Comparación de dos técnicas multivariadas en la categorización de textos: Sistema de clasificación Bagging y Método del vecino más cercano.

Contrastive analysis of two multivariate techniques in text categorization: Bagging Classification System and Nearest Neighbour Method

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

Abstract

This work compares two multivariate techniques which purpose is the classification of units in previously defined categories. In this case, the Bagging Classification System (BCS) and the Nearest Neighbour Method (NNM) to classify texts are evaluated. The criterion of classification is the genre of the text (Scientific/Non Scientific) and the text characterization is based on the frequency distribution of the morphosyntactic categories.

The BCS showed a global error rate of 26%; a 21% for the scientific texts and a 33% for the non scientific. As regards precision rate and coverage rate, the SCIENTIFIC genre showed a 78% and 79% while the NON SCIENTIFIC showed a 68% and 67%, respectively.

The NNM showed a prediction of a global error of 13%; a 9% for the scientific genre and a 20% for the non scientific. As regards precision rate and coverage rate, the scientific genre showed an 87% and 91% while the non scientific showed an 86% and 80%, respectively.

Key words: Bagging – Nearest neighbour – Text classification

Resumen

En este trabajo se comparan dos técnicas multivariadas cuyo objetivo es la clasificación de unidades en categorías definidas previamente. En este caso se evalúan los desempeños del Sistema de

Clasificación Bagging (SCB) y el Método del Vecino más Cercano (MVC) para clasificar textos. El criterio de clasificación es el género al que pertenece el texto (Científico / No Científico) y la caracterización de los textos está basada en la distribución de frecuencias de las categorías morfo-sintácticas.

En el SCB se halló una tasa de error global de 26%, siendo 21% para los textos científicos y 33% para los no científicos. Respecto a la precisión y cobertura fueron de 78% y 79% para el género CIENTÍFICO y de 68% y 67% para los textos NO CIENTÍFICOS, respectivamente.

Para el MVC el error global en la predicción resultó ser del 13%, correspondiendo un 9% para el género Científico y un 20% para el No Científico y respecto a la precisión y cobertura fueron de 87% y 91% para el género CIENTÍFICO y de 86% y 80% para los textos NO CIENTÍFICOS, respectivamente.

Palabras claves: Bagging, vecino más cercano, clasificación de textos.

1. INTRODUCCION

Una de las tareas de la lingüística computacional que ha adquirido mayor importancia debido a su utilidad es la clasificación de documentos. Esto se debe a la gran cantidad de información disponible en la web. La clasificación automática de textos tiene por objetivo categorizar documentos dentro de un número fijo de categorías definidas previamente en función de su contenido. Mediante el aprendizaje automático se logra aprender a clasificar a partir de ejemplos que permitan hacer la asignación a la categoría automáticamente. Para llevarlo a cabo es necesario disponer de un conjunto de documentos cuya categoría de pertenencia se conozca como así también la información que se utilizará para caracterizarlos. Durante el entrenamiento se evalúan las condiciones de pertenencia a cada una de las categorías.

En este trabajo se evalúan los desempeños del Sistema de Clasificación Bagging (SCB) y el Método del Vecino más Cercano (MVC) para clasificar documentos. El criterio de clasificación es el género al que pertenece el texto (Científico / No Científico) y la caracterización de los textos está basada en distribución de frecuencias de las categorías morfo-sintácticas.

2. MATERIAL Y METODOS

2.1. Diseño de la muestra

El conjunto de textos que participan en la investigación, el corpus, corresponde a distintos tipos de acuerdo a los requerimientos de los objetivos planteados. Estos textos fueron agrupados de la siguiente manera definiendo así 4 estratos:

- Noticias de tipo general, en español.
- Resúmenes, en español, de trabajos científicos presentados a congresos o revistas de Biometría/Estadística.
- Resúmenes, en español, de trabajos científicos presentados a congresos o revistas de Lingüística.
- Resúmenes, en español, de trabajos científicos presentados a congresos o revistas de Filosofía.

