Estudio exploratorio para la comparación de distintos tipos de textos: Textos Científicos y Textos No Científicos

Exploratory Study to Compare Different Types of Texts: Scientific Texts and Non Scientific Texts

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

Abstract

This work proposes the accomplishment of automatic analysis of scientific and non scientific texts. The scientific texts correspond to abstracts published in scientific magazines and proceedings of conferences in several academic disciplines; the non scientific texts correspond to newspaper reports on general subjects published by online Argentine newspapers. The incoming information of morphological analysis of the text is employed to arrange a database to which the principal component technique is applied. The study allows an exploratory analysis that shows the characteristics discriminated by the corpora of the text under consideration. The first three components give an explanation of the 75% of the total variation for data. The number of adverbs in the text is the main variable of such separation.

Keywords: principal components – text comparison – automatic analysis of text.

Resumen

Este trabajo se propone la realización del análisis automático de textos científicos y no científicos. Los textos científicos corresponden a resúmenes de publicaciones en revistas científicas y actas de congresos de distintas disciplinas y los textos no científicos corresponden a noticias periodísticas de interés general publicadas en páginas web de periódicos argentinos. 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 componentes principales. Este estudio permite un análisis exploratorio en el cual se evidencian las características que discriminan los corpus de textos en estudio. Las tres primeras componentes explican un 75% de la variación total de los datos. El número de adverbios en el texto es la variable de mayor importancia en dicha separación.

Palabras claves: Componentes principales, comparación de textos, análisis automático de textos.

1. INTRODUCCION

Este trabajo se propone la realización del análisis automático de distintos tipos de textos: científicos (C) y no científicos (NC). Los textos científicos corresponden a resúmenes de publicaciones en revistas científicas y actas de congresos de distintas disciplinas y los textos no científicos corresponden a noticias periodísticas de interés general publicadas en páginas web de periódicos argentinos. 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 la técnica de componentes principales. Este tipo de análisis es una herramienta útil para caracterizar unidades considerando un gran número de variables cuantitativas medidas sobre ellas. En este caso es un estudio exploratorio que permite caracterizar los textos provenientes de estos dos géneros de modo tal de evidenciar aquellas características provenientes del análisis automático de los textos que son más discriminatorias con el propósito de lograr limitar el número de mediciones retenidas en estudios posteriores.

2. MATERIAL Y METODOS

2.1. Diseño de la muestra

El marco muestral para la selección de la muestra de los textos científicos está compuesto por textos académicos, resúmenes de trabajos presentados a congresos y revistas, extraídos de internet pertenecientes a distintas disciplinas. Los textos periodísticos fueron seleccionados de un corpus mayor utilizado por el equipo de investigación INFOSUR. Este corpus se construyó con noticias extraídas de las páginas web de periódicos argentinos (noticias de tipo general, no especializadas en español). La unidad de muestreo fue el texto y la selección de la muestra se llevó a cabo empleando un diseño muestral estratificado.

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 para este trabajo quedó conformada de la siguiente manera:

Tabla 1. Conformación de la muestra final

Muestra	Nro. de textos	Cantidad de palabras
Científico	90	14.554
No científico	60	8.080

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. 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

La información que contiene la base de datos es el resultado del análisis de Smorph-Mps almacenada en un archivo de texto. La información resultante del análisis morfológico se dispuso 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 obtuvo una base de datos que posee la información del texto, ocurrencia, lema y etiqueta asignada, como muestra la tabla 2.

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

Tabla 2. Fragmento de la base de datos obtenida

CORPUS	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	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

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 Componentes principales

Análisis de Componentes Principales (ACP) es una de las técnicas multivariadas de análisis exploratorio de datos más ampliamente utilizada, introducida por Pearson en 1901 y posteriormente desarrollada por Hotelling en 1933. Esta técnica también puede ser vista como un caso particular de los métodos de búsqueda de proyección, los cuales seleccionan proyecciones de poca dimensión de datos multivariados. La selección de proyecciones de poca dimensión usualmente se realiza optimizando algún índice que mida una característica de interés en los datos bajo todas las direcciones de proyección. Para el caso del Análisis de Componentes Principales la característica que se optimiza es la variancia de los datos.
Supongamos que tenemos una población y un vector aleatorio de dimensión px1 que puede ser medido sobre todos los individuos de la población:

$$\mathbf{x} = (x_1\ x_2\ .....\ x_p)'$$

Sea **Σ** la matriz de variancias-covariancias (de tamaño pxp) de p variables x_1, x_2, ..., x_p:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & . & . & . & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & . & . & . & \sigma_{2p} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ \sigma_{p1} & \sigma_{p2} & . & . & . & \sigma_{pp} \end{bmatrix}$$

