Una Medida Cuantitativa de la Diversidad Estilística: Aplicación al Análisis de Textos Españoles[1]

Ignacio Moreno-Torres Sánchez
Universidad de Málaga — 29071 Málaga — España

Abstract

Our objective is to obtain an objective measure of the gramatical diversity of a text. We will use as a starting point a transition matrix (TM). A MT is a table which contains the frequencies of every pair of lexical categories (proper noun, infinitive, gerund, adjective, etc.) used consecutively in a text. We see a TM as a map of the gramatical resources used in a text. So we can hypothesise that different types of texts will produce different types of matrixes.

We propose two measures: *DivEst*, which measures the degree of variation in a text, and *DifEst*, which measures the distance between two texts. In order to prove the interest of these two measures, we will examine the values obtained for different types of texts (*journalistic, oral* and *narrative* texts). Furthermore we will show that there is a clear relation between *DivEst* and the traditional stylistic opposition *nominal/verbal* style maintained among others by Batjin (1990) and Nuñez Lavedeze (1991).

Resumen

El objetivo de este trabajo es proponer una medida objetiva de la diversidad gramatical de un texto. Para ello tomamos como punto de partida una matriz de transiciones (MT). Una MT recoge el número de veces que cada par de categorías léxicas (nombre propio, infinitivo, gerundio, adjetivo, etc.) ha aparecido en un texto o conjunto de textos. Por tanto podemos ver un MT como un mapa de los recursos gramaticales empleados en un conjunto de textos. Ello nos permite suponer que diferentes tipos de textos deben dar lugar a diferentes tipos de matrices.

Proponemos dos formas de medir la diversidad: *DivEst*, que mide el grado de variación del texto, y *DifEst*, que mide la distancia entre dos textos.

Para mostrar el interés de estas dos medidas examinamos los valores obtenidos para 3 tipos de textos (*periodísticos, orales y narrativos*). Además, mostramos que hay una relación clara entre el parámetro DivEst y la oposición entre el estilo *nominal/verbal* mantenida entre otros por Batjin (1990) y Nuñez Lavedeze (1991).

Palabras clave: statistics, stylistics, nominal style, verbal style, transition matrix

1. Introducción

En los años 80 se desarrolló en lingüística computacional (Garside et al. 1987) una técnica de desambiguación léxica basada en cadenas de markov. Al emplear esta técnica se crean unas matrices de transición (MT) que recogen las frecuencias o pesos con las que dos categorías léxicas cualesquiera aparecen seguidas en un corpus de aprendizaje. Después de trabajar con estas técnicas durante algunos años y desarrollar varios sistemas de etiquetado nos hemos planteado si los datos recogidos en las MT podrían servir para identificar tipos de textos o propiedades de éstos. En particular lo que motiva nuestro trabajo ha sido la necesidad de

[1] Para la realización de este trabajo nos hemos servido del entorno de lematización semiautomático Ayda (Albalá, Cappelli, Marrero y Moreno-Torres; 1996)

responder a preguntas como las siguientes: ¿En qué medida se diferencian dos matrices de un mismo autor, de un mismo dialecto, de una misma lengua?; ¿qué textos dan lugar a MT iguales y qué textos dan lugar a MT diferentes?, etc. Para poder responder a estas preguntas, necesitamos una forma de comparar las diferentes matrices: esto es necesitamos una función que mida las características de cada matriz.

Debe notarse que esta aproximación se diferencia claramente del análisis multidimensional de Biber (1995a, 1995b). Este autor selecciona un conjunto de fenómenos lingüísticos a partir de los conocimientos del propio investigador, mide sus frecuencias y, por último, estudia sus relaciones. En nuestro caso partimos del supuesto de que una MT recoge todos los fenómenos lingüísticos que pueden darse en una lengua. De tal forma que si un fenómeno no se produce encontraremos un cero en la matriz y si ocurre con frecuencia encontraremos un valor alto. Esto es, entendemos que una MT es un *mapa de los recursos lingüísticos* de una lengua.

El resto del artículo se organiza como sigue. En el apartado 2 describimos brevemente las MT y proponemos dos parámetros para cuantificar su diversidad y compararlas. En el apartado 3 mostramos los valores obtenidos al analizar un conjunto de textos españoles. Una vez confirmada la hipótesis de que podemos medir la variación, en el apartado 4 nos planteamos el problema de por qué son más simples algunas matrices; mostraremos cómo la diversidad/simplicidad de las matrices puede asociarse a la clásica oposición planteada entre otros por Batjin (1990) o Nuñez Lavedeze (1991) entre *estilo nominal/estilo verbal*.

