Análisis discriminante aplicado a textos académicos: Biometría y Filosofía

Discriminant Analysis Applied to Academic Texts: Biometrics and Philosophy

Celina Beltrán

Universidad Nacional de Rosario
Facultad de Ciencias Agrarias
Rosario, Argentina
beltranc@dat1.net.ar

Abstract

The aim of this work is to carry out an automatic analysis of academic texts from different scientific fields: Biometrics and Philosophy. The resulting information of the morphological analysis of these texts is used to shape a database on which the discriminant analysis technique is applied. This study allows for an analysis which makes it clear those characteristics which discriminate the text corpora in study. The linear discriminant function is mainly determined by two morphosyntactic categories: adverbs and clitics. The global error rate estimated by cross validation is 20%.

Keywords: discriminant analysis – multivariate analysis – automatic analysis of texts

Resumen

Este trabajo se propone la realización del análisis automático de textos académicos provenientes de distintas áreas científicas: Biometría y Filosofía. La información resultante del análisis morfológico de dichos textos es utilizada para conformar una base de datos sobre la cual se aplica la técnica de análisis discriminante. Este estudio permite un análisis en el cual se evidencian aquellas características que discriminan los corpus de textos en estudio. La función lineal discriminante está determinada principalmente por dos categorías morfosintácticas: adverbios y clíticos. La tasa de error global estimada por validación cruzada es del 20%.

Palabras claves: Análisis discriminante, análisis multivariado, análisis automático de textos.

1. INTRODUCCION

Este trabajo se propone la realización del análisis automático de textos académicos provenientes de distintas áreas científicas: Biometría y Filosofía. Se recurre al analizador morfológico Smorph, implementado como etiquetador, para asignar categoría a todas las ocurrencias lingüísticas.

La información resultante del análisis morfológico de dichos textos es utilizada para conformar una base de datos sobre la cual se aplica un de análisis discriminante. Este tipo de análisis es una herramienta útil para construir una regla de clasificación de unidades en varias poblaciones considerando un gran número de variables medidas sobre ellas.

Este estudio permite hallar las características 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 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”.

Luego de obtener las muestras de los dos estratos, fueron evaluadas y comparadas respecto al número medio de palabras por texto. Se requiere esta evaluación para evitar que la comparación entre las disciplinas se vea afectada por el tamaño de los textos.

La muestra final quedó conformada de la siguiente manera:

Tabla 1. Conformación de la muestra final

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

2.2. Etiquetado 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.

En el archivo **modelos**, se introduce la información correspondiente a los modelos de flexiones morfológicas. A un conjunto de terminaciones se le asocia el correspondiente conjunto de definiciones morfológicas. El esquema para definir los modelos es el siguiente:

<nombre_modelo> -<cantidad de caracteres a sustraer>

<terminación 1> <definición morfológica para terminación 1>

<terminación 2> <definición morfológica para terminación 2>

...

<terminación k> <definición morfológica para terminación k>

Se declara en primer lugar el nombre del modelo, luego la cantidad de caracteres que hay que sustraer a la forma lematizada. En tercer lugar se consigna la terminación, declarada previamente en el archivo terminaciones. La declaración morfológica corresponde a una cadena de caracteres sin espacios en blanco.

En el archivo **terminaciones** es necesario declarar todas las terminaciones que son necesarias para definir los modelos de flexión, se declaran una a continuación de otra, separadas por un punto.

Para construir los modelos se recurre a rasgos morfológico- sintácticos. En el archivo **rasgos**, se organizan jerárquicamente las etiquetas. En el archivo de códigos **ASCII** se especifican, entre otros, los caracteres separadores, las equivalencias entre mayúsculas y minúsculas. El archivo “data”, contiene los nombres de cada uno de los cinco archivos descriptos anteriormente.

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).

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.

