Análisis de regresión logística aplicado a la clasificación textos académicos: Biometría y Filosofía

Analysis of Logistic Regression Applied to the Classification of Academic Texts: Biometrics and Philosophy

Celina Beltrán
Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina
beltranc36@yahoo.com.ar

Abstract

This work is aimed to continue with the application of the multivariate static analysis carried out in Beltrán (2010). This paper employs the resulting information of the automatic analysis of academic texts coming from different scientific areas (Biometrics and Philosophy) to constitute a database on which the logistic regression technique is applied. This application differs from the discriminant analysis used on a previous work mainly in the required assumptions about the distributions of variables and the resulting information of the evaluated model. This study allows an analysis that shows those characteristics that discriminate the analyzed texts corpora as working with the absolute frequency of different morphosyntactic categories. The significant variables constituting the proposed model correspond to two categories: adverbs and clitics. The estimated *odds ratio* shows that the possibilities of classifying a text in a corpus of Biometrics increase to 62% if raising the amount of clitics in the unit, while the possibilities of classifying a text in a corpus of Philosophy increase to 41% if raising the amount of adverbs in the unit. The global error rate estimated by cross validation is 19%.

Keywords: multivariate logistic regression – multivariate analysis – automatic text analysis.

Resumen

Este trabajo pretende continuar la aplicación del análisis estadístico multivariado llevada a cabo en Beltrán (2010). En este artículo se utiliza la información resultante del análisis automático de textos académicos provenientes de distintas áreas científicas (Biometría y Filosofía) para conformar una base de datos sobre la cual se aplica la técnica de regresión logística. Esta aplicación presenta diferencias respecto al análisis discriminante aplicado en un trabajo anterior principalmente en los supuestos requeridos sobre las distribuciones de las variables y en la información resultante del modelo estimado. El estudio permite un análisis en el cual se evidencian aquellas características que discriminan los corpus de textos analizados trabajando con las frecuencias absolutas de las distintas categorías morfosintácticas. Las variables significativas que conforman el modelo propuesto corresponden a dos categorías: adverbios y clíticos. Los *odds ratio* estimados evidencian que la

chance de clasificar a un texto dentro del corpus de Biometría se incrementa en un 62% al aumentar en número de clíticos en una unidad, mientras que la chance de clasificarlo en el corpus de Filosofía aumenta un 41% al incrementarse en una unidad el número de adverbios. La tasa de error global estimada por validación cruzada es del 19%.

Palabras claves: Regresión logística multivariada, análisis multivariado, análisis automático de textos.

1. INTRODUCCION

Este trabajo pretende continuar el análisis estadístico multivariado llevado a cabo en Beltrán (2010). El analizador morfológico Smorph, implementado como etiquetador, es utilizado para asignar una categoría morfológica a todas las ocurrencias lingüísticas.

Se utiliza la información resultante del análisis automático de textos académicos provenientes de distintas áreas científicas (Biometría y Filosofía) para conformar una base de datos sobre la cual se aplica la técnica de regresión logística. Esta aplicación presenta diferencias respecto al análisis discriminante aplicado en trabajos previos principalmente en los supuestos requeridos sobre las distribuciones de las variables y en la información resultante del modelo estimado.

El análisis discriminante y la técnica de regresión logística son técnicas ampliamente utilizadas cuando se tiene por objetivo identificar el grupo al cual pertenece una unidad experimental. En este caso la regresión logística pretende predecir el corpus al cual pertenece un texto en función de la información relevada en el análisis automático de los mismos. A diferencia del análisis discriminante no se requiere el supuesto de normalidad multivariada del conjunto de variables regresoras, lo cual permite trabajar con las variables originales que resultan del análisis morfológico sin necesidad de transformarlas.

Mediante la interpretación de los coeficientes del modelo estimado se busca hallar las características, considerándolas simultáneamente a todas ellas, provenientes del análisis automático de los textos que son más discriminatorias de las áreas científicas de las cuales provienen.

2. MATERIAL Y METODOS

2.1. Diseño de la muestra

El marco muestral para la selección de la muestra es el utilizado en trabajos anteriores. Está compuesto por textos académicos, resúmenes de trabajos presentados a congresos y revistas, extraídos de internet pertenecientes a las disciplinas: Biometría y Filosofía. La unidad de muestreo fue el texto y la selección de la muestra se llevó a cabo empleando un diseño muestral estratificado con selección proporcional al tamaño, siendo la medida de tamaño el “número de palabras del texto”.