2. Matrices de transición

El siguiente gráfico muestra una MT obtenida a partir de un texto etiquetado con las categorías: *Sustantivo, Verbo personal, Infinitivo, Gerundio, Adjetivo* (hemos eliminado otras categorías por falta de espacio)

	Sust.	Verbo	Infi.	Ger.
Sustantivo	6	24	6	0
Verbo personal	**46**	**0**	1	0
Infinitivo	0	41	3	0
Gerundio	5	12	3	0

En esta matriz de transiciones, el valor 46 en negrita indica que en los textos analizados aparecieron 46 verbos en forma personal seguidos de sustantivo, mientras que no apareció ningún verbo personal seguido de otro verbo personal.

2.1 Métrica de la diversidad estilística

Si atendemos a la distribución de los valores en las matrices, obtenemos dos tipos de matrices extremas: las *Matrices de homogeneidad máxima* (máxima diversidad) —donde las frecuencias de todas las secuencias coinciden—, y las *Matrices de heterogeneidad máxima* (mínima diversidad) —donde una sola secuencia se repite siempre.

Heterogeneidad Máxima

2	2
2	2

Homogeneidad Máxima

0	0
8	0

Podemos cuantificar la diversidad estilística mediante la siguiente fórmula:

$$DivEst=\sum^{i,j}(c_{ij})^2$$

Mediante esta fórmula las matrices en las que hay mayor homogeneidad (tienen frecuencias más grandes y más ceros) dan lugar a un valor mayor, mientras que las matrices heterogéneas (tienen frecuentes más bajas y menos ceros) dan lugar a un valor menor. Ahora bien, para evitar que el resultado sea sensible al número de casos procesados (la suma de las frecuencias), dividiremos cada frecuencia por el número total de casos. De esta forma el valor máximo teórico será siempre 1. La fórmula obtenida así es la siguiente:

$$DivEst=\sum^{i,j}(c_{ij}/Ncasos)^2$$

Esta fórmula nos va a permitir comparar resultados entre diferentes matrices con el mismo conjunto de categorías léxicas. Sin embargo, si cambiara el número de categorías léxicas los resultados no serían comparables. Queda pendiente por tanto obtener un valor normalizado para este caso.

2.2 Diferencias estilísticas

El índice DivEst nos permite comprobar el grado de variación, pero es posible que dos textos cuyo DivEst sea similar tengan comportamientos locales diferenciados. Por ello necesitamos un parámetro (al menos) que nos permita establecer la distancia entre dos matrices (o dos textos). Podemos comparar dos matrices mediante la siguiente fórmula:

$$DifEst=\sum^{i,j}((c_{1ij} - c_{2ij})/Ncasos_2)^2$$

O sea, sumamos la distancia de cada frecuencia (dividida por el número de casos *Ncasos*) al cuadrado. Los valores extremos oscilarán en este caso entre 0 el 2.

3. Análisis de textos españoles

3.1 Corpus

Para comprobar la validez de tales fórmulas seleccionamos los textos con estos criterios. Por un lado empleamos textos bien diferenciados diastráticamente. Suponemos que los textos pertenecientes a sublenguajes muy específicos tales como los textos informativos deben ser distribucionalmente más simples. Por el contrario, podemos suponer que la riqueza expresiva de los textos literarios debe reflejarse en una mayor variedad distribucional y, por lo tanto, en un valor mayor del índice DivEst. Junto a estos dos casos hemos seleccionado un texto perteneciente al habla espontánea[2]. Por otro lado, hemos tomado pares de textos de uno de los dos primeros tipos con el fin de comprobar si los valores coincidían para textos diferentes.

Los textos empleados son los siguientes.

[2] Pertenecientes al corpus VUM (Universidad de Málaga). Agradecemos a J.Villena y A. Avila la amable cesión de material sin el cual no podríamos haber realizado este trabajo.