En Beltrán (2009) se presentó una función definida en el sistema estadístico R que logra captar la información resultante del análisis morfológico y la dispone 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 que retiene la información del texto, ocurrencia, lema y etiqueta asignada, como muestra 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
…	…	…	…	…
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	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 Discriminante de Fisher

2.4.1. El método

El análisis discriminante de Fisher considera el problema de obtener una función discriminante cuando se tienen 2 poblaciones. El objetivo es hallar una regla de clasificación lineal mediante la cual se diferencien lo más posible las dos poblaciones en estudio. Esta técnica asume matrices de variancias y covariancias iguales para las dos poblaciones Sea π_t una población con distribución normal multivariada p-dimensional con vector de medias poblacional $\boldsymbol{\mu}_t$, con t=1,2, y matriz de variancias y covariancias $\boldsymbol{\Sigma}$ y sea **x** una observación p-dimensional a ser clasificada en una de las 2 poblaciones. Si **a**'**x** es una regla de clasificación lineal en **x** (una combinación lineal de los componentes de **x**) se pretende hallar el vector **a** tal que maximice la distancia entre las dos poblaciones, es decir, el vector **a** para el cual la distancia entre E(**a**'**x**) en π_1 y E(**a**'**x**) en π_2 sea máxima. Puesto que la distancia en cuestión puede ser aumentada multiplicando el vector **a** por una constante positiva, es apropiado eliminar esta ambigüedad estableciendo una condición adicional sobre **a**. Se adiciona entonces la condición que **a** sea estandarizado y que **a**'**x** tenga variancia unitaria. Esto es:

$$var\ (\mathbf{a}'\mathbf{x}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = 1$$

De esta manera, el método de Fisher consiste en elegir un vector **a** que maximice la distancia entre $\mathbf{a}'\boldsymbol{\mu}_1$ y $\mathbf{a}'\boldsymbol{\mu}_2$, $|\ \mathbf{a}'\boldsymbol{\mu}_1 - \mathbf{a}'\boldsymbol{\mu}_2\ |$, sujeto a la condición $\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a} = 1$.

La teoría de álgebra matricial establece que la elección óptima de **a** debe ser proporcional a $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Como el objetivo es clasificación, ignoramos la constante de proporcionalidad y **a** puede ser tomado como:

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

De esta manera, se clasifica una observación **x** en π_1 si $\mathbf{a}'\mathbf{x} \geq h$ y en π_2 de otra forma, donde $h = \mathbf{a}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Esta elección de h es esencialmente igual a elegir una regla la cual clasifica a **x** en la más cercana de las dos poblaciones cuando consideramos la distancia euclídea.

Puesto que se trabaja con datos muestrales, la matriz $\boldsymbol{\Sigma}$ será reemplazada por su estimador **S** y el vector de medias $\boldsymbol{\mu}_t$ (t=1,2) por el vector de medias muestrales $\bar{\mathbf{x}}_t$ (t=1,2).

2.4.2. Estimación de la tasa de error

La tasa de error estimada puede ser utilizada para evaluar el esquema de discriminación elegido.

Si se tiene un conjunto de datos con varias poblaciones y el número de observaciones en cada una de ellas es conocido, el número de observaciones de la t-ésima población clasificados erróneamente en la s-ésima población puede proporcionar alguna idea acerca de la probabilidad de clasificar una unidad en la población s cuando en realidad esta unidad pertenece a la población t, esto es P(s/t). La proporción de clasificaciones erróneas desde la t-ésima población se puede expresar como:

$$\hat{ER}(t) = \sum_{s=1, s\neq t}^{k} \hat{P}(s/t)$$

Esta cantidad proporciona una estimación de la tasa de error para la t-ésima población.

No obstante, es importante remarcar que si el conjunto de datos utilizados para evaluar esta tasa de error es diferente de aquel usado para obtener la función discriminante, las estimaciones de la tasa de error son no viciadas. Sin embargo, si el comportamiento de la función discriminante está siendo examinado sobre el mismo conjunto de datos, las estimaciones son viciadas (subestiman la verdadera tasa de error) y por lo tanto estas estimaciones son excesivamente optimistas.

Lachenbruch (1975) sugiere un método alternativo para estimar la tasa de error a través de la validación cruzada. El mismo consiste en dejar una observación afuera y construir una regla discriminante con el resto de los datos. Esta regla se emplea para clasificar la observación que fue dejada afuera. Este procedimiento se repite para cada observación y finalmente se cuentan los números de observaciones mal clasificadas para cada población y se calculan las tasas de error individuales como las respectivas proporciones. La tasa de error global puede ser computada como el promedio ponderado de estas proporciones, utilizando como pesos las probabilidades a priori. Este procedimiento conduce a estimaciones que tienen un vicio considerablemente más pequeño.

