Una propuesta para el tratamiento de los enclíticos en NooJ

A proposal to analyze the enclitics using NOOJ Tools

Rodolfo Bonino
Grupo INFOSUR
Universidad Nacional de Rosario, Argentina
rodolfobonino@yahoo.com.ar

Abstract

The fact that enclitics form a graphic unit with the verb and, in several cases, may produce graphic alterations (apocopation and chances in the accent) places these sequences between morphology and syntax. The present work tries to solve the empirical problem in the treatment of verb and enclitic sequences using the NooJ tools. In NooJ, the morphological grammars are used to analyze the internal structure of the lexical units and the syntax grammars are used for the relations between the different lexical units. If it is proper to deal with the object of study by means of syntactic grammars, the previous step is to create a productive grammar to analyze that object as a graphical unit formed by two lexical units. Thus, the enclitics are recognized as syntax elements and, consequently, it is possible to elaborate grammatical syntaxes analogous to those employed to treat proclitic and verbal sequences. It is also posed the modifications that must be introduced in the morphological grammar of the verbal system to obtain the graphical forms adopted by verbs when associated with enclitics. Finally, it is presented a brief analysis of the enclitic inserted in the compound forms and the verbal periphrasis.

Key words: NooJ, Spanish language, enclitics, verbs, compound tenses, verbal periphrasis.

Resumen

El hecho de que los enclíticos formen una unidad gráfica con el verbo y, en muchos casos, se produzcan alteraciones gráficas (apócope y cambios de tilde) sitúa a estas secuencias entre la morfología y la sintaxis. En el presente trabajo se intenta resolver el problema empírico que presenta el tratamiento de secuencias de verbos y enclíticos mediante la herramienta NooJ. En NooJ, las gramáticas morfológicas se utilizan para analizar la estructura interna de las unidades léxicas y las gramáticas sintácticas para las relaciones entre distintas unidades léxicas. Si se considera adecuado tratar el objeto de estudio mediante gramáticas sintácticas, el paso previo es crear una gramática productiva que lo analice como una unidad gráfica formada por dos unidades léxicas; así, los enclíticos son reconocidos como elementos de la sintaxis y, consecuentemente, es posible elaborar gramáticas sintácticas análogas a las que se utilizan para tratar secuencias de proclíticos y verbos. También se plantean las modificaciones que se deben introducir en la gramática morfológica del sistema verbal para obtener las formas gráficas que adoptan los verbos cuando se asocian con enclíticos. Finalmente, se presenta brevemente el análisis de los enclíticos insertos en las formas compuestas y las perífrasis verbales.

Palabras claves: NooJ, español, enclíticos, verbos, tiempo compuesto, perífrasis verbales.

Referencias

Aitchison J. 1983. The Statistical Analysis of Compositional Data. Chapman & Hall, London.

Beltrán, C., Bender, C., Bonino, R., Deco, C., Koza, W., Méndez, B., Moro, Stella Maris. 2008 Recursos informáticos para el tratamiento lingüístico de textos. Ediciones Juglaría. Rosario.

Beltrán, C. 2009 Modelización lingüística y análisis estadístico en el análisis automático de textos. Ediciones Juglaría. Rosario.

Beltrán, C. 2010 Estudio y comparación de distintos tipos de textos académicos: Biometría y Filosofía. Revista de Epistemología y Ciencias Humanas. Grupo IANUS. Rosario.

Beltrán, C. 2013 Estudio exploratorio para la comparación de distintos tipos de textos: Textos Científicos y Textos No Científicos. 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.

probabilidad de clasificar un texto como No Científico se incrementa en un 22% (1/0.82=0.22) al aumentar en una unidad la cantidad de nombres en el texto.

