Revista de Epistemología y Ciencias Humanas

Estudio y comparación de distintos tipos de textos académicos: Biometría y Filosofía

A COMPARISON STUDY OF DIFFERENT TYPES OF ACADEMIC TEXTS: BIOMETRICS AND PHILOSOPHY

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

Abstract

The aim of this work is to carry out 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 principal component technique is applied. This study allows an exploratory analysis which makes clear the characteristics that differentiate the text corpora in study. The two first components explained the 63% of the total data variation while the second component is the dimension that separates the texts of both disciplines. The number of the clitics and the adverbs in the text are the variables of main importance in that separation.

Keywords: Principal components; multivariate analysis; automatic text analysis.

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 componentes principales. Este estudio permite un análisis exploratorio en el cual se evidencian las características que discriminan los corpus de textos en estudio. Las dos primeras componentes explican un 63% de la variación total de los datos, observándose que la segunda componente es la dimensión que separa los textos de ambas disciplinas. El número de clíticos y el de adverbios en el texto son las variables de mayor importancia en dicha separación.

Palabras claves: Componentes principales, 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 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.

El análisis presentado en este trabajo es un estudio exploratorio que permite también hallar las características provenientes del análisis automático de los textos que son más discriminatorias para limitar el número de mediciones retenidas en caracterizaciones posteriores.

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
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 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 $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma}$ ($tr\ \mathbf{\Sigma}$), la cual es la suma de los elementos de la diagonal principal de la matriz $\mathbf{\Sigma}$:

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

La primer componente principal de un vector $\mathbf{x} = (x_1\ x_2\ .....\ x_p)'$ de dimensión px1 es una combinación lineal:

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

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

y tal que la variancia de $(\mathbf{a}_1'\mathbf{x})$ es la máxima entre todas las combinaciones posibles de los elementos de $\mathbf{x}$ 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 $\mathbf{x}$ 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 $\mathbf{\Sigma}$, 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 $\mathbf{\Sigma}$ (tr $\mathbf{\Sigma}$) 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:

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

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
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=0.796, 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 5). El número de clíticos es mayor en los textos de biometría y el número de adverbios es superior en los textos de filosofía.

Tabla 5. 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 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 dos primeras componentes explican un 63% 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	0,592	38,6%	38,6%
2	0,374	24,4%	63,1%
3	0,274	17,9%	80,9%
4	0,131	8,6%	89,5%
5	0,079	5,2%	94,6%
6	0,040	2,6%	97,3%
7	0,023	1,5%	98,8%
8	0,016	1,1%	99,8%
9	0,003	0,2%	100,0%

Tabla 7. Coeficientes correspondientes a las dos primeras componentes

Variable	CP1	CP2
adj_transf	-0,036	0,029
adv_transf	0,791	0,523
cl_transf	0,374	-0,736
cop_transf	-0,360	0,410
det_transf	-0,055	0,080
nom_transf	-0,101	-0,051
prep_transf	-0,133	-0,081

v_transf	0,204	-0,011
OTRO_transf	0,178	0,009

La tabla 7 presenta los coeficientes de las variables sobre las dos primeras componentes. La primer componente se caracteriza por coeficientes altos positivos para las variables referidas a los adverbios, clíticos y verbos; y valores negativos para la variable referida a copulativos. Esto significa que esta primer componente (CP1) presentará valores altos cuando un texto presente con mayor frecuencia adverbios, verbos y clíticos y en menor número copulativos. Si bien esta CP1 es la que explica un porcentaje mayor de variancia no es la dimensión en la que se observa la diferencia entre corpus. Esto ocurre en la segunda componente (CP2). La CP2 presenta principalmente un coeficiente alto positivo para la variable de adverbios y alto pero negativo para la variable referida a los clíticos. Al proyectar los textos analizados sobre estas dos dimensiones o primeras componentes (Gráfico 1), se observa que la segunda dimensión (eje vertical) discrimina a los textos dejando por encima del eje horizontal en su mayoría aquellos pertenecientes al corpus de filosofía y por debajo a los textos de biometría. Por la conformación de la segunda componente, esta disposición en el plano de proyección evidencia que los textos procedentes del corpus de biometría presentan un mayor número de clíticos que los de filosofía. Asimismo, los textos de filosofía presentan un mayor número de adverbios frente a los textos del otro grupo.

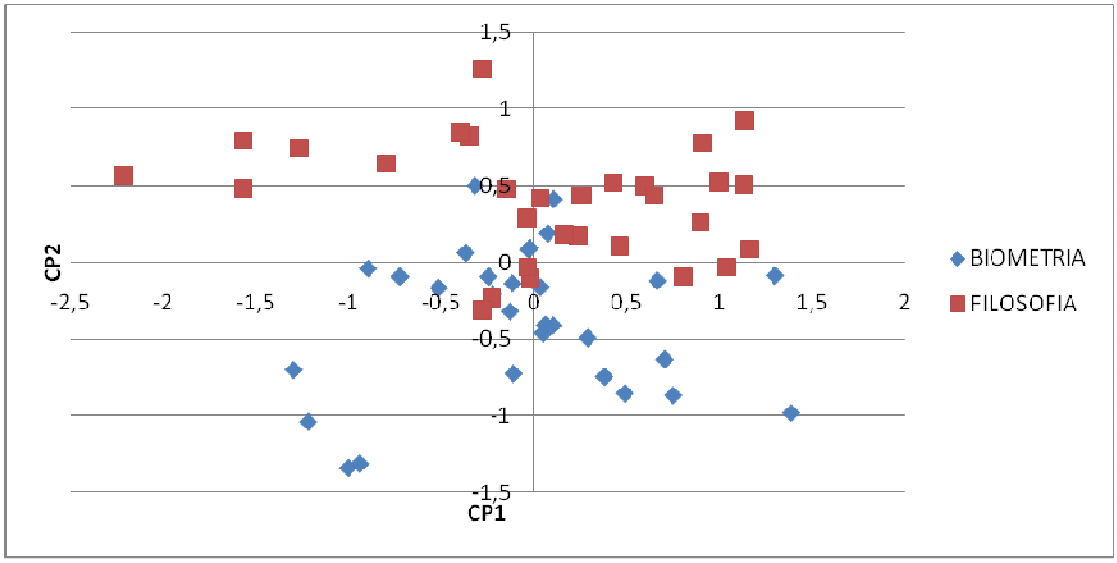

Gráfico 1: Proyección de los textos sobre las dos primeras componentes principales

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 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.

La proyección de los textos sobre un gráfico bidimensional permitió la visualización de las diferencias halladas entre los dos corpus.

Las dos primeras componentes explicaron un 63% de la variación total de los datos, observándose que la segunda componente es la dimensión que separa los textos de ambas disciplinas. El número de clíticos y adverbios en el texto son las variables de mayor importancia en dicha separación. En los textos de biometría hay más clíticos que en los humanísticos por la frecuencia de expresiones impersonales o pasivas con "se" como "se ajusta un modelo lienal" , "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.

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.