Aplicación del análisis de regresión logística multinomial en la clasificación de textos académicos: Biometría, Filosofía y Lingüística informática.

APPLICATION OF THE MULTINOMIAL LOGISTIC REGRESSION ANALYSIS TO CLASSIFY ACADEMIC TEXTS: BIOMETRICS, PHILOSOPHY AND LINGUISTIC INFORMATICS

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

Abstract

This work aims at extending the application of the multivariate statistic analysis carried out in Beltrán (2010). The outcome information of the automatic analysis of academic texts of different scientific areas (Biometrics, Philosophy and Linguistic Informatics) is used to generate a database on which the multinomial logistic regression technique is applied. While in a previous paper logistic regression in dichotomous response variable was used to classify two types of texts was used, in this work an analysis for three categories is generalized. The study allows an analysis showing those characteristics discriminated by the corpora of the texts analyzed when the absolute frequencies of different morphosyntactic categories are employed. The significant variables considered in the proposed model correspond to three categories: adverbs, nouns, determiners, verbs, clitics and interactions between the latter two. The *odds ratio* estimated to compare each corpus with that of Biometrics (the reference corpus of the model) proves that:

- The possibility of classifying a text within the corpus of Philosophy versus Biometrics increases to 43% if raising the amount of adverbs in the unit, while the possibility of classifying a text in the corpus of Biometrics versus Linguistics increased to 18% if raising the amount of adverbs in the unit.
- The possibility of classifying a text within the corpus of Biometrics versus Linguistics increases to 16% if raising the amount of nouns in the unit. The amount of nouns does not discriminate the corpus of Biometrics and Philosophy.
- The possibility of classifying a text within the corpus of Philosophy versus Biometrics increases to 11% if raising the amount of determiners in the unit, while the possibility of classifying a text in the corpus of Biometrics versus Linguistics increased to15% if raising the amount of determiners in the unit.
- The possibility of classifying a text within the corpus of Philosophy versus Biometrics with regard to the amount of clitics is altered by the number of verbs (interaction). If the text has a verb frequency upper to 15%, the possibility of classifying in Philosophy versus Biometrics increases with the number of clitics.

  However, if the text shows low verb frequency, the opposite effect is obtained. With respect to the classification of Linguistics versus Biometrics, the situation is alike.

The global error rate estimated by cross validation is 14%.

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

Resumen

Este trabajo pretende continuar la aplicación del análisis estadístico multivariado llevada a cabo en Beltrán (2010). Se utiliza la información resultante del análisis automático de textos académicos provenientes de distintas áreas científicas (Biometría, Filosofía y Lingüística informática) para conformar una base de datos sobre la cual se aplica la técnica de regresión logística multinomial. Mientras en un trabajo anterior se utilizó regresión logística para variable respuesta dicotómica para clasificar dos tipos de textos, en este trabajo se generaliza el análisis para tres categorías. 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 tres categorías: adverbios, nombres, determinantes, verbos, clíticos y la interacción de estas dos últimas categorías. Los *odds ratio* estimados para comparar cada corpus con el de Biometría (corpus de referencia en el modelo) evidencian:

- La chance de clasificar a un texto dentro del corpus de Filosofía versus Biometría se incrementa en un 43% al aumentar en número de adverbios en una unidad, mientras que la chance de clasificarlo en el corpus de Biometría versus Lingüística aumenta un 18% al incrementarse en una unidad el número de adverbios.
- La chance de clasificar a un texto dentro del corpus de Biometría versus Lingüística se incrementa en un 16% al aumentar en número de nombres en una unidad. El número de nombres no discrimina los corpus de Biometría y Filosofía.
- La chance de clasificar a un texto dentro del corpus de Filosofía versus Biometría se incrementa en un 11% al aumentar en número de determinantes en una unidad, mientras que la chance de clasificarlo en el corpus de Biometría versus Lingüística aumenta un 15% al incrementarse en una unidad el número de determinantes.
- La chance de clasificar un texto dentro del corpus de Filosofía versus Biometría respecto al número de clíticos se ve afectado por el número de verbos (interacción). Cuando la frecuencia de verbos en el texto es superior al 15%, la chance de clasificar en Filosofía versus Biometría se incrementa con el número de clíticos. Sin embargo, cuando el texto presenta una frecuencia baja de verbos, el efecto es inverso. Con respecto a la clasificación en Lingüística versus Biometría la situación es la misma.