Las muestras de los dos estratos fueron evaluadas y comparadas respecto al número medio de palabras por texto. Esta comparación se requiere para evitar que la discriminación entre las disciplinas se vea afectada por el tamaño de los textos.

La conformación de la muestra final se presenta en la tabla 1.

Tabla 1. Conformación de la muestra final

Muestra	Nro. de textos	Cantidad de palabras
Biometría	30	5047
Filosofía	30	5513

2.2. Etiquetado: Análisis morfológico de los textos

El software Smorph, analizador y generador morfosintáctico desarrollado en el Groupe de Recherche dans les Industries de la Langue (Universidad Blaise-.Pascal, Clermont II) por Salah Aït-Mokhtar (1998) realiza en una sola etapa la tokenización y el análisis morfológico. A partir de un texto de entrada se obtiene un texto lematizado con las formas correspondientes a cada lema (o a un subconjunto de lemas) con los valores correspondientes. Se trata de una herramienta declarativa, la información que utiliza está separada de la maquinaria algorítmica, en consecuencia, puede adaptarse a distintos usos. Con el mismo software se puede tratar cualquier lengua si se modifica la información lingüística declarada en sus archivos.

Smorph compila, minimiza y compacta la información lingüística que queda disponible en un archivo binario. Los códigos fuente se dividen en cinco archivos: Códigos ASCII, Rasgos, Terminaciones, Modelos y Entradas.

En el archivo **entradas**, se declaran los ítems léxicos acompañados por el modelo correspondiente. Este indicador de modelo oficia de enlace con el archivo modelos, en el que se especifica la información morfológica y las terminaciones que se requieren en cada ítem. El archivo **modelos**, es el que introduce la información correspondiente a los modelos de flexiones morfológicas, mientras que en el archivo **terminaciones** es necesario declarar todas las terminaciones que son necesarias para definir los modelos de flexión. Las etiquetas correspondientes a los rasgos morfológico-sintácticos son organizadas jerárquicamente en el archivo **rasgos**. Por último, en el archivo de códigos **ASCII** se especifican, entre otros, los caracteres separadores y las equivalencias entre mayúsculas y minúsculas.

El módulo post-smorph **MPS** es un analizador que recibe en entrada una salida Smorph (en formato Prolog) y puede modificar las estructuras de datos recibidos. Ejecuta dos funciones principales: la Recomposición y la Correspondencia, que serán útiles para resolver las ambigüedades que resulten del análisis de Smorph.

La información contenida en estos archivos es la presentada en Beltrán (2009) para implementar el etiquetador.

2.3. Diseño y desarrollo de la base de datos

El resultado del análisis de Smorph-Mps se almacena en un archivo de texto. Esta es la información que contendrá la base de datos.
Mediante una función definida en el sistema estadístico R se logra captar la información resultante del análisis morfológico y disponerla en una matriz de dimensión: tantas filas como cantidad de objetos lingüísticos tenga el texto y tantas columnas como ocurrencia+lema+valores. De esta manera se obtiene una base de datos con la estructura que se muestra en la tabla 2.

Tabla 2. Fragmento de la base de datos obtenida

MUESTRA	TEXTO	OCURRENCIA	LEMA	ETIQUETA
1	1	El	El	det
1	1	problema	problema	nom

1	1	de	De	prep
1	1	las	El	det
1	1	series	Serie	nom
1	1	de	De	prep
1	1	tiempo	Tiempo	nom
1	1	se	Lo	cl
…	…	…	…	…
2	1	Uno	Uno	pron
2	1	de	De	prep
2	1	los	El	det
2	1	agentes	Agente	nom
2	1	que	Que	rel
2	1	ha	Haber	aux
2	1	provocado	provocar	v
2	1	una	Una	det
2	1	verdadera	verdadera	adj
2	1	transformación	transformación	nom
2	1	en	En	prep
…	…	…	…	…

Abreviaturas:
‘adj’: adjetivo ‘art’: artículo ‘nom’: nombre ‘prep’: preposición ‘v’: verbo ‘adv’: adverbio ‘cl’: clítico ‘aux’: auxiliar ´cop’: copulativo ‘pun’: signo de puntuación

Luego, a partir de esta base de datos por palabra (cada unidad o fila es una palabra analizada del texto), se confecciona la base de datos por documento que será analizada estadísticamente. Esta es una nueva base, donde cada unidad es el texto, que retiene la información de las variables indicadas en la tabla 3.a con la estructura presentada en la tabla 3.b.