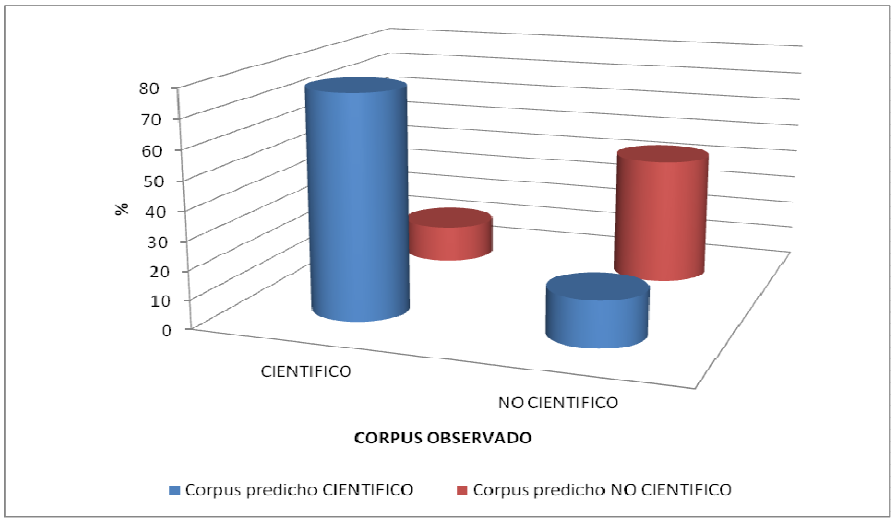

Grafico 3: Clasificación de textos según género mediante Regresión logística

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 (TMC), la precisión (PR) y la cobertura (CO), calculadas sobre una muestra de textos no incluidos en la estimación del modelo y construcción del árbol. El árbol de clasificación presentó una TMC inferior a la del modelo logístico logrando clasificar con mayor precisión los textos científicos. Para el AC la TMC, PR y CO resultaron 4%, 84% y 96% para los textos científicos y 28%, 92% y 72% para los textos no científicos, respectivamente. Para el modelo de RL la TMC, PR y CO resultaron 14%, 83% y 86% para los textos científicos y 26%, 77% y 74% para los textos no científicos, respectivamente.

La diferencia en la tasa de mala clasificación sólo se diferenció en el corpus de textos científicos para el cual con el árbol se obtuvo un 4% de mala clasificación versus un 14% para el modelo de regresión logística.

En ambos tipos de análisis, las diferencias entre los dos tipos de textos están centradas principalmente en el porcentaje de adverbios, adjetivos, nombres y preposiciones presentes. Sin embargo, en el modelo de regresión logística han intervenido otras variables en la discriminación como los determinantes y conjunciones copulativas; mientras que el árbol de clasificación utiliza el porcentaje de verbos, categoría morfológica no utilizada en la regresión.

Una ventaja observada en el árbol de clasificación es la adaptación para recoger el comportamiento no aditivo de las variables predictoras, de manera que las interacciones se incluyen de manera automática. Sin embargo, en esta técnica se pierde información al tratar a las variables predictoras continuas como variables dicotómicas.

$$P(\in Cien/X) = \pi(x) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_j x_j+...+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_j x_j+...+\beta_p x_p}}$$

$$P(\in No\ Cien/X) = \pi(x) = \frac{1}{1+e^{\beta_0+\beta_1 x_1+...+\beta_j x_j+...+\beta_p x_p}}$$

Con este criterio un texto es asignado al corpus cuya probabilidad es máxima.

La bondad del ajuste se evaluó mediante el test de Hosmer-Lemeshow y la tasa de error de clasificación. Con el modelo de regresión logística obtenido durante la selección de variables se obtuvo una tasa de error global, estimada sobre el corpus de prueba, del 20% y la probabilidad asociada en el test de bondad de ajuste es p=0.9696 evidenciando lo adecuado del modelo. La tabla 4 presenta las medidas de precisión, cobertura y tasa de error para cada género.

Tabla 4: Tasa de error estimada, Precisión y Cobertura

Medidas de evaluación		
	CIENTIFICO	**NO CIENTIFICO**
Tasa de error	14%	26%
Precisión	83%	77%
Cobertura	86%	74%

Tabla 5: Razones de odds estimadas

Razón de odds			
Efecto	**Estimación puntual**	**IC 95%**	
adjetivos	1.185	1.061	1.323
adverbios	0.733	0.627	0.857
Conj. Cop.	1.319	1.069	1.628
determinantes	1.129	1.031	1.237
nombres	0.819	0.748	0.897
preposiciones	1.171	1.052	1.302

Los coeficientes del modelo de regresión logística permiten la interpretación de la misma. La razón de odds para el número de adjetivos es 1.19 lo cual indica que la chance de clasificar a un texto como Científico se incrementa en un 19% al aumentar en número de adjetivos en una unidad. Con respecto al número de adverbios la razón de odds es menor a la unidad por lo tanto si se interpreta el recíproco, 1/0.73=1.36, significa que la chance de clasificar un texto en el corpus No Científico aumenta un 36% al incrementarse en una unidad el número de adverbios. Si analizamos el efecto de las conjunciones copulativas, determinantes y preposiciones, al incrementar en una unidad cada una de estas categorías gramaticales, la chance de clasificar un texto como Científico se incrementa en un 32%, 13% y 17% respectivamente. Al igual que el efecto del número de adverbios, la