La tasa de error global estimada por validación cruzada es del 14%.

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

1. INTRODUCCION

Este trabajo pretende continuar el análisis estadístico multivariado llevado a cabo en Beltrán (2010), generalizando en esta oportunidad a la clasificación en tres áreas científicas de pertenencia de los textos. El analizador morfológico Smorph, implementado como etiquetador, es utilizado para

asignar una categoría morfológica a todas las ocurrencias lingüísticas.

La información resultante del análisis automático de textos académicos provenientes de distintas áreas científicas (Biometría, Lingüística y Filosofía) es utilizada para definir y construir una base de datos sobre la cual se aplica la técnica de regresión logística multinomial.

Si bien tanto el análisis discriminante como la técnica de regresión logística son técnicas ampliamente utilizadas cuando se tiene por objetivo identificar el grupo al cual pertenece una unidad experimental, a diferencia del análisis discriminante, la regresión logística no 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.

En esta aplicación, la regresión logística multinomial pretende predecir el corpus al cual pertenece un texto en función de la información relevada en el análisis automático de los mismos, cuando la variable respuesta (corpus) presenta más de dos categorías.

La interpretación de los coeficientes del modelo estimado permitirá hallar las categorías morfológicas, considerándolas simultáneamente a todas ellas, que son más discriminatorias de las áreas científicas de las cuales provienen los textos.

2. MATERIAL Y METODOS

2.1. Diseño de la muestra

El marco muestral para la selección de la muestra está compuesto por textos académicos, resúmenes de trabajos presentados a congresos y revistas, extraídos de internet pertenecientes a las disciplinas: Biometría, Lingüística informática 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 tres 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.

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 1.

Tabla 1. 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
…	…	…	…	…
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
…	…	…	…	…
3	1	permitió	permitir	v
3	1	el	el	det
3	1	análisis	análisis	nom
3	1	automático	automático	adj
…	…	…	…	…

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 2.a con la estructura presentada en la tabla 2.b.

Tabla 2.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 2.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
…	…	…	…	…	…	…	…	…	…	…	…
3	28	11	6	9	7	31	43	32	31	22	192
3	29	7	1	3	4	22	26	16	33	26	138
3	30	11	2	6	3	25	33	26	30	21	157

2.4. Análisis de regresión Logística multinomial

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 politómica, y un conjunto de variables explicativas que pueden ser tanto categóricas como cuantitativas.

Sea **x** un vector de p variables independientes, esto es, $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. En este caso la variable respuesta es el corpus al cual pertenece el texto y presenta 3 categorías. Si se define al corpus Biometría como la categoría de referencia, los logits generalizados compararán cada uno de los otros dos corpus con el de referencia. Asignando Y=0 al corpus de Biometría (referencia), Y=1 al corpus de Filosofía y por último Y=2 al de Lingüística informática, las dos funciones logit se expresan de la siguiente manera:

$$g_1(x) = \ln\left(\frac{P(Y=1/x)}{P(Y=0/x)}\right) = \beta_{10} + \beta_{11}x_1 + \ldots + \beta_{1p}x_p$$

$$g_2(x) = \ln\left(\frac{P(Y=2/x)}{P(Y=0/x)}\right) = \beta_{20} + \beta_{21}x_1 + \ldots + \beta_{2p}x_p$$

donde

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.

La probabilidad condicional de que la variable *y* tome el valor j (para j=1,2), dado valores de las covariables **x** es:

$$P(y = j/\mathbf{x}) = \pi_j(\mathbf{x}) = \frac{e^{g_j(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

y para la categoría de referencia es

$$P(y = 0/\mathbf{x}) = \pi_0(\mathbf{x}) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

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 dos niveles de la variable respuesta (categoría j versus categoría de referencia) se denomina odds y se expresa como:

$$\frac{P(y=j|\mathbf{x})}{P(y=0|\mathbf{x})} = e^{\beta_{j0}+\beta_{j1}x_1+\ldots+\beta_{jp}x_p} \qquad j=1,2$$

Si se aplica el logaritmo natural, se obtienen los logits generalizados:

$$\log\left(\frac{P(y=j|\mathbf{x})}{P(y=0|\mathbf{x})}\right) = \log\left(e^{\beta_{j0}+\beta_{j1}x_1+\ldots+\beta_{jp}x_p}\right) =$$

$$= \beta_{j0}+\beta_{j1}x_1+\beta_{j2}x_2+\ldots+\beta_{jp}x_p \qquad j=1,2$$

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 $\boldsymbol{\beta}' = (\beta_{10}, \beta_{11}, \beta_{12}, ..., \beta_{1p}, \beta_{20}, \beta_{21}, \beta_{22}, ..., \beta_{2p})$ por el método de Máxima Verosimilitud.