2.4.3. Selección de variables en Análisis discriminante:

Una cuestión importante en el Análisis Discriminante 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 las poblaciones. Dado que las medidas registradas sobre la misma unidad es probable que estén correlacionadas, es probable también que compartan información acerca de los miembros de la población. 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. Necesitamos un procedimiento que pueda ser usado para decidir cual subgrupo particular de variables debería ser usado para discriminación y cuales variables pueden ser consideradas redundantes en presencia de ese subgrupo. Existen varios algoritmos de selección:

Método forward: comenzar por seleccionar la variable más importante y continuar 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: comenzar con el modelo más grande posible. En cada paso descartar la variable menos importante, una por vez, usando un criterio similar a la selección forward. Continuar hasta que ninguna de las variables pueda ser descartada.

Selección paso a paso: combinar los 2 procedimientos anteriores. En un paso, una variable puede entrar o salir desde la lista de variables importantes, de acuerdo a cierto criterio pre-establecido para estas selecciones.

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
nombres	44,6	45,0	**44,8**	0,55400
verbos	16,1	18,4	**17,2**	0,85882
TOTAL_PALABRAS	165,8	182,9	**174,4**	0,79578

En cada caso la distribución de las variables se alejó significativamente de la Normal ($p<0.05$).

3.2. Análisis de Discriminante

Se realizó un análisis discriminante de Fisher para obtener una función lineal que permita clasificar 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. Para lograr variables con distribución Normal y matriz de variancias y covariancias común a los dos conjuntos de textos, supuestos requeridos para aplicar la técnica descripta, se recurre a una transformación de las variables. En este caso se trabajó con los logaritmos de las proporciones de cada categoría morfosintáctica, dado que fue la más adecuada para lograr el cumplimiento de estos supuestos:

$$v_j = \log(x_j),\ \text{j=1,2,3,4,5}$$

donde x_j es la proporción de la categoría j en el texto.

Se evaluó la normalidad de las variables transformadas de acuerdo a lo trabajado en Beltrán (2010) hallando un ajuste satisfactorio en todos los casos ($p>0.05$).

Considerando todas las categorías se obtiene la siguiente función lineal discriminante:

Tabla 5: Coeficientes de la función lineal discriminante

Variable	Coeficiente
Constante	13.28
Log(prop_adj)	0.08
Log(prop_adv)	-4.56
Log(prop_cl)	6.71
Log(prop_nom)	4.00
Log(prop_v)	1.30

Esta función es usada como regla de clasificación para un texto de la siguiente manera:

- si $\mathbf{a}'\mathbf{v} \geq 0$ entonces el texto correspondiente pertenece al corpus de Biometría, en caso contrario pertenece al corpus humanístico, siendo

$$a'v = 13.28 + 0.08 \times \log(prop\_adj) - 4.56 \times \log(prop\_adv) + 6.71 \times \log(prop\_cl) + \\ + 4.00 \times \log(prop\_nom) + 1.30 \times \log(prop\_v)$$

Aplicando esta regla de clasificación y estimando por validación cruzada, la tasa de error global que se obtiene es del 21.7% (Tabla 6).

Tabla 6: Tasa de error estimada

Tasa de error por corpus			
	BIOMETRIA	**FILOSOFIA**	**Total**
Tasa	20%	23.3%	21.7%

Para determinar cuáles categorías gramaticales eran 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. La función lineal discriminante luego de la selección de variables se muestra en la tabla 7.

Esta función es usada como regla de clasificación para un texto de la siguiente manera:

- si **a**'**v** ≥ 0 entonces el texto correspondiente pertenece al corpus de Biometría, en caso contrario pertenece al corpus de Filosofía.

Esta función resultante expresada en término de estas dos variables seleccionadas es:

$$a'v = 4.42 - 5.06 \times \log(prop\_adv) + 6.65 \times \log(prop\_cl) .$$

Si el valor de esta función valorizada en un texto es positiva, entonces el texto se clasifica como perteneciente a biometría, en caso contrario se lo clasifica en el corpus de filosofía. Con la función lineal discriminante obtenida durante la selección de variables se obtiene una tasa de error global del 15% mediante validación cruzada (Tabla 8).