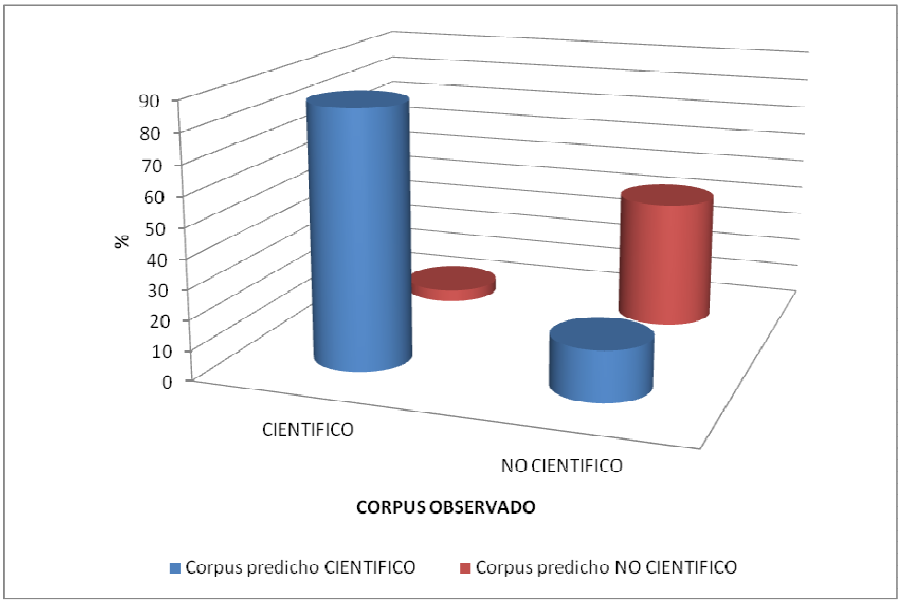

Grafico 2: Clasificación de textos según género mediante Árboles de Clasificación

3.2. Análisis de Regresión Logística

Se realizó un análisis de regresión logística para obtener una regla de clasificación que permita asignar los textos en estas dos poblaciones, definidas por el género al que pertenecen (Científico / No científico), en base a la frecuencia de cada categoría gramatical en el texto.

Para determinar cuáles categorías gramaticales son 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 adjetivos, adverbios, conjunciones copulativas, determinantes, nombres y preposiciones.

Tabla 3: Coeficientes del modelo de regresión logística

Estimación máximo verosímil					
Coeficiente	**gl**	**Estimador**	**Error estándar**	**Est. Chi-cuadrado**	**Prob. asociada**
Intercepto	1	-0.6868	0.5507	1.5554	0.2123
adjetivos	1	0.1694	0.0562	9.0777	0.0026
adverbios	1	-0.3106	0.0800	15.0862	0.0001
Conj. Cop.	1	0.2769	0.1073	6.6566	0.0099
determinantes	1	0.1216	0.0464	6.8795	0.0087
nombres	1	-0.1995	0.0464	18.5044	<.0001
preposiciones	1	0.1575	0.0544	8.3925	0.0038

Este modelo permite, mediante la utilización de los coeficientes estimados, calcular para cada texto la probabilidad de pertenecer a cada uno de los corpus definidos por el género.

encuentran 10 perfiles de textos (que corresponden a los 10 nodos terminales) asociados con una de los dos géneros. Estos son:

- Textos con un porcentaje de adverbios inferior al 2.9% y un porcentaje de nombres menor a 25% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25% y de adjetivos inferior al 5% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición menor al 17% y de verbos menor al 15% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición entre 17% y 19%, y de verbos menor al 15% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición inferior al 19%, y de verbos mayor al 15% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios inferior al 2.9%, un porcentaje de nombres mayor a 25%, de adjetivos superior al 5%, de preposición mayor al 19% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos inferior al 4.7% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos entre 4.7% y 5.7% son clasificados como textos CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos entre 5.7% y 9.5% son clasificados como textos NO CIENTÍFICOS.
- Textos con un porcentaje de adverbios superior al 2.9% y un porcentaje de adjetivos superior al 9.5% son clasificados como textos CIENTÍFICOS.

