, 1991).

La exploración de los «mundos léxicos» —terminología usada por Reinert— nos permite acceder a la representación estructurada que subyace a ellos. En el caso particular en que hemos aplicado esta metodología, ha contribuido a identificar los tipos de «mundos» desde los que los profesores se expresan, indicando los aspectos desde los que la reforma puede ser valorada por el profesorado: la enseñanza en sí misma, y el contexto en que ésta se produce. La organización detectada permite afirmar que los significados relativos a la enseñanza se estructuran teniendo como referencia la figura del alumno o la figura del profesor, mientras que el contexto es contemplado desde un punto de vista físico-material o estructural-institucional.

Tales temas serían objeto de atención en nuevas aproximaciones al contenido de los discursos que pudieran llevarse a cabo siguiendo métodos basados no ya en las

características formales, sino en enfoques puramente semánticos e interpretativos. De alguna forma, el mapa que serviría de guía para el análisis de las opiniones del profesorado acerca de la reforma educativa queda trazado.

REFERENCIAS BIBLIOGRÁFICAS

BENZÉCRI, J. P. (1973): *L'Analyse des Données*. Paris: Dunod.

BOLASCO, S. (1993): Sur différentes stratégies dans une analyse des formes textuelles: une expérimentation à partir de données d'enquête. En Bécue, M.; Lebart, L. y Rajadell, N. *Jornades Internacionals d'Anàlisi de Dades Textuals* (pp. 69-88). Barcelona: Servicio de Publicaciones de la UPC.

GIL FLORES, J. (1992): *Análisis de Datos Cualitativos. Aplicación al caso de Datos procedentes de Grupos de Discusión*. Tesis Doctoral inédita. Universidad de Sevilla.

GUIRAUD, P. (1960): *Problèmes et méthodes de la statistique linguistique*. París: PUF.

IKER (1975): *Words, system manual*. Rochester, NY: Computer Print.

LAFON, P. (1984): *Dépouillements et statistiques en lexicométrie*. París: Slatkine-Champion.

LEBART, L. y SALEM, A. (1988): *Analyse Statistique des Données Textuelles. Questions ouvertes et Lexicométrie*. París, Bordas.

LEBART, L.; MORINEAU, A. y BÉCUE, M. (1989): *SPAD-T. Système portable pour l'analyse des données textuelles. Manuel de l'utilisateur*. París: CISIA.

MOCHMANN (1985): Análisis de Contenido mediante Ordenador Aplicado a las Ciencias Sociales. *Revista Internacional de Sociología, 43* (1), 11-44.

MULLER, C. (1968): *Initiation à la statistique linguistique*. París: Larousse.

REINERT, M. (1985): Classification descendente hiérarchique: un algorithme pour le traitement des tableaux logiques de grandes dimensions. Comunicación a las *Quatrièmes Journées Internationales «Analyse des Données et Informatique»*. Versailles.

REINERT, M. (1986): Un logiciel d'analyse lexical (ALCESTE). *Les Cahiers de l'Analyse des Données, XI (4)*, 471-484.

REINERT, M. (1991): La méthodologie d'analyse des données textuelles ALCESTE; application à l'analyse des poésies d'A. Rimbaud. En Maurand, G. (Ed.). *Poésie et Modernité. Colloques d'Albi Langages et Signification* (pp. 303-325). Toulouse: Université de Toulouse-le-Mirail.

REINERT, M. (1992a): *Notice du logiciel ALCESTE, version 2.0*. Toulouse.

REINERT, M. (1992b): La methodologie ALCESTE et l'analyse d'un corpus de 304 récits de cauchemars d'enfants. Comunicación presentada al *Convegno Internazionale Ricerca Qualitativa e Computer nelle Scienze Sociali*. Roma.

SALEM, A. (1990): *LEXICO1*. Ecole Normale de Fontenay-Saint Cloud.

STONE, P. J. et al., (1966): *The General Inquirer: a computer approach to content analysis*. Cambridge, MA: M.I.T. Press.

YULE, G. U. (1944): *A statistical study of vocabulary*. Cambridge: University Press.

ZIPF, G. K. (1946): *The psychobiology of language, an introduction to dynamic philology*. Boston: Houghton-Miffin.