Tabla 7: Coeficientes de la función lineal discriminante final

Variable	**Coeficiente**
Constante	4.42
Log(prop_adv)	-5.06
Log(prop_cl)	6.65

Tabla 8: Tasa de error estimada

Tasa de error por corpus			
	BIOMETRIA	**FILOSOFIA**	**Total**
Tasa	16.7%	13.3%	15%

A modo de ejemplo, considérese el caso de los seis nuevos textos correspondiente a la tabla 9.

Tabla 9: Valores de las variables de 6 textos a clasificar.

CORPUS	**TEXTO**	**adverbios**	**clíticos**	**Log(prop_adv)**	**Log(prop_cl)**
B	T1	2	3	-2,80	-2,71
B	T2	1	4	-2,99	-2,54
B	T3	5	2	-2,55	-2,79
F	T4	3	1	-2,57	-2,99
F	T5	3	1	-2,57	-2,83
F	T6	4	5	-2,30	-2,18

Los valores de la función discriminante en cada uno de ellos se calculan y se muestran en la tabla 10, donde se observan dos textos mal clasificados, T3 y T6. La figura 1 visualiza la clasificación.

Tabla 10: Función discriminante valorizada en los nuevos textos.

CORPUS	TEXTO	adverbios	clíticos	Log(prop_adv)	Log(prop_cl)	F. Discriminante	Clasificado en:
B	T1	2	3	-2,80	-2,71	0,5533	B
B	T2	1	4	-2,99	-2,54	2,6684	B
B	T3	5	2	-2,55	-2,79	-1,2123	**F**
F	T4	3	1	-2,57	-2,99	-2,4599	F
F	T5	3	1	-2,57	-2,83	-1,3789	F
F	T6	4	5	-2,30	-2,18	1,5615	**B**

Los puntos marcados con una “estrella” corresponden a los dos textos mal clasificados. Se observa que para el texto 3 (perteneciente al corpus de Biometría), el valor de la función discriminante está más próximo al valor correspondiente al centro del corpus de Filosofía y para el texto 6, ocurre lo contrario, esto es, el valor de la variable discriminante para el texto 6 está cercano al valor del centro de Biometría (cuando pertenece al corpus de Filosofía).

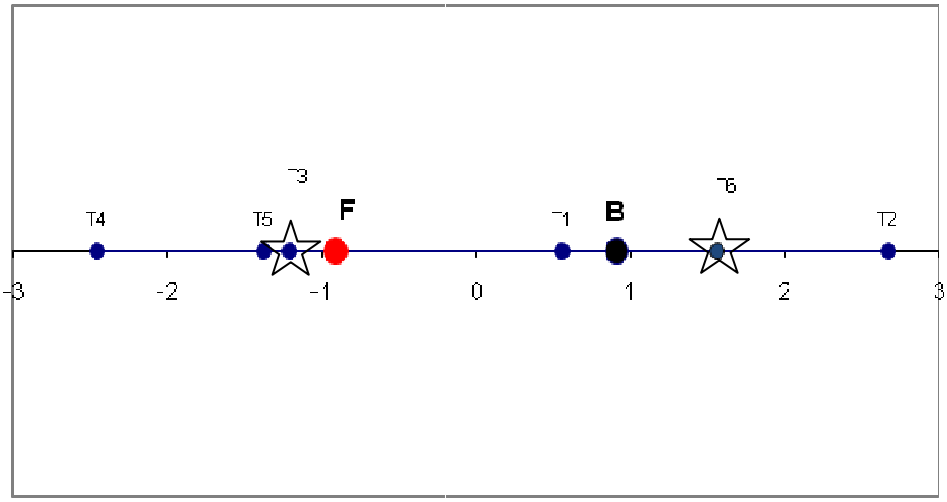

Figura 1: Representación gráfica de la clasificación mediante la función lineal discriminante

Los coeficientes de la función lineal discriminante 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. Un texto será clasificado dentro del área de Biometría cuando presente un número alto de clíticos y pocos adverbios. Por el contrario, un texto se clasificará como del área de Filosofía al presentar un número superior de adverbios y pocos clíticos.

La figura 2, muestra una simulación de textos pertenecientes a ambos corpus que fueron clasificados utilizando la función lineal discriminante seleccionada. Cada uno de ellos fue clasificado y “marcado” en el texto con diferentes colores para observar cómo se disponen estas dos poblaciones respecto de las variables incluidas en la discriminación. Se observa que valores altos para la frecuencia de clíticos y bajos para la frecuencia de adverbios, textos sobre la diagonal, son clasificados a Biometría (color negro) y textos ubicados por debajo de esta

diagonal, con baja frecuencia de clíticos y gran cantidad de adverbios, se clasifican en Filosofía (color rojo).