El árbol final fue evaluado utilizando la muestra de prueba, no fue utilizada en la construcción del mismo, hallando una tasa de mala clasificación del 14%, siendo 4% para los textos científicos y 28% para los no científicos. Respecto a la precisión y cobertura fueron de 84% y 96% para el género CIENTÍFICO y de 92% y 72% para los textos NO CIENTÍFICOS, respectivamente.

Tabla 2: Tasa de error estimada, Precisión y Cobertura

Medidas de evaluación		
	CIENTIFICO	**NO CIENTIFICO**
Tasa de error	4%	28%
Precisión	84%	92%
Cobertura	96%	72%

3. RESULTADOS

En Beltrán 2013 se realizó un análisis exploratorio en el cual se evidencian las características que discriminan los corpus de textos en estudio. En dicho estudio se evidenció que existen diferencias significativas entre los corpus respecto al tamaño de los textos (número de palabras por texto). Esta situación llevó a realizar las sucesivas comparaciones sobre los porcentajes o proporciones de las categorías gramaticales, hallando diferencias significativas ($p<0.05$) para todas las categorías gramáticas excepto la proporción de clíticos y de verbos en los documentos analizados. Asimismo, en un análisis de componentes principales, se dispusieron los textos en el plano de proyección demostrando que los textos procedentes del corpus No Científico presentan un mayor número de adverbios, respecto a las restantes categorías, que los textos Científicos.

3.1. Árboles de Clasificación

Se aplicó la técnica de Árboles de Clasificación para obtener reglas de clasificación que permitan asignar los textos en dos poblaciones, definidas por el género al que pertenecen: CIENTÍFICO y NO CIENTÍFICO. De la misma manera que en el apartado previo, los predictores utilizados corresponden a la distribución de las distintas categorías morfológicas halladas en el análisis automático de los textos (proporción de cada categoría morfológica).

Las variables que mostraron una buena discriminación de los grupos son la proporción de adverbios, nombres, adjetivos, preposición y verbos. El árbol final presenta 10 nodos terminales.

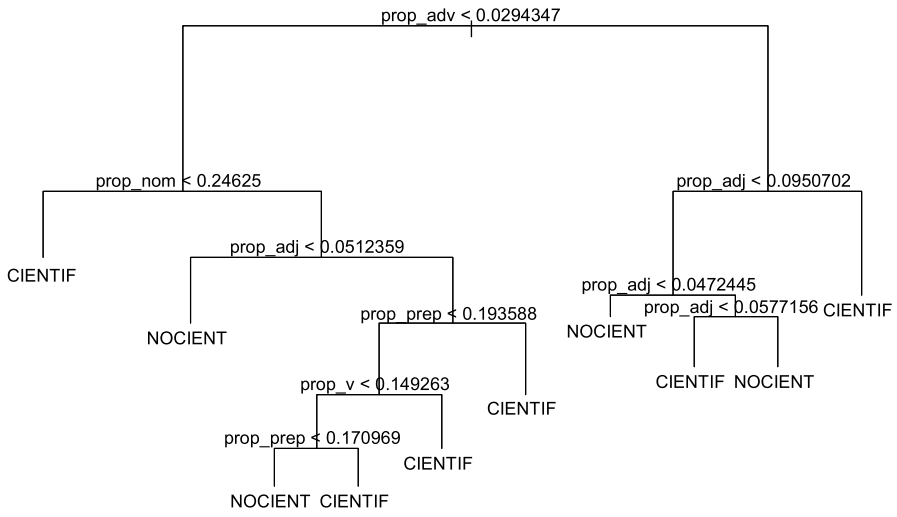

Gráfico 1: Árbol de clasificación

El gráfico 1 muestra el árbol resultante de esta aplicación. La variable regresora más fuertemente asociada con el género es la proporción de adverbios en el texto, categorizada como superiores e inferiores a 0.029 (2.9%). El corpus general queda así dividido en dos grupos, los que presentan más del 2.9% de adverbios y los que no. Luego intervienen en las sucesivas subdivisiones el número de nombres, adjetivos, verbos y preposiciones. Interpretando el árbol resultante, se

$$g(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

β_0 es la constante del modelo o término independiente

p el número de covariables

β_i los coeficientes de las covariables

x_i las covariables que forman parte del modelo.