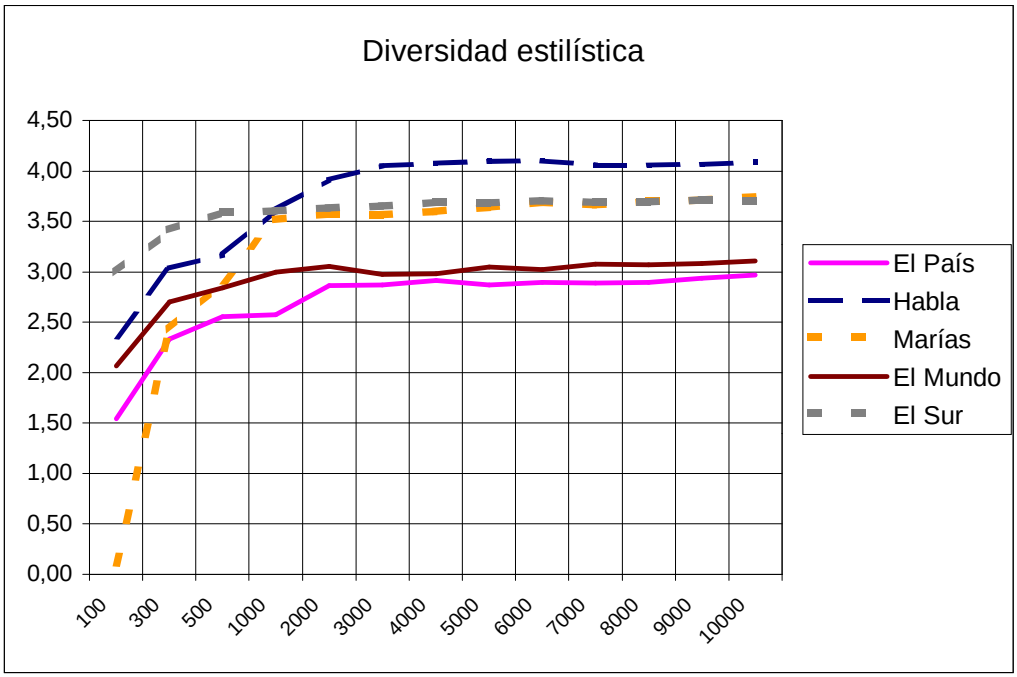
a) Textos informativos. Todas las noticias de las tres primeras páginas de la versión electrónica de los diarios *El País* y *El Mundo* de la semana del 17 al 23 de marzo de 1999. En todos los casos se han seleccionado sólo los textos cuyos autores son redactores de los periódicos seleccionados, esto es, no son textos de agencias o traducciones.

b) Textos literarios. Hemos escogido dos novelas españolas recientes: *El Sur*, de Adelaida García Morales, y *Corazón tan blanco*, de Javier Marías.

c) Habla espontánea: Emisiones de habla espontánea de sujetos de nivel cultural medio.

3.2 Índice de diversidad estilística

A continuación mostramos los valores de DivEst para los textos estudiados[3]:



Como se ve en el gráfico, a partir de los 3000 casos aproximadamente, el valor DivEst se estabiliza. Además, los valores obtenidos nos permite distinguir:

a) Textos informativos (líneas continuas)**:** Son los textos menos variados. Usan un número menor de las posibilidades combinatorias de la lengua. Su valor de DivEst es próximo a 3.

b) Habla espontánea (trazos largos)**:** Son los textos más heterogéneos. En ellos se da el mayor grado de variedad global. Su valor de DivEst es aproximadamente 4.

c) Textos literarios (trazos cortos): Se sitúan más próximos al habla que a los textos informativos. Su valor DivEst es de 3,7 aproximadamente.

---

[3] Dado que los valores reales obtenidos son muy pequeños, para facilitar la lectura de los datos hemos hecho algunas modificaciones sobre la fórmula recogida anteriormente. En concreto, hemos restado el valor obtenido de 0,05 y lo hemos multiplicado por 100:

$$DifEst=(0{,}05-\sum_{i,j}((c_{1ij} - c_{2ij})/Ncasos_2)^2)*100$$

Con ello los datos siguen siendo comparables, pero los valores manejados parecen más claros.