Los textos periodísticos fueron seleccionados de un corpus mayor utilizado por el equipo de investigación INFOSUR utilizado en mi tesis de doctorado. Este corpus se construyó con noticias

extraídas de las páginas web de periódicos argentinos. Por otro lado, los textos científicos fueron seleccionados a partir de un marco muestral compuesto por textos académicos, resúmenes de trabajos presentados a congresos y revistas, extraídos de internet pertenecientes a las disciplinas: Biometría, Lingüística y Filosofía. La unidad de muestreo fue el texto y la selección de la muestra se llevó a cabo empleando un muestreo aleatorio estratificado. La muestra final contiene 60 textos académicos de cada estrato (de modo de poder utilizar 30 de ellos para estimar los modelos o entrenar los sistemas y los restantes para evaluar la tasa de error de clasificación en cada caso) y 120 textos periodísticos (de modo de utilizar 60 de ellos durante el entrenamiento de cada sistema y los restantes para la etapa de evaluación). La base actual contiene 300 textos y 42.491 palabras.

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 por palabra, esto es, cada unidad o fila es una palabra analizada del texto. Luego 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 1 con la estructura presentada en la tabla 2.

Tabla 1. Variables de la base de datos por documento

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

Tabla 2. Fragmento de la base de datos para análisis estadístico

GÉNERO	TEXTO	adj	adv	cl	cop	det	nom	prep	v	OTRO	TOTAL_PAL
C	1	0,11	0,02	0,02	0,04	0,16	0,26	0,18	0,09	0,11	185
C	2	0,13	0,00	0,05	0,04	0,13	0,25	0,18	0,08	0,15	110
C	3	0,09	0,03	0,06	0,03	0,15	0,26	0,14	0,10	0,14	181
…	…	…	…	…	…	…	…	…	…	…	…
NC	1	0,08	0,01	0,02	0,03	0,16	0,32	0,21	0,09	0,09	186
NC	2	0,10	0,00	0,03	0,04	0,17	0,28	0,18	0,09	0,11	141
NC	3	0,10	0,03	0,01	0,03	0,19	0,27	0,16	0,10	0,11	183
…	…	…	…	…	…	…	…	…	…	…	…

2.4. Sistema de Clasificación Bagging

La técnica BAGGING (Breiman, 1994) tiene por objetivo combinar distintos clasificadores generados a partir de un mismo conjunto de datos y así lograr una mejora en la predicción de la categoría de pertenencia. Busca mejorar el desempeño de los Árboles de clasificación.

Este procedimiento obtiene muchas muestras de entrenamiento obtenidas, por muestreo Bootstrap, a partir de un único conjunto de datos. Con cada conjunto de entrenamiento obtiene un árbol de clasificación y combina las predicciones de cada uno de ellos para obtener la categoría de pertenencia de una nueva observación. En cada caso se estimó previamente el número de árboles a combinar de modo que el porcentaje de error en la clasificación sea aceptable.

Sea el conjunto de datos E={$(\mathbf{x_1}, Y_1)\,(\mathbf{x_2}, Y_2)\,(\mathbf{x_3}, Y_3)\ldots(\mathbf{x}_n, Y_n)$} de tamaño n. A partir de dicho conjunto se generan M muestras mediante el método de Bootstrap, esto es, M muestras aleatorias simples con reposición de tamaño n de E, Ek (k=1,2,…,M) donde cada elemento del conjunto E tiene una probabilidad aproximada de 0.63 de ser seleccionado.

En cada una de las muestras Ek, se obtiene un predictor basado en árboles de clasificación y estos predictores individuales son combinados para obtener una predicción final (predicción Bagging). La predicción por Bagging será la categoría más frecuente hallada en los M predictores individuales.

El algoritmo se resume en los siguientes pasos (Figura 1):

1- Sea E={ $(\mathbf{x_1}, Y_1)\,(\mathbf{x_2}, Y_2)\,(\mathbf{x_3}, Y_3)\ldots(\mathbf{x}_n, Y_n)$ }el conjunto de datos.

2- Se construyen M muestras Bootstrap $E_1, E_2, \ldots E_k$ de tamaño n.