La variancia total de estas variables es definida como la traza de **Σ** (*tr* **Σ**), la cual es la suma de los elementos de la diagonal principal de la matriz **Σ**:

$$tr\Sigma = \sum_{i=1}^{p} \sigma_{ii}$$

La primer componente principal de un vector **x** = (x_1 x_2 ..... x_p)' de dimensión px1 es una combinación lineal:

$$\mathbf{a}_1'\mathbf{x} = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p$$

donde $\mathbf{a}_1' = (a_{11}\ a_{12}\ .....\ a_{1p})'$ con $\mathbf{a}_1'\mathbf{a}_1 = 1$

y tal que la variancia de ($\mathbf{a}_1'\mathbf{x}$) es la máxima entre todas las combinaciones posibles de los elementos de **x** con los coeficientes del vector $\mathbf{a}_1$ cumpliendo la condición de longitud igual a 1. Por lo tanto, la primer componente principal así obtenida explica la máxima variación. La segunda componente principal $\mathbf{a}_2'\mathbf{x}$ de **x** con $\mathbf{a}_2'\mathbf{a}_2 = 1$ es tal que no está correlacionada con la primera y su variancia es la mayor entre todas las combinaciones lineales no correlacionadas con la primera componente. En forma similar se obtienen las restantes. La última componente principal (p-ésima) $\mathbf{a}_p'\mathbf{x}$ no está correlacionada con todas las (p-1) componentes principales anteriores y es la que menos explica de la variancia total, por este motivo se dice que es la componente menos informativa.

Los coeficientes de estas combinaciones lineales (las componentes principales) se obtienen de la siguiente manera:

Sean $\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_p > 0$ los autovalores y $\mathbf{a}_1\ \mathbf{a}_2\ .....\ \mathbf{a}_p$ los correspondientes autovectores de **Σ**, cada uno de ellos con longitud igual a 1, esto es $\mathbf{a}_i'\mathbf{a}_i = 1$, para i = 1,2, ..., p. La primer componente será $y_1 = \mathbf{a}_1'\mathbf{x}$, la segunda componente es $y_2 = \mathbf{a}_2'\mathbf{x}$ y la p-ésima componente está dada por $y_p = \mathbf{a}_p'\mathbf{x}$. Asimismo, las variancias de las componentes principales serán cada uno de los autovalores, $var(y_1) = \lambda_1$ $var(y_2) = \lambda_2$ ..... $var(y_p) = \lambda_p$. Por lo tanto, dado que la variancia total es la traza de la matriz **Σ** (tr **Σ**) es también la misma que la suma de todos sus autovalores ($\sum_{i=1}^{p} \lambda_i$).

Los elementos de los autovectores tienen interpretaciones útiles para el análisis. Por ejemplo, la covariancia entre la i-ésima variable x_i y la j-ésima componente principal y_j es $\lambda_j a_{ji}$ y por lo tanto el coeficiente de correlación corr (x_i, y_j) entre ellos es:

$$corr(x_i, y_j) = a_{ji}\sqrt{\frac{\lambda_i}{\text{var}(x_i)}}$$

Esto significa que las variables con coeficientes de gran magnitud en una componente principal tienen una mayor contribución en dicha componente.

Si bien no es equivalente, es posible trabajar con la matriz de correlaciones en lugar de la matriz de variancias y covariancias. Esto será apropiado cuando las medidas sobre diferentes variables no estén en la misma escala y las variancias sean de magnitudes muy diferentes.

Al trabajar con datos muestrales, las matriz **Σ** es reemplazada por sus respectivo estimador **S**.