Tabla 3.a. Variables de la base de datos por documento

CORPUS	Corpus al que pertenece el texto
TEXTO	Identificador del texto dentro del corpus
adj	cantidad de adjetivos del texto
adv	cantidad de adverbios del texto
cl	cantidad de clíticos del texto
cop	cantidad de copulativos del texto
det	cantidad de determinantes del texto
nom	cantidad de nombres (sustantivos) del texto
prep	cantidad de preposiciones del texto
v	cantidad de verbos del texto
otro	cantidad de otras etiquetas del texto
total_pal	cantidad total de palabras del texto

Tabla 3.b. Fragmento de la base de datos para análisis estadístico

CORPUS	TEXTO	adj	adv	cl	cop	det	nom	prep	v	OTRO	TOTAL_PAL
1	1	21	4	4	8	30	48	33	17	20	185

1	2	14	0	5	4	14	27	20	9	17	110
1	3	16	5	11	5	28	47	26	18	25	181
…	…	…	…	…	…	…	…	…	…	…	…
2	28	14	2	3	6	30	60	39	16	16	186
2	29	14	0	4	5	24	40	26	12	16	141
2	30	18	5	2	5	35	49	30	19	20	183

2.4. Análisis de regresión Logística

2.4.1. El modelo

La regresión logística es utilizada en situaciones en las cuales el objetivo es describir la relación entre una variable respuesta categórica, en este caso dicotómica, y un conjunto de variables explicativas que pueden ser tanto categóricas como cuantitativas.
Sea **x** un vector de p variables independientes, esto es, **x´**= (x_1, x_2,…, x_p). La probabilidad condicional de que la variable *y* tome el valor 1 (presencia de la característica estudiada), dado valores de las covariables **x** es:

$$P(y=1/X)=\pi(x)=\frac{e^{g(x)}}{1+e^{g(x)}}$$

donde $g(x)=\beta_0+\beta_1x_1+...+\beta_px_p$

β_0 es la constante del modelo o término independiente
p el número de covariables
βi los coeficientes de las covariables
xi las covariables que forman parte del modelo.

Si alguna de las variables independientes es una variable discreta con k niveles se debe incluir en el modelo como un conjunto de k-1 "variables de diseño" o "variables dummy". El cociente de las probabilidades correspondientes a los dos niveles de la variable respuesta se denomina odds y su expresión es:

$$\frac{P(y=1|X)}{1-P(y=1|X)}=e^{\beta_0+\beta_1x_1+...+\beta_px_p}$$

Si se aplica el logaritmo natural, se obtiene el modelo de regresión logística:

$$\log\left(\frac{P(y=1|X)}{1-P(y=1|X)}\right)=\log(e^{\beta_0+\beta_1x_1+...+\beta_px_p})=\beta_0+\beta_1x_1+...+\beta_px_p$$

En la expresión anterior el primer término se denomina logit, esto es, el logaritmo de la razón de proporciones de los niveles de la variable respuesta.

2.4.2. Estimación y significación de los coeficientes del modelo

Sea una muestra aleatoria de n observaciones independientes de pares $(\mathbf{x}_i, y_i)$ para i=1,2,...,n. El objetivo es estimar el vector de parámetros $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2, ..., \beta_p)$ por el método de Máxima Verosimilitud.

Las p+1 ecuaciones a resolver se obtienen derivando la función de verosimilitud respecto a cada uno de los parámetros del modelo e igualando a cero. Las ecuaciones quedan expresadas de la siguiente manera:

$$\sum_{i=1}^{n} (y_i - \pi(\mathbf{x}_i)) = 0$$

$$\sum_{i=1}^{n} x_{ej} (y_i - \pi(\mathbf{x}_i)) = 0 \qquad j = 1,2,..., p$$

Las soluciones de estas ecuaciones son los estimadores máximo verosímiles de cada uno de los componentes del vector de parámetros $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2, ..., \beta_p)$, que se simboliza $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$. Asimismo, de acuerdo al método de estimación por máxima verosimilitud, los estimadores de las variancias y covariancias se obtienen a partir de las derivadas parciales segundas de la función de verosimilitud.

Para comprobar la significación estadística de cada uno de los coeficientes de regresión en el modelo se puede utilizar, entre otros, el test de Wald y el test de razón de verosimilitudes.