Si alguna de las variables independientes es una variable discreta con k niveles se debe incluir en el modelo como un conjunto de k-1 "variables de diseño" o "variables dummy". El cociente de las probabilidades correspondientes a los dos niveles de la variable respuesta se denomina odds y su expresión es:

$$\frac{P(y=1|X)}{1-P(y=1|X)} = e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}$$

Si se aplica el logaritmo natural, se obtiene el modelo de regresión logística:

$$\log\left(\frac{P(y=1\mid X)}{1-P(y=1\mid X)}\right) = \log(e^{\beta_0+\beta_1 x_1+...+\beta_p x_p}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

En la expresión anterior el primer término se denomina logit, esto es, el logaritmo de la razón de proporciones de los niveles de la variable respuesta.

Los coeficientes del modelo se estiman por el método de Máxima Verosimilitud y la significación estadística de cada uno de los coeficientes de regresión en el modelo se puede verificar utilizando, entre otros, el test de Wald y el test de razón de verosimilitudes.

Una cuestión importante en este tipo de análisis es determinar si todas las variables consideradas en la función de discriminante contienen información útil y si solamente algunas de ellas son suficientes para diferenciar los grupos. Dado que las variables utilizadas para explicar la respuesta es probable que estén correlacionadas, es posible también que compartan información. Por lo tanto, se puede buscar un subgrupo de variables mediante algún criterio de modo tal que las variables excluidas no contengan ninguna información adicional. Existen varios algoritmos de selección de variables, entre ellos podemos citar:

Método forward: comienza por seleccionar la variable más importante y continúa seleccionando las variables más importantes una por vez, usando un criterio bien definido. Uno de estos criterios involucra el logro de un nivel de significación deseado pre-establecido. El proceso termina cuando ninguna de las variables restantes encuentra el criterio pre-especificado.

Método backward: comienza con el modelo más grande posible. En cada paso descarta la variable menos importante, una por vez, usando un criterio similar a la selección forward. Continúa hasta que ninguna de las variables pueda ser descartada.

Selección paso a paso: combina los dos procedimientos anteriores. En un paso, una variable puede entrar o salir desde la lista de variables importantes, de acuerdo a algún criterio pre-establecido.

En este trabajo se utilizó como evaluación del ajuste del modelo el test de Hosmer-Lemeshow y, dado que el modelo es utilizado para clasificar unidades, se utilizó también la tasa de mala clasificación calculada sobre la muestra independiente excluida en la etapa de estimación.

recursivas logran formar grupos homogéneos respecto a la variable respuesta (en este caso el género a la que pertenece el texto). El árbol determinado puede ser utilizado para clasificar nuevos textos.

Es un método no-paramétrico de segmentación binaria donde el árbol es construido dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. Comienza el algoritmo con un nodo inicial o raíz a partir del cual se divide en dos sub-grupos o sub-nodos, esto es, se elige una variable y se determina el punto de corte de modo que las unidades pertenecientes a cada nuevo grupo definido sean lo más homogéneas posible. Luego se aplica el mismo procedimiento de partición a cada sub-nodo por separado. En cada uno de estos nodos se vuelve a repetir el proceso de seleccionar una variable y un punto de corte para dividir la muestra en dos partes homogéneas.

Las divisiones sucesivas se determinan de modo que la heterogeneidad o impureza de los sub-nodos sea menor que la del nodo de la etapa anterior a partir del cual se forman. El objetivo es particionar la respuesta en grupos homogéneos manteniendo el árbol lo más pequeño posible.

La función de impureza es una medida que permite determinar la calidad de un nodo, esta se denota por i(t). Si bien existen varias medidas de impureza utilizadas como criterios de partición, una de las más utilizadas está relacionada con el concepto de entropía

$$i(t) = \sum_{j=1}^{K} p(j/t).\ln p(j/t)$$

donde j = 1, ..., k es el número de clases de la variable respuesta categórica y p(j|t) la probabilidad de clasificación correcta para la clase j en el nodo t. El objetivo es buscar la partición que maximiza

$$\Delta i(t) = -\sum_{j=1}^{K} p(j/t).\ln p(j/t)\,.$$

De todos los árboles que se pueden definir es necesario elegir el óptimo. El árbol óptimo será aquel árbol de menor complejidad cuya tasa de mala clasificación (TMC) es mínima. Generalmente esta búsqueda se realiza comparando árboles anidados mediante validación cruzada. La validación cruzada consiste, en líneas generales, en sacar de la muestra de aprendizaje o entrenamiento una muestra de prueba, con los datos de la muestra de aprendizaje se calculan los estimadores y el subconjunto excluido es usado para verificar el desempeño de los estimadores obtenidos utilizándolos como "datos nuevos".