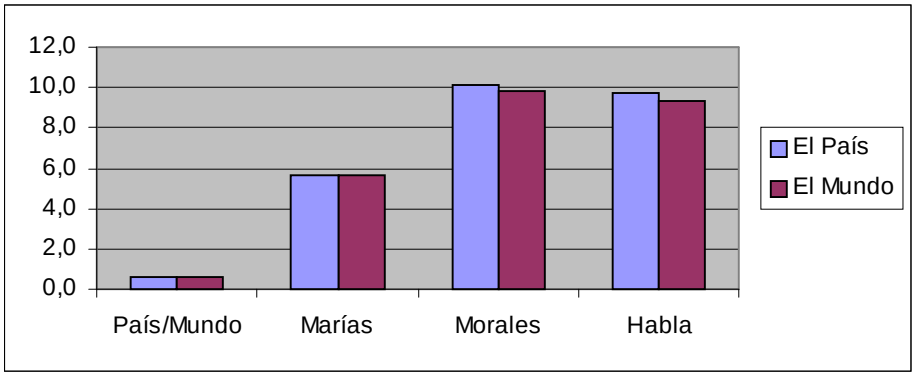
Estos datos nos permiten afirmar que el valor DivEst es significativo lingüísticamente ya que: textos independientes dan valores similares si pertenecen a una misma variedad y, además, en todos los casos el valor obtenido se estabiliza a partir de una determinada cantidad de casos.

3.2 Índice de diferencias distribucionales

A continuación mostramos el resultado de comparar algunos de los textos escogidos. Con el fin de que los valores obtenidos sean más intuitivos, los multiplicamos por 100.

3.2.1 Textos periodísticos

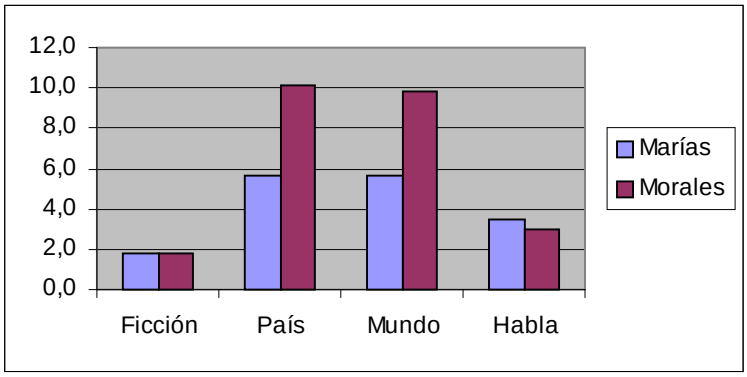
Los datos para los textos periodísticos de noticias son los siguientes. Para cada par de columnas, la columna izquierda muestra la distancia de el periódico El País con otro texto; la columna derecha muestra la distancia del periódico El Mundo con otro texto.



Como se ve, los textos periodísticos son muy parecidos entre sí (de ahí que el primer par de columnas sean muy pequeñas). El mayor tamaño de los siguientes pares de columnas –y especialmente del par que resulta de comparar con el texto de Morales–, muestra que los textos de ambos periódicos se diferencian mucho de los otros textos. Y, lo que es más importante, que las diferencias son siempre similares. O sea, ambos son muy diferentes del texto de Morales y del *habla* y ambos son razonablemente diferentes del texto de Marías. Estos hechos confirman que los dos textos seleccionados debe emplear el mismo tipo de recursos estilísticos.

3.2.2 Textos literarios

En el siguiente gráfico el primer par de filas recoge las diferencias entre los dos textos literarios. Los siguientes pares filas recogen las diferencias con los restantes textos:



A diferencia del caso anterior, estos dos textos no resultan tan parecidos entre sí (así, el primer par de columnas es tres veces mayor que el par de los textos periodísticos). Además, no se diferencias en la misma medida de los textos periodísticos. El texto de Marías es más próximo al texto periodísticos que el de Morales. En la medida en la que estos datos son representativos,

apuntan por tanto a que si bien estos dos textos presentan un grado de variación similar (como capta el índice DivEst), emplean las técnicas diferentes para alcanzar esta variedad.

4. Tipos de matrices y la oposición *estilo nominal/estilo verbal*

A menudo se ha observado que hay una oposición estilística básica entre el lenguaje periodístico y el lenguaje oral (por ejemplo, Criado de Val (1966), Martínez Albertos (1974), Nuñez Lavedeze (1991) o Avila Muñoz (1998)): el primero tiene una tendencia clara a lo nominal, mientras que el segundo es claramente verbal. Nuñez Lavedeze considera, siguiendo a Batjin (1990), que la oposición entre *estilo nominal/estilo verbal* debe ser la base a partir de la cual avanzar en la descripción estilística del uso lingüístico. La nominalidad o verbalidad se manifiesta mediante diversos rasgos:

- **Rasgos de lo nominal**: lo estático, la voz pasiva, formas impersonales del verbo, participios, etc.
- **Rasgos de lo verbal**: lo activo, la voz activa, formas personales del verbo, gerundios, etc.