2.4.3. Interpretación de los coeficientes estimados

Los β_j estimados representan tasa de cambio de una función de la variable dependiente *y* por unidad de cambio de la variable independiente *x*.
El coeficiente β_i expresa el cambio resultante en la escala de medida de la variable *y* para un cambio unitario de la variable *x*. Por ejemplo, para la variable x_1, $\beta_1 = g(x_1+1) - g(x_1)$ representa el cambio en el logit frente a un incremento de una unidad en la variable x_1. La interpretación se hace en términos de la razón de Odds (OR).

$$OR = \frac{\left(\frac{P(Y=1/\mathrm{x_j}+1)}{P(Y=0/\mathrm{x_j}+1)} \right)}{\left(\frac{P(Y=1/\mathrm{x_j})}{P(Y=0/\mathrm{x_j})} \right)} = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_j(x_j+1)+...+\beta_p x_p}}{e^{\beta_0+\beta_1 x_1+...+\beta_j x_j+...+\beta_p x_p}} = e^{\beta_j}$$

2.4.4. Selección de variables

Una cuestión importante en este tipo de análisis es determinar si todas las variables consideradas en la función de discriminante contienen información útil y si solamente algunas de ellas son suficientes para diferenciar los grupos (en este caso las disciplinas). Dado que las variables utilizadas para explicar la respuesta es probable que estén correlacionadas, es posible también que

compartan información. Por lo tanto, se puede buscar un subgrupo de variables mediante algún criterio de modo tal que las variables excluidas no contengan ninguna información adicional. Existen varios algoritmos de selección de variables, entre ellos podemos citar:
Método forward: comienza por seleccionar la variable más importante y continúa seleccionando las variables más importantes una por vez, usando un criterio bien definido. Uno de estos criterios involucra el logro de un nivel de significación deseado pre-establecido. El proceso termina cuando ninguna de las variables restantes encuentra el criterio pre-especificado.
Método backward: comienza con el modelo más grande posible. En cada paso descarta la variable menos importante, una por vez, usando un criterio similar a la selección forward. Continúa hasta que ninguna de las variables pueda ser descartada.
Selección paso a paso: combina los dos procedimientos anteriores. En un paso, una variable puede entrar o salir desde la lista de variables importantes, de acuerdo a algún criterio pre-establecido.

2.4.5. Bondad de ajuste del modelo:

En este trabajo se utilizó como evaluación del ajuste del modelo el test de Hosmer-Lemeshow. Esta prueba propone un agrupamiento basado en las probabilidades estimadas por el modelo. Agrupa los n sujetos en g patrones según criterios estadísticos. La estadística del test se obtiene calculando la estadística chi-cuadrado de Pearson que compara las frecuencias observadas y las estimadas, en cada grupo y categoría de la variable respuesta, esto es, la estadística chi-cuadrado calculada a partir de una tabla 2 x g. La ausencia de significación de la misma indica un buen ajuste del modelo.
Otra medida que permite evaluar el modelo cuando es utilizado para clasificar unidades en dos grupos es la tasa de error estimada por validación cruzada.

3. RESULTADOS

3.1. Análisis preliminar.

La primera comparación que se realiza, como ya se mencionó al describir la muestra, es la del número de palabras por texto. La misma se lleva a cabo mediante el test no paramétrico de Wilcoxon para muestras independientes arrojando una probabilidad asociada p=0.796, evidenciando que no existen diferencias significativas entre los corpus respecto al tamaño de los textos.

Comparaciones similares entre los corpus se llevan a cabo para las restantes variables hallando diferencias significativas ($p<0.05$) para el número de clíticos y de adverbios en los documentos analizados (Tabla 4). El número de clíticos es mayor en los textos de biometría y el número de adverbios es superior en los textos de filosofía.

Tabla 4. Comparación mediante test de Wilcoxon

Número promedio de:	**BIOMETRIA**	**FILOSOFIA**	**General**	**Valor de p**
adjetivos	17,9	21,3	**19,6**	0,54861
adverbios	2,9	5,9	**4,4**	0,01046
clíticos	4,1	2,7	**3,4**	0,00698
Copulativos	4,7	6,0	**5,4**	0,11850
Determinantes	26,8	32,4	**29,6**	0,35490
Nombres	44,6	45,0	**44,8**	0,55400
Preposición	30,0	29,7	**29,9**	0,67317
Verbos	16,1	18,4	**17,2**	0,85882
Otro	18,8	21,4	**20,1**	0,85318