3- Para cada una de ellas se obtiene el predictor $g(\mathbf{x}, E_1)$, $g(\mathbf{x}, E_2), \ldots, g(\mathbf{x}, E_M)$

4- Se calcula el estimador Bagging mediante

$$gBagg(\mathbf{x}) = \arg\max_{y}(\#\{k : g(\mathbf{x_1}, E_k) = y\}) \text{ para } k=1,2,..,M$$

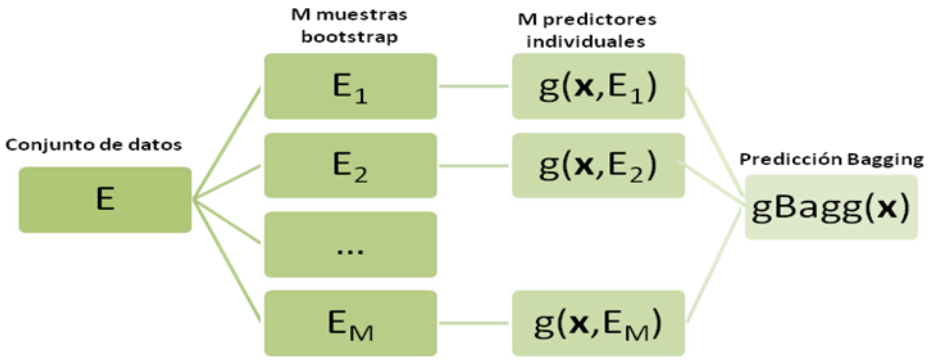

Figura 1: Esquema del algoritmo de clasificación Bagging

2.5. Método del vecino más cercano

La clasificación por el método del vecino más cercano es una de las técnicas no paramétricas de clasificación más utilizadas. La idea en la cual está basado el método es muy simple, para predecir la categoría a la cual pertenece una nueva unidad (clasificar) sólo considera las k unidades del grupo de entrenamiento más cercanas o parecidas a dicha unidad. Este método clasifica a la nueva unidad al grupo al cual pertenece la mayoría de los k vecinos más cercanos del grupo de entrenamiento.

Sea $(\mathbf{x_1}, Y_1)\,(\mathbf{x_2}, Y_2)\,(\mathbf{x_3}, Y_3)\ldots(\mathbf{x}_n, Y_n)$ la muestra de entrenamiento, donde la variable Y es la que se refiere a la variable de clasificación y sus niveles corresponden a las distintas categorías a las cuales pertenecen las unidades, y el vector x contiene las covariables utilizadas para asignar la categoría de la variable Y a la cual pertenece la unidad.

La muestra que será utilizada como validación es similar a la de entrenamiento pero sin considerar la variable Y, la cual es conocida pero será utilizada luego de aplicar el sistema para evaluar su desempeño.

Puesto que requiere reconocer las unidades más cercanas a la unidad a clasificar es necesario definir una medida de distancia entre unidades. Esta medida debe ser calculada en función del conjunto de covariables x cuya información se considera relevante para la clasificación. Para variables cuantitativas, como las utilizadas en esta aplicación, algunas de las medidas de distancia usuales son la distancia Euclídea, la distancia de Mahalanobis y otras variantes.

La distancia euclídea entre el punto P y un punto fijo Q con coordenadas P=(x1, x2, ...,xp) y Q=(y1, y2, ...,yp) está dada por

$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_p - y_p)^2}$$

Una característica de la distancia euclídea es que cada una de las coordenadas contribuye de la misma manera en el cálculo de la distancia. Sin embargo, en muchas situaciones las coordenadas representan mediciones de diferente magnitud y es deseable que el "peso" de cada coordenada tome en consideración la variabilidad de las mediciones. Esto sugiere distintas definiciones de distancia.