Las ecuaciones a resolver se obtienen derivando la función de verosimilitud respecto a cada uno de los parámetros del modelo e igualando a cero. Las soluciones de estas ecuaciones son los estimadores máximo verosímiles de cada uno de los componentes del vector de parámetros. 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_{jk} estimados representan tasa de cambio de una función de la variable dependiente *y* por unidad de cambio de la variable independiente x_k.

El coeficiente β_{ik} expresa el cambio resultante en la escala de medida de la variable *y* para un cambio unitario de la variable x_k. Por ejemplo, para la variable X_k, $\beta_{j1} = g(x_k+1) - g(x_k)$ representa el cambio en el logit, correspondiente a la categoría Y=j versus la categoría de referencia Y=0, frente a un incremento de una unidad en la variable X_k. La interpretación se hace en términos de la razón de Odds (OR).

$$OR = \frac{\left(\frac{P(Y=j/\mathrm{x_k}+1)}{P(Y=0/\mathrm{x_k}+1)}\right)}{\left(\frac{P(Y=j/\mathrm{x_k})}{P(Y=0/\mathrm{x_k})}\right)} = \frac{e^{\beta_{j0}+\beta_{j1}x_1+\ldots+\beta_{jk}(x_k+1)+\ldots+\beta_{jp}x_p}}{e^{\beta_{j0}+\beta_{j1}x_1+\ldots+\beta_{jk}x_k+\ldots+\beta_{p}x_p}} = e^{\beta_{jk}}$$

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 la estadística del cociente o razón de verosimilitud. 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 Kruskal Wallis, arrojando una probabilidad asociada $p=0.16$, 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 3). El número de clíticos es mayor en los textos de biometría y el número de adverbios es superior en los textos de filosofía.

Tabla 3. Comparación mediante test de Kruskal Wallis

Número promedio de:	BIOMETRIA	FILOSOFIA	LINGÜÍSTICA INFORMÁTICA	Valor de p
adjetivos	17,9	21,3	11,1	**0.0031**
adverbios	2,9	5,9	2,33	**0.0007**
clíticos	4,1	2,7	2,44	**0.0072**
copulativos	4,7	6,0	4,0	**0.0122**
determinantes	26,8	32,4	20,9	**0.0031**
nombres	44,6	45,0	30,2	**0.0010**
preposición	30,0	29,7	21,5	**0.0077**
verbos	16,1	18,4	24,0	**0.2592**
otro	18,8	21,4	16,7	**0.6324**
TOTAL_PALABRAS	165,8	182,9	155,1	**0.1664**

3.2. Análisis de Regresión Logística multinomial

Se realizó un análisis de regresión logística multinomial para obtener una regla de clasificación que permita asignar los textos en estas tres poblaciones, definidas por el área científica a la que pertenecen, en base a la frecuencia de cada categoría gramatical en el texto.

La selección del modelo se llevó a cabo mediante el procedimiento backward. El modelo final, cuyos coeficientes estimados se presentan en la tabla 4, evidenció un buen ajuste (Razón de verosimilitud=106,83 p=0.99). Los efectos incorporados en el modelo son:

- Número de adverbios
- Número de nombres
- Número de determinantes
- Número de clíticos
- Número de verbos
- Interacción verbos*clíticos

$$g_j(x) = \ln\left(\frac{P(Y=j/x)}{P(Y=0/x)}\right) = \beta_{j0} + \beta_{j1}adv + \beta_{j2}nom + \beta_{j3}\,\det + \beta_{j4}cl + \beta_{j5}v + \beta_{j6}cl*v$$

para j=1,2.