TOTAL_PALABRAS	165,8	182,9	**174,4**	0,79578

3.2. Análisis de Regresión Logística

Se realizó un análisis de regresión logística para obtener una regla de clasificación que permita asignar los textos en estas dos poblaciones, definidas por el área científica a la que pertenecen, en base a la frecuencia de cada categoría gramatical en el texto.
Considerando todas las categorías se obtiene el siguiente modelo de regresión logística:

Tabla 5: Coeficientes del modelo de regresión logística

Estimación máximo verosímil					
Coeficiente	**gl**	**Estimador**	**Error estándar**	**Est. Chi-cuadrado**	**Prob. asociada**
Intercepto	1	0.6223	1.1398	0.2981	0.5851
Adj	1	-0.0474	0.0763	0.3863	0.5342
Adv	1	-0.3986	0.2103	3.5901	0.0581
Cl	1	0.3275	0.2197	2.2222	0.1360
Cop	1	-0.3867	0.2122	3.3189	0.0685
Det	1	-0.1288	0.0770	2.7969	0.0944
Nom	1	0.1108	0.0777	2.0359	0.1536
Prep	1	0.0126	0.0772	0.0265	0.8707
V	1	0.0546	0.0873	0.3912	0.5317

Este modelo permite, mediante la utilización de los coeficientes estimados, calcular para cada texto la probabilidad de pertenecer a cada uno de los corpus.

$$P(y=1/X)=\pi(x)=\frac{e^{\beta_0+\beta_1x_1+\ldots+\beta_j(x_j+1)+\ldots+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+\ldots+\beta_j(x_j+1)+\ldots+\beta_px_p}}$$

$$P(y=0/X)=\pi(x)=\frac{1}{1+e^{\beta_0+\beta_1x_1+\ldots+\beta_j(x_j+1)+\ldots+\beta_px_p}}$$

Con este criterio un texto es asignado al corpus cuya probabilidad es máxima. Aplicando este modelo como regla de clasificación y estimando por validación cruzada, la tasa de error global que se obtiene es del 18% (Tabla 6).

Tabla 6: Tasa de error estimada

Tasa de error por corpus			
	BIOMETRIA	**FILOSOFIA**	**Total**
Tasa	16.7%	20%	18.3%

Para determinar cuáles categorías gramaticales son las responsables de la discriminación se utilizaron los tres algoritmos de selección de variables. En todos los casos se llega al mismo resultado: las variables que logran diferenciar a los dos corpus de textos analizados son el número de adverbios y de clíticos.

El modelo final estimado luego de la selección de variables se muestra en la tabla 7 y las razones de odds resultantes en la tabla 8.

Tabla 7: Coeficientes del modelo de regresión logística final

Estimación máximo verosímil					
Coeficiente	**gl**	**Estimador**	**Error estándar**	**Est. Chi-cuadrado**	**Prob. asociada**
Intercepto	1	-0.3150	0.7128	0.1952	0.6586
adv	1	-0.3418	0.1551	4.8561	0.0275
cl	1	0.4828	0.1737	7.7277	0.0054

Tabla 8: Razones de odds estimadas

Razón de odds			
Efecto	**Estimación puntual**	**IC 95%**	
adv	0.711	0.524	0.963
cl	1.621	1.153	2.278

La bondad del ajuste de evalúa mediante el test de Hosmer-Lemeshow y la tasa de error de clasificación. Con el modelo de regresión logística obtenido durante la selección de variables se obtiene una tasa de error global del 19% mediante validación cruzada (Tabla 9) y la probabilidad asociada en el test de bondad de ajuste es p=0.2543 evidenciando lo adecuado del modelo.

Tabla 9: Tasa de error estimada

Tasa de error por corpus			
	BIOMETRIA	**FILOSOFIA**	**Total**
Tasa	18%	20%	19%

Los coeficientes del modelo de regresión logística permiten la interpretación de la misma. Las diferencias entre los dos tipos de textos se basan fundamentalmente en el número de clíticos y de adverbios presentes. La razón de odds para el número de clíticos es 1.62 lo cual indica que la chance de clasificar a un texto dentro del corpus de Biometría se incrementa en un 62% al aumentar en número de clíticos en una unidad. Con respecto al número de adverbios la razón de odds es menor a la unidad por lo tanto si se interpreta el recíproco, 1/0.71=1.41, significa que la chance de clasificar un texto en el corpus de Filosofía aumenta un 41% al incrementarse en una unidad el número de adverbios.