Una distancia "estadística" que tenga en cuenta las distintas variabilidades de las variables se puede construir a partir de las coordenadas estandarizadas, $x_j^* = x_j / \sqrt{S_{jj}}$, para el punto P y $y_j^* = y_j / \sqrt{S_{jj}}$, para el punto Q, con j=1,2,...,p

$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{S_{11}} + \frac{(x_2 - y_2)^2}{S_{22}} + \ldots + \frac{(x_p - y_p)^2}{S_{pp}}}$$

el "peso" que se le da a la j-ésima coordenada es kj=1/Sjj, para j=1,2,...,p. Si S11=S22=...=Spp, entonces la distancia euclídea es conveniente.

Otro aspecto importante de este método es determinar el valor de k. Si toma un valor grande se corre el riesgo de hacer la clasificación de acuerdo a la mayoría global del conjunto de entrenamiento y no a los parecidos de esta manera se obtendría una predicción constante para toda unidad a clasificar. Por otro lado, si el valor es chico puede perderse exactitud debido a la presencia de ruido en los datos. En esta aplicación el valor de k fue seleccionado buscando minimizar el error de clasificación.

El método del vecino más cercano para clasificar una unidad P se puede describir enunciando unos simples pasos:

1. Se define la distancia entre unidades (puntos) que se va a utilizar
2. Calcular la distancia de P a cada uno de las unidades del conjunto de entrenamiento.
3. Registrar las k unidades más próximas a P.
4. Calcular la frecuencia (cantidad de puntos o unidades), de los k vecinos más cercanos, que pertenecen a cada una de las categorías.
5. Clasificar a la unidad P en la categoría que presente mayor frecuencia.

3. RESULTADOS

3.1. Sistema de Clasificación Bagging

Se lleva a cabo el procedimiento para obtener un clasificador de textos según el género: Científico y No científico. En este caso el número de árboles a ensamblar óptimo fue 45 árboles. Con este sistema se obtiene una tasa de error global de 26%. (Tabla 3), siendo 21% para los textos científicos y 33% para los no científicos. Respecto a la precisión y cobertura fueron de 78% y 79% para el género CIENTÍFICO y de 68% y 67% para los textos NO CIENTÍFICOS, respectivamente.

Tabla 3: Tasa de error estimada, Precisión y Cobertura

Medidas de evaluación		
	CIENTIFICO	**NO CIENTIFICO**
Tasa de error	21,1%	33,3%
Precisión	78,0%	67,8%
Cobertura	78,9%	66,7%

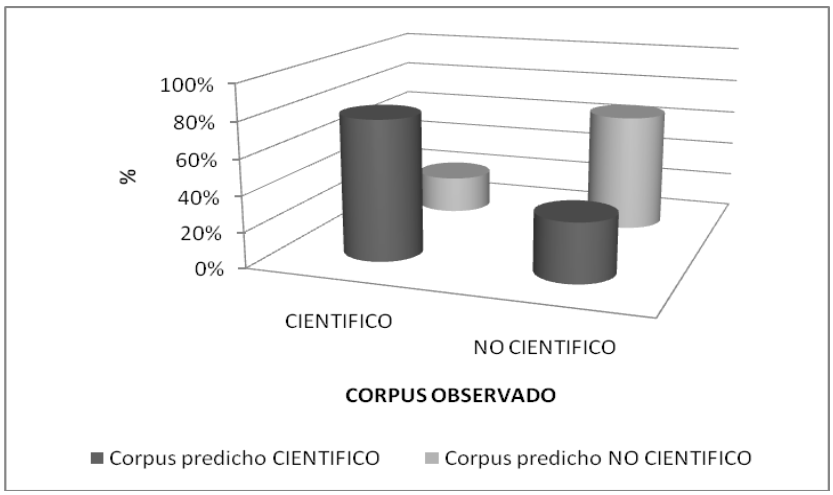

Gráfico 1: Clasificaciín según género mediante Bagging.

3.2. Método del Vecino más Cercano

En este método, el número de textos cercanos a considerar en la predicción del género al que pertenece, resultó ser de 3. El error global en la predicción bajo este sistema resultó ser del 13%, correspondiendo un 9% para el género Científico y un 20% para el No Científico.