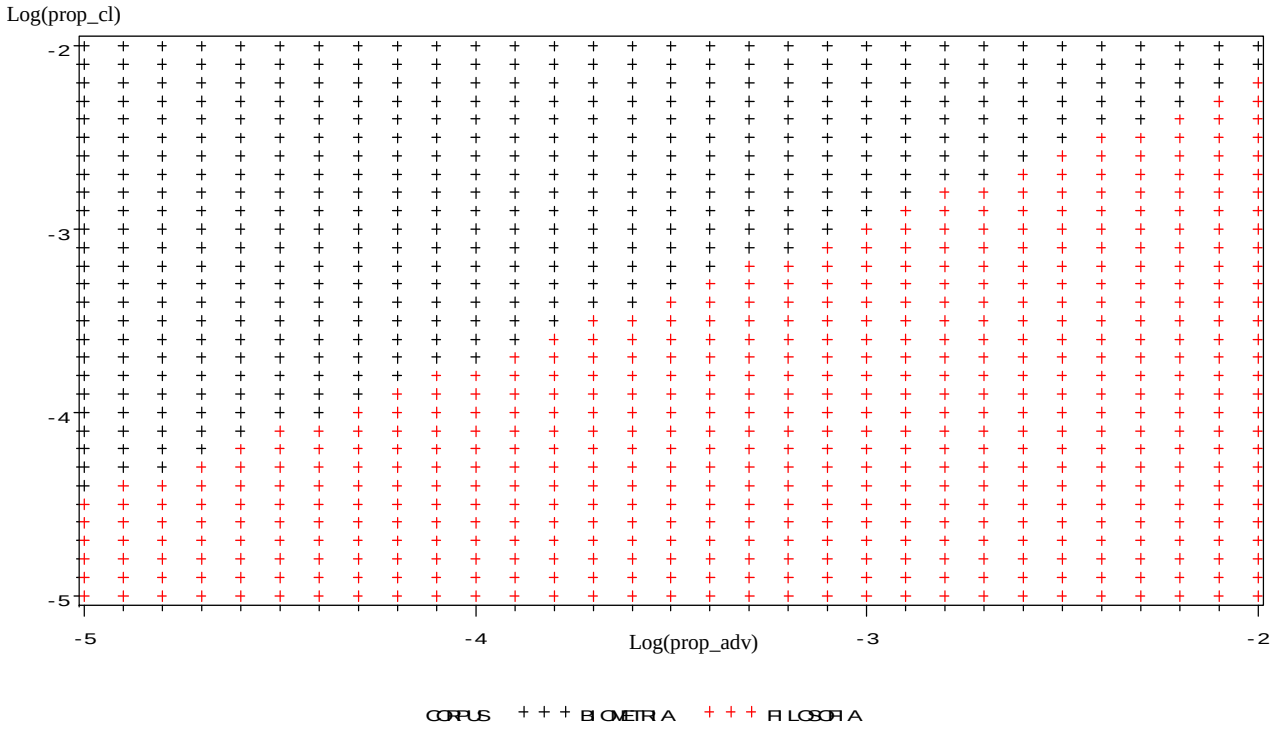

Figura 2: Simulación y clasificación de textos mediante la función lineal discriminante

4. CONCLUSIONES

Los resultados del análisis morfológico de los textos se analizaron teniendo en cuenta simultáneamente todas las mediciones realizadas sobre ellos.

El análisis multivariado aplicado en este trabajo presenta una modalidad de análisis estadístico no muy frecuente en la investigación lingüística. El mismo permitió hallar las características de los textos que discriminan los dos grupos definidos por la disciplina a la que pertenecen.

Las diferencias entre los dos tipos de textos se basan fundamentalmente en el número de clíticos y de adverbios presentes. Un texto será clasificado dentro del área de Biometría cuando presente un número alto de clíticos y pocos adverbios. Por el contrario, un texto se clasificará como del área de Filosofía al presentar un número superior de adverbios y pocos clíticos.

Esto puede deberse a que, en los textos de biometría 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 lineal”

“se estima el promedio poblacional”

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

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.

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.

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

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.