6. CONCLUSIONES

Los resultados del análisis morfológico de los textos se analizaron teniendo en cuenta simultáneamente todas las mediciones realizadas sobre ellos sin aplicar ninguna transformación a las variables.

El análisis de regresión logística aplicado en este trabajo presenta una modalidad de análisis estadístico para discriminar grupos no muy habitual en la investigación lingüística. El mismo permitió hallar las categorías gramaticales cuyas frecuencias observadas en los textos permiten discriminar los dos grupos definidos por la disciplina a la que pertenecen.

Las diferencias entre los dos tipos de textos está centrada principalmente en el número de clíticos y de adverbios presentes. Los *odds ratio* estimados evidencian que la chance de clasificar a un texto dentro del corpus de Biometría se incrementa en un 62% al aumentar en número de clíticos en una unidad, mientras que la chance de clasificarlo en el corpus de Filosofía aumenta un 41% al incrementarse en una unidad el número de adverbios.

Similares resultados se hallaron en Beltrán (2010) utilizando un análisis discriminante sobre las variables transformadas. De la misma manera que en aquella instancia, se cree que puede deberse a que, en los textos de biometría/estadística hay más clíticos que en los humanísticos por la frecuencia de expresiones impersonales o pasivas con el clítico "se" del tipo:
"se ajusta un modelo cuadrático"

"se estima la variancia poblacional"

Mientras en los textos de filosofía se manifiesta la presencia de mayor proporción de adverbios.

Esta metodología puede ser generalizada a un número mayor de disciplinas, de las cuales provienen los textos, mediante una extensión del modelo de regresión logística para variable respuesta multinomial. En este sentido se continuará el trabajo presentado.

Referencias

Aitchison J. 1983. The Statistical Analysis of Compositional Data. Chapman & Hall, London.

Beltrán, C., Bender, C., Bonino, R., Deco, C., Koza, W., Méndez, B., Moro, Stella Maris. 2008 Recursos informáticos para el tratamiento lingüístico de textos. Ediciones Juglaría. Rosario.

Beltrán, C. 2009 Modelización lingüística y análisis estadístico en el análisis automático de textos. Ediciones Juglaría. Rosario.

Beltrán, C. 2010 Estudio y comparación de distintos tipos de textos académicos: Biometría y Filosofía. Revista de Epistemología y Ciencias Humanas. Grupo IANUS. Rosario.

Beltrán, C. 2010 Análisis discriminante aplicado a textos académicos: Biometría y Filosofía. Revista INFOSUR. Grupo INFOSUR. Rosario.

Bès,Gabriel, Solana, Z y Beltrán, C. 2005 Conocimiento de la lengua y técnicas estadísticas en el análisis lingüístico en Desarrollo, implementación y uso de modelos para el procesamiento automático de textos (ed. Víctor Castel) Facultad de Filosofía y Letras, UNCUYO

Cuadras, C.M. 2008 NUEVOS MÉTODOS DE ANÁLISIS MULTIVARIANTE. CMC Editions. Barcelona, España.

Hosmer, D.W.; Lemeshow, S. (1989) Applied Logistic Regression. John Wiley & Sons. New York.

Johnson R.A. y Wichern D.W. 1992 Applied Multivariate Statistical Análisis. Prentice-Hall International Inc.

Khattre R. y Naik D. 1999 Applied Multivariate Statistics with SAS Software. SAS. Institute Inc. Cary, NC. USA.

Khattre R. y Naik D. (2000) Multivariate Data Reduction and Discriminatio with SAS Software. SAS Institute Inc. Cary, NC. USA

Pogliano, A.M. (2010) "Análisis Estadístico de Datos Aplicados al Estudio de Calidad en Servicios de Traducción". Tesis Lic. en estadística. Facultad de Cs. Económicas y estadística. UNR.

Rodrigo Mateos, José Lázaro y Bès, Gabriel G. 2004 Análisis e implementación de clíticos en una herramienta declarativa de tratamiento automático de corpus. En VI Congreso de Lingüística General, Santiago de Compostela.

Solana, Z. Beltrán, C., Bender, C., Bonino, R., Deco, C., Koza, W., Méndez, B., Rodrigo, A., Tramallino, C. 2009 La interlengua de los aprendientes de español como L2. Aportes de la Lingüística Informática. GRUPO INFOSUR- Ediciones Juglaría.