Cuando se trabaja con datos composicionales, los porcentajes de los elementos para cada muestra suman 100 (ó 1 si son proporciones) y por lo tanto hay una restricción entre las medidas de las variables. Por este motivo es que se debe tener cuidado al analizar estos datos.
Supongamos que x_1, x_2, ..., x_p son las medidas (ó porcentajes) tomados sobre p variables, con $\sum_{i=1}^{p} x_i = 100$ (ó = 1). Debido a esta restricción, exactamente uno de los autovalores de la matriz de variancias-covariancias de $x = (x_1\ x_2\ .....\ x_p)'$ será cero. Esto hace que la interpretación usual de las variancias y las covariancias se pierde. Por lo tanto, Aitchison (1983) sugirió que el Análisis de Componentes Principales esté basado en la matriz de variancias-covariancias muestral de los p logaritmos-contrastes de las variables originales:

$$v_j = \log(x_j) - (\frac{1}{p})\sum_{i=1}^{p}\log(x_i),\ \text{j = 1, 2, ... , p}$$

en lugar de estar basado en la matriz de variancias-covariancias muestral de los porcentajes originales.
En este trabajo se aplica ACP para datos composicionales sobre las transformaciones mencionadas para las variables siguientes:

Tabla 4. Variables utilizadas en ACP

CORPUS	**CORPUS**	Corpus al que pertenece el texto (CIENTIFICO - NO CIENTIFICO)
TEXTO	**TEXTO**	Identificador del texto dentro del corpus
x1	**adj**	porcentaje de adjetivos del texto
x2	**adv**	porcentaje de adverbios del texto
x3	**cl**	porcentaje de clíticos del texto
x4	**cop**	porcentaje de copulativos del texto
x5	**det**	porcentaje de determinantes del texto
x6	**nom**	porcentaje de nombres (sustantivos) del texto
x7	**prep**	porcentaje de preposiciones del texto
x8	**v**	porcentaje de verbos del texto
x9	**OTRO**	porcentaje de otras etiquetas del texto

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=0.0062, evidenciando que existen diferencias significativas entre los corpus respecto al tamaño de los textos. Esta situación lleva a realizar las sucesivas comparaciones sobre los porcentajes o proporciones de las categorías gramaticales.

Comparaciones similares entre los corpus se llevan a cabo para las restantes variables (tomando las proporciones de cada categoría gramatical) hallando diferencias significativas ($p<0.05$) para todas

las categorías gramáticas excepto la proporción de clíticos y de verbos en los documentos analizados (Tabla 5).

Tabla 5. Comparación mediante test de Wilcoxon

Categoría Gramatical	Proporción promedio en corpus Científico	Proporción promedio en corpus No Científico	Valor de p
adjetivos	0,09	0,07	0,0001
adverbios	0,02	0,04	<0,0001
clíticos	0,02	0,02	0,2365
copulativos	0,03	0,02	0,0011
determinantes	0,16	0,15	0,0008
nombres	0,24	0,26	0,0109
preposición	0,17	0,15	0,0117
verbos	0,14	0,14	0,3195
otro	0,12	0,15	<0,0001

3.2. Análisis de Componentes principales

El ACP se realizó sobre la matriz de variancias y covariancias de las variables transformadas según lo establecido en la sección 2.4. para datos composicionales. En el ACP (Tabla 6) se puede observar que las tres primeras componentes explican un 75% de la variación total de los datos.

Tabla 6. Porcentaje de variancia explicada por las componentes principales

AUTOVALORES DE LA MATRIZ DE VARIANCIAS Y COVARIANCIAS			
CP	**Autovalor**	**% variancia explicada**	**% variancia explicada acumulado**
1	0,65	35%	35%
2	0,49	26%	61%
3	0,26	14%	75%
4	0,21	11%	86%
5	0,14	7%	93%
6	0,07	4%	97%
7	0,03	2%	99%
8	0,02	1%	100%
9	0	0%	100%

Tabla 7. Coeficientes correspondientes a las tres primeras componentes

Variable	**CP1**	**CP2**	**CP3**
adj_transf	-0,18	-0,24	0,67
adv_transf	0,82	-0,28	-0,12
cl_transf	0,02	0,86	0,06
cop_transf	-0,47	-0,22	-0,64
det_transf	-0,12	-0,11	0,11
nom_transf	-0,1	-0,14	0,03
prep_transf	-0,15	-0,11	0,16
v_transf	0,02	0,13	0,03
OTRO_transf	0,15	0,12	-0,29

Tabla 8. Correlaciones de las tres primeras componentes principales con las variables originales

Variable	**CP1**	**CP2**	**CP3**
adj_transf	-0,31	-0,37	0,73
adv_transf	0,93	-0,27	-0,09
cl_transf	0,02	0,97	0,05
cop_transf	-0,68	-0,28	-0,57
det_transf	-0,38	-0,31	0,21
nom_transf	-0,33	-0,42	0,06
prep_transf	-0,44	-0,29	0,29
v_transf	0,07	0,34	0,05
OTRO_transf	0,29	0,2	-0,35