Tabla 4: Coeficientes del modelo de regresión logística multinomial

Efecto	Parámetro (j)	Estimador	Error estándar	Est. Chi-cuadrado	Prob. asociada
Intercepto	**1**	5.4082	2.2028	6.03	0.0141
	2	6.1627	2.7743	4.93	0.0263
adv	**3**	0.3610	0.1707	4.47	0.0345
	4	-0.1713	0.2170	0.62	0.4298
nom	**5**	-0.0855	0.0496	2.98	0.0844
	6	-0.1526	0.0544	7.87	0.0050
det	**7**	0.1195	0.0681	3.08	0.0792
	8	-0.1358	0.0906	2.25	0.1340
cl	**9**	-1.6551	0.5501	9.05	0.0026
	10	-1.2251	0.6580	3.47	0.0626
v	**11**	-0.2650	0.1041	6.48	0.0109
	12	0.1104	0.1293	0.73	0.3935
v*cl	**13**	0.0588	0.0220	7.15	0.0075
	14	0.0565	0.0276	4.18	0.0408

Este modelo permite, mediante la utilización de los coeficientes estimados, calcular para cada texto la probabilidad de pertenecer a cada uno de los corpus. 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 14% (Tabla 5).

Tabla 5: Tasa de error estimada

Tasa de error por corpus				
	BIOMETRIA	**FILOSOFIA**	**LINGÜÍSTICA**	**Total**
Tasa	16%	8%	17%	13.7%

Los coeficientes del modelo de regresión logística permiten la interpretación de la misma. Las categorías gramaticasles útiles para la discriminación de las áreas científicas a la que pertenecen los textos son: el número de adverbios, determinantes, nombres, clíticos y verbos.

Para los primeros efectos mencionados se estima que:

- La chance de clasificar a un texto dentro del corpus de Filosofía versus Biometría se incrementa en un 43% al aumentar en número de adverbios en una unidad, mientras que la chance de clasificarlo en el corpus de Biometría versus Lingüística aumenta un 18% al incrementarse en una unidad el número de adverbios.
- La chance de clasificar a un texto dentro del corpus de Biometría versus Lingüística se

incrementa en un 16% al aumentar en número de nombres en una unidad. El número de nombres no discrimina los corpus de Biometría y Filosofía.

- La chance de clasificar a un texto dentro del corpus de Filosofía versus Biometría se incrementa en un 11% al aumentar en número de determinantes en una unidad, mientras que la chance de clasificarlo en el corpus de Biometría versus Lingüística aumenta un 15% al incrementarse en una unidad el número de determinantes.

Sin embargo es importante notar que el modelo presenta una interacción clítico*verbo. Esto significa que el efecto del número de clíticos dependerá de la cantidad de verbos que tenga el texto.

- La chance de clasificar un texto dentro del corpus de Filosofía versus Biometría respecto al número de clíticos se ve afectado por el número de verbos (interacción). Cuando la frecuencia de verbos en el texto es superior al 15%, la chance de clasificar en Filosofía versus Biometría se incrementa con el número de clíticos. Sin embargo, cuando el texto presenta una frecuencia baja de verbos, el efecto es inverso. Con respecto a la clasificación en Lingüística versus Biometría la situación es la misma.

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 multinomial aplicado en este trabajo presenta una generalización de esta modalidad de análisis estadístico para discriminar más de dos grupos. El mismo permitió hallar las categorías gramaticales cuyas frecuencias observadas en los textos permiten discriminar los tres grupos definidos por la disciplina a la que pertenecen.

Las diferencias entre los dos tipos de textos está centrada principalmente en el número de adverbios, nombres, determinantes, clíticos y verbos presentes.

Los textos de Filosofía presentan, respecto a los de Biometría, una mayor cantidad de adverbios y una mayor cantidad de determinantes. Con respecto al número de clíticos, la chance de clasificar al texto en Biometría se incrementa con el número de clíticos presentes siempre y cuando el texto presente una proporción de verbos superior al 15%.

Los textos de Lingüística Informática presentan, respecto a los de Biometría, una menor cantidad de adverbios, una menor cantidad de nombres y una menor cantidad de determinantes. Con respecto al número de clíticos, se observa el mismo comportamiento que para el grupo de Filosofía, la chance de clasificar al texto en Biometría se incrementa con el número de clíticos presentes siempre y cuando el texto presente una proporción de verbos superior al 15%.

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.

Stokes, M. E., Davis, C.S., Koch, G.G. 1999 Categorical Data Analysis using SAS® System. WA (Wiley-SAS).