Nuñez Lavedeze habla además de dos estilos nominales: el informativo y el conceptista, y critica el abuso del primero[4]. Podemos preguntar ahora si hay alguna relación clara entre la nominalidad que critica Nuñez Lavedeze y la diversidad que hemos medido en el apartado anterior.

4.1 La oposición estilo nominal/estilo verbal y el índice DivEst

Podemos comprobar, al menos parcialmente, el carácter nominal de un texto examinando la frecuencia de las categorías gramaticales propias del estilo nominal y examinando las posibles correspondencias entre estos valores y el índice DivEst. Para ello empleamos una función estadística común: el *índice de correlación*. La siguiente tabla muestra las frecuencias entre algunas categorías nominales y el valor de DivEst obtenido:

Relación entre las categorías nominales y DivEst

Categorías gram. / Textos	Sust.	Nomb. propio	Adjet.	Part/ Adjet.	Prepo-sición	Total	DivEst
El País	16,8%	9,6%	5,5%	1,6%	14,2%	**47.9%**	**2,96**
El Mundo	16,1%	10,9%	5,9%	1,3%	14,2%	**48,5%**	**3,1**
Habla	11,2%	2,4%	2,1%	,4%	8,8%	**25,0%**	**4,09**
Morales	13,8%	0,8%	3,5%	,8%	11,6%	**30,8%**	**3,71**
Marías	15,9%	1,2%	5,6%	1,7%	11,4%	**35,9%**	**3,74**
Correlación con DivEst	**0,870**	**0,872**	**0,811**	**0,693**	**0,983**	**0,978**	

[4] Nuñez Lavedeze critica abiertamente el estilo nominal informativo: "El recurso a deícticos anafóricos y a locuciones prepositivas para prolongar preposicional y no conjuntivamente la oración también coopera con la degradación de la finura modalizadora de los enlaces conjuntivos y ayuda a deteriorar el ritmo sintáctico y las inflexiones características de la oración compleja. […] También es más fácil, pero menos elegante y más inexpresivo, añadir complementos al verbo que sustituirlos por giros verbales y subordinaciones. En general el verbo está ligado a la conjunción y el estilo verbal es a la vez personal y conjuntivo. Por el contrario el nombre está ligado a la preposición." (págs. 156-7).

Los datos[5] muestran que al aumentar el número de elementos nominales disminuye la variedad (menor DivEst). Así, en los textos periodísticos, en los elementos nominales representan aproximadamente un 48%, la variedad es muy baja –del orden de 3. Por el contrario, en el texto de habla espontánea, en el que los elementos nominales representan tan solo un 25%, la variedad es mucho mayor –superior a 4. Esta correspondencia es la que captan matemáticamente los valores de la última fila. Así, el valor *0,978* de la columna Total indica que hay una relación muy marcada entre lo nominal y la variedad medida por DivEst.

Mostramos ahora una tabla semejante a la anterior que muestra la relación entre los elementos verbales y el índice DivEst.

Relación entre las categorías verbales y DivEst

Texto	Haber	Verbo f. personal	Parti-cipio	Gerun-dio	Infini-tivo	Copula	Con-junción	Sumas	DivEst
País	0,4%	6,2%	0,5%	0,3%	1,8%	1,0%	0,6%	**11,0%**	**2,96**
Mundo	0,3%	6,5%	0,3%	0,2%	2,2%	0,7%	0,5%	**11,0%**	**3,1**
Habla	0,7%	9,3%	0,6%	0,4%	2,3%	3,0%	2,7%	**19,5%**	**4,09**
Morales	0,8%	10,7%	0,8%	0,9%	3,5%	1,0%	1,4%	**19,4%**	**3,71**
Marías	0,7%	7,6%	0,6%	0,7%	2,5%	1,3%	1,1%	**14,7%**	**3,74**
Correlación	**-0,861**	**-0,776**	**-0,711**	**-0,570**	**-0,488**	**-0,783**	**-0,917**	**-0,911**	

Los datos muestran que el aumento del número de elementos verbales es paralelo al aumento de la diversidad (mayor DivEst). Así, en los textos periodísticos los elementos verbales representan tan solo un 11% y la variedad es escasa. Por el contrario, en el texto oral el número de elementos verbales sube hasta el 19,5% y la variedad aumenta al valor máximo. Esta correspondencia inversa es la que captan matemáticamente los valores de la última fila. Así, el valor *-0,911* de la columna Total indica que hay una relación inversa muy alta entre lo verbal y la variedad medida por DivEst.