La tabla 7 presenta los coeficientes de las variables sobre las tres primeras componentes, mientras que la tabla 8 presenta las correlaciones entre las componentes y las variables originales. La primera componente se caracteriza por presentar un coeficiente alto positivo para la variable referida a los adverbios y valores negativos para las variables referidas a las restantes categorías. Esto significa que esta primer componente (CP1) está altamente correlacionada con la proporción de adverbios en el texto y presentará valores altos cuando un texto presente con mayor frecuencia esta categoría. La segunda componente (CP2) presenta principalmente un coeficiente alto positivo para las variables de clíticos y verbos y negativo para las restantes. Esto significa que esta componente presentará valores altos cuando el texto evidencie una mayor cantidad de clíticos y verbos. Respecto a la tercer componente (CP3), ésta se asocia positivamente con la presencia de adjetivos y preposiciones en mayor frecuencia en el texto y se asocia negativamente con la frecuencia de copulativos.

Al proyectar los textos analizados sobre estas tres dimensiones o tres primeras componentes (Gráfico 1, 2 y 3), se observa que en el gráfico 1, los textos NC se encuentran desplazados a la derecha sobre el eje horizontal. Por la conformación de la primera componente, esta disposición en el plano de proyección evidencia que los textos procedentes del corpus NC presentan un mayor

número de adverbios, respecto a las restantes categorías, que los textos C. En relación a las otras dos dimensiones no se visualiza una clara discriminación de los textos respecto al género.

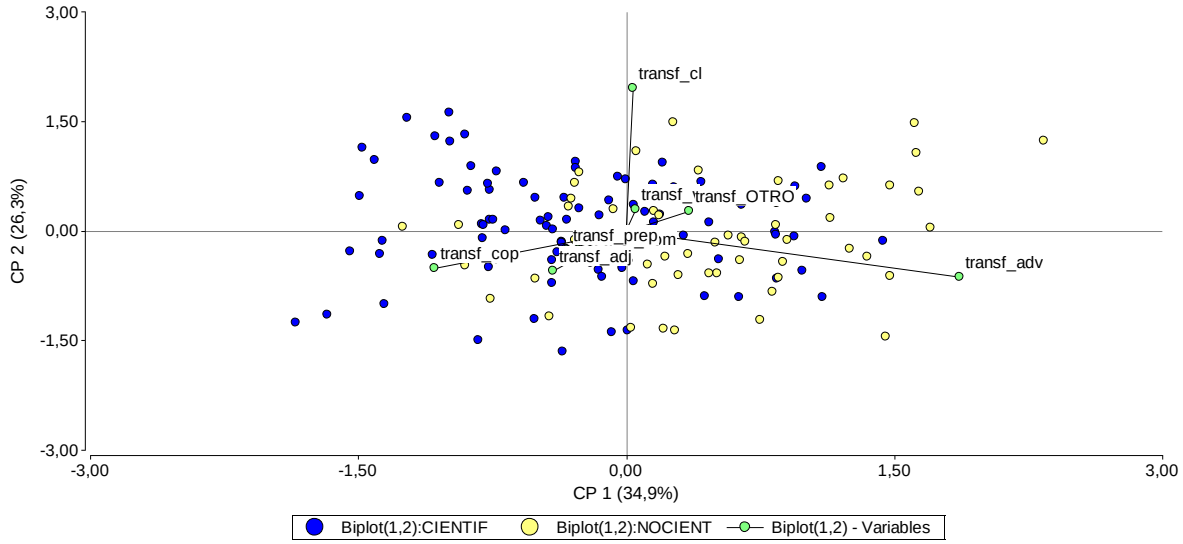

Gráfico 1: Proyección de los textos sobre las dos primeras componentes principales