El error global en la predicción bajo este sistema resultó ser del 13%, correspondiendo un 9% para el género Científico y un 20% para el No Científico. Respecto a la precisión y cobertura fueron de 87% y 91% para el género CIENTÍFICO y de 86% y 80% para los textos NO CIENTÍFICOS, respectivamente.

Tabla 4: Tasa de error estimada, Precisión y Cobertura

Medidas de evaluación		
	CIENTIFICO	**NO CIENTIFICO**
Tasa de error	8,9%	20,0%
Precisión	87,2%	85,7%
Cobertura	91,1%	80,0%

Se investigó variando el número de vecinos a considerar en la predicción sin hallar mejoras en las predicciones.

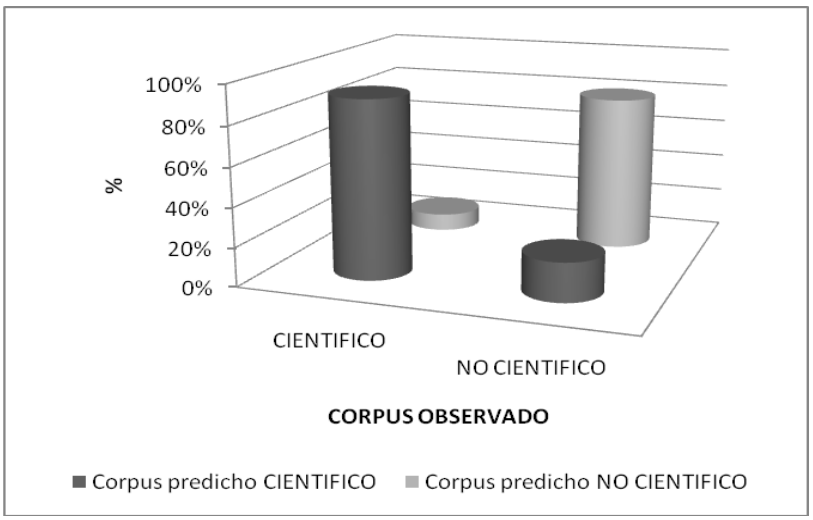

Gráfico 2: Clasificaciín según género mediante el método del Vecino más cercano.

4. CONCLUSIONES

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

El desempeño de las técnicas fue medido con la tasa de mala clasificación, la precisión y la cobertura. Estas medidas fueron calculadas sobre una muestra de textos de evaluación.

En el SCB se halló una tasa de error global de 26%, siendo 21% para los textos científicos y 33% para los no científicos. Respecto a la precisión y cobertura fueron de 78% y 79% para el género CIENTÍFICO y de 68% y 67% para los textos NO CIENTÍFICOS, respectivamente.

Para el MVC el error global en la predicción resultó ser del 13%, correspondiendo un 9% para el género Científico y un 20% para el No Científico y respecto a la precisión y cobertura fueron de 87% y 91% para el género CIENTÍFICO y de 86% y 80% para los textos NO CIENTÍFICOS, respectivamente.

La técnica de Bagging no mostró superioridad frente al método del vecino más cercano. En estos datos no logró discriminar los textos mostrando una tasa de error superior.

El método del vecino más cercano tuvo un desempeño mejor. Sin embargo, la ventaja de éste radica en lo simple de su aplicación y en la estabilidad de su comportamiento.

Referencias

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.

Beltrán, C. 2011. Aplicación del análisis de regresión logística multinomial en la clasificación de textos académicos: Biometría, Filosofía y Lingüística informática. Revista INFOSUR. Grupo INFOSUR. Rosario.

Beltrán, C. 2012 Aplicación de redes neuronales artificiales en la clasificación de textos académicos según disciplina: Biometría, Filosofía y Lingüística informática. 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

Catena, A.; Ramos, M.M; Trujillo, H.M. 2003. ANALISIS MULTIVARIADO. UN MANUAL PARA INVESTIGADORES. Bibiloteca Nueva S.L. España.

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

Flórez López, R.; Fernández Fernández, J.M. 2008. LAS REDES NEURONALES ARTIFICIALES. FUNDAMENTOS TEORICOS Y APLICACIONES PRACTICAS. Netbiblio S.L. España.

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

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).