Riqueza estilística y variedad estilística

Los textos más variados son los orales y no los literarios. Ello nos lleva a hacer una distinción entre variedad estilística y riqueza estilística. Los textos orales son más variados pero no más ricos que los textos orales, cuya variación proviene en parte de usos anómalos. Dicho de otra forma, hay un punto a partir del cual la variedad no es indicio de riqueza sino de desorden.

Estilo nominal y poca variedad

El efecto empobrecedor del estilo nominal puede deberse a que el verbo fuerza la aparición de más recursos gramaticales. Como elemento central del sintagma verbal y de la oración, al usar un verbo usamos también otros elementos como: pronombres personales átonos y tónicos, otros pronombres, adverbios, conjunciones, preposiciones y, por supuesto, sustantivos.

Sin embargo, el sustantivo es el núcleo del sintagma nominal. Así, al utilizarlo sólo nos vemos obligados a usar elementos asociados al sintagma nominal: artículos, adjetivos gramaticales o determinantes, adjetivos, pero no adverbios de modo, conjunciones o verbos.

De esa forma podríamos decir que la nominalización conlleva una simplificación generalizada en lo que se refiere a la variedad de elementos gramaticales empleados. Se trata por tanto de

[5] El índice de correlación oscila entre -1 y 1. Cuando es próximo a 1, quiere decir que la relación entre los dos conjuntos de datos es muy alta.

una tendencia no sólo criticable en términos semánticos (como hace Nuñez Lavedeze) sino también en términos meramente gramaticales ya que es la causa de la limitación del número de recursos gramaticales empleado.

5. Conclusiones

Nos planteábamos al principio del artículo la dificultad que planteaba medir la variedad de recursos empleados por un texto. En este artículo hemos mostrado cómo lograrlo a partir de una matriz de transiciones. Hemos visto también que podemos asociar diferentes valores a tipos de textos y que tales valores son estables entre dos textos específicos del mismo tipo.

En concreto los textos oscilan entre la máxima variedad propia de los textos orales espontáneos hasta la máxima rigidez de los textos periodísticos. En un punto intermedio, pero bastante más cercano a la lengua oral, quedan los textos literarios analizados.

Además, hemos analizado las causas que llevan al empobrecimiento estilístico y hemos podido mostrar, apoyando así a Nuñez Lavedeze, que el estilo nominal tiene un efecto empobrecedor sobre los textos periodísticos.

Estos resultados nos hacen plantearnos numerosas preguntas y otras posibilidades, algunas de las cuales están siendo objeto de estudio en estos momentos. Entre otras, la medida que proponemos podría aplicarse al estudio del lenguaje infantil, como instrumento objetivo para conocer el desarrollo lingüístico de un niño, o podría emplearse para mejorar los propios sistemas de lematización automática –ya que el sistema podría tomar una MT diferente según el tipo de texto que lematice.

Bibliografía

Albalá, M.J., Capelli,G.,Marrero, M.V. e I. Moreno-Torres (1996). "Sistema de análisis informático del español", comunicación presentada al 26º Congreso de la Sociedad Española de Lingüística Madrid.

Ávila Muñoz, A. (1998) *Elaboración, anotación y análisis del corpus oral del proyecto VUM*. Tesis doctoral inédita, Universidad de Málaga.

Batjin, M.M. (1990). *Estética de la creación verbal*, México, Siglo XXI.

Biber, D. (1995a). *Dimensions of register Variation: A cross linguistic comparison*. Cambridge, , CUP.

Biber, D. (1995b). "On the role of computational, statistical and interpretive techniques in multi-dimensional analyses of register variation: A reply to Watson." *Text* 15 (3), págs. 341-370.

Criado de Val, M. (1980). *Estructura general del coloquio*. Madrid, SGEL.

Garside, G. et al. (1987). *The Computational Analysis of English: a corpus based apporach*, Londres, Longman.

Martínez Albertos; J.L. (1974) *Redacción periodística los estilos y géneros en la prensa escrita*. Barcelona, A.T.E.

Nuñez Lavedeze, L. (1991). *Teoría y práctica de la construcción del texto*. Barcelona, Ariel Comunicación.

Romero Gualda, M.V. (1993). *El Español en los medios de comunicación*, Madrid, Arco Libros.