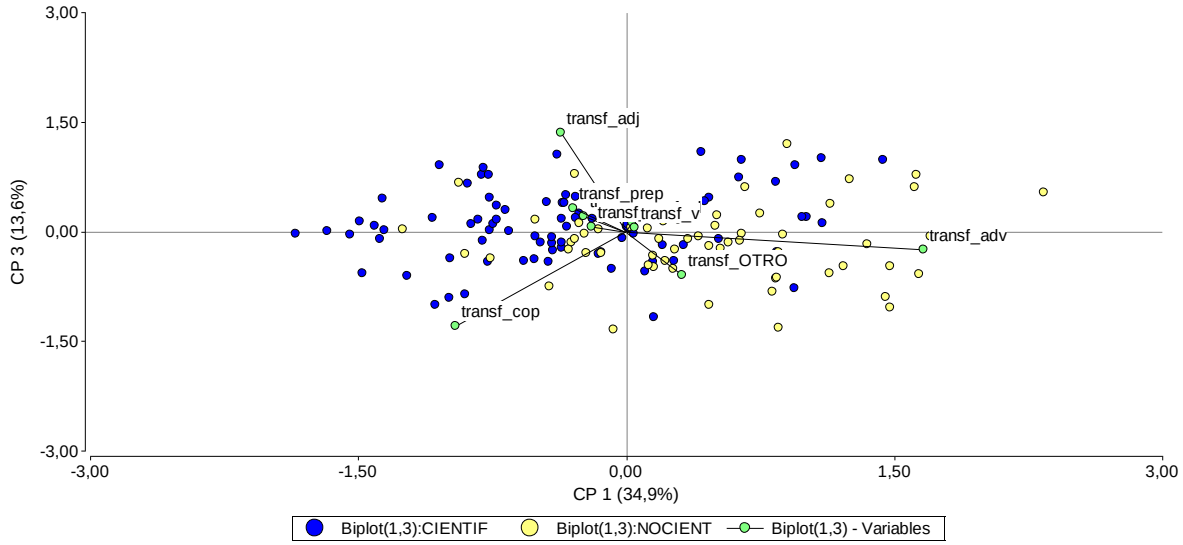

Gráfico 2: Proyección de los textos sobre la primera y tercer componente principal

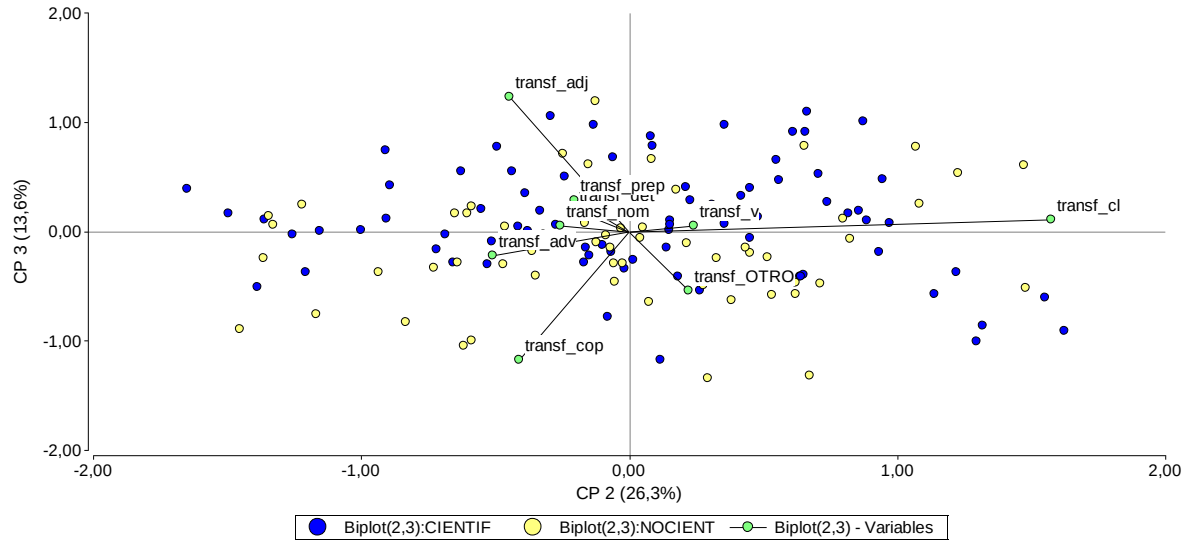

Gráfico 3: Proyección de los textos sobre la segunda y tercera componente principal

6. 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 permitió hallar las características de los textos que discriminan los dos grupos definidos por el género al que pertenecen: Científico y No Científico.

La proyección de los textos sobre gráficos bidimensionales fue muy útil para visualizar las diferencias halladas entre los dos corpus. Las tres primeras componentes explicaron un 75% de la variación total de los datos, observándose que la primera componente es la dimensión que separa los textos científicos y no científicos. El número de adverbios en el texto es la variable de mayor importancia en dicha separación.

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.

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.