2.4.2. Regresión logística

La regresión logística es utilizada en situaciones en las cuales el objetivo es describir la relación entre una variable respuesta categórica, en este caso dicotómica, y un conjunto de variables explicativas que pueden ser tanto categóricas como cuantitativas.

Sea **x** un vector de p variables independientes, esto es, $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. La probabilidad condicional de que la variable *y* tome el valor 1 (presencia de la característica estudiada), dado valores de las covariables **x** es:

$$P(y = 1/X) = \pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

donde

por documento que es analizada estadísticamente. La información registrada en esta base corresponde a las siguientes variables:

- 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

2.4. Metodología Estadística

Uno de los problemas de los que se ocupan las técnicas estadísticas multivariadas es la clasificación de objetos o unidades en grupos o poblaciones. Dos enfoquen agrupan a las técnicas de clasificación. Uno de ellos es cuando se conocen los grupos o categorías y se pretende ubicar los individuos dentro de estas categorías a partir de los valores de ciertas variables. El segundo enfoque, que no es el utilizado en este trabajo, ocurre cuando no se conocen los grupos de antemano y se pretende establecerlos a partir de los datos observados.

Referido al primer enfoque, una de las técnicas más utilizadas es la Regresión Logística. La regresión logística estima la probabilidad de un suceso en función de un conjunto de variables explicativas. Este modelo expresa matemáticamente la probabilidad de pertenencia a uno de los grupos, de manera que es posible calcularlas y asignar cada unidad al grupo cuya probabilidad de pertenencia estimada sea mayor. Otra técnica muy utilizada son los Árboles de Clasificación que crean una serie de reglas basadas en las variables predictoras que permiten asignar una nueva observación a una de las categorías o grupo.

2.4.1. Árboles de Clasificación

Los árboles de clasificación (AC) se emplean para asignar unidades experimentales a las clases de una variable dependiente a partir de sus mediciones en uno o más predictores. En esta aplicación, los AC se emplean para asignar textos al género al que corresponde: CIENTIFICO – NO CIENTIFICO a partir de la información relevada en el análisis morfológico automático de los textos.

Es un algoritmo que genera un árbol de decisión en el cual las ramas representan las decisiones y cada una de ellas genera reglas sucesivas que permiten continuar la clasificación. Estas particiones

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. 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

desempeño de las técnicas fue medido con la tasa de mala clasificación (TMC), la precisión (PR) y la cobertura (CO), calculadas sobre una muestra de textos no incluidos en la estimación del modelo y construcción del árbol. El árbol de clasificación presentó una TMC inferior a la del modelo logístico logrando clasificar con mayor precisión los textos científicos.

Para el AC la TMC, PR y CO resultaron 4%, 84% y 96% para los textos científicos y 28%, 92% y 72% para los textos no científicos, respectivamente.

Para el modelo de RL la TMC, PR y CO resultaron 14%, 83% y 86% para los textos científicos y 26%, 77% y 74% para los textos no científicos, respectivamente.

Palabras claves: Regresión logística multivariada, árboles de clasificación, análisis automático de textos, clasificación de textos.

1. INTRODUCCION

Este trabajo se propone la realización del análisis automático de distintos tipos 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. 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 aplican las técnicas multivariadas de clasificación: Regresión logística y Árboles de clasificación.

El desempeño de las técnicas es evaluado con tres medidas: la tasa de mala clasificación (TMC), la precisión (PR) y la cobertura (CO), calculadas sobre una muestra de textos, muestra de prueba, no incluidos en la estimación del modelo y construcción del árbol.

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

La muestra final para este trabajo quedó conformada de la siguiente manera:

Técnicas estadísticas de clasificación. Una aplicación en la clasificación de textos según el género: Textos Científicos y Textos No Científicos

Statistical Techniques in Classification. An Application to Text Classification According to Gender: Scientific – Non scientific

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

Abstract

The problem of unit classification in groups or known population samples offers great interest to Statistics; as a consequence of this several techniques have been developed to fulfill this purpose. This work is aimed to classify scientific and non scientific texts comparing the analysis of classification (CT) and the logistic regression (LR) trees. The scientific texts are abstracts of papers published on journals and conference proceedings coming from different disciplines. The non scientific texts are news report of general interest published on the web page of Argentine newspapers. The information obtained from the morphological analysis of these texts is employed as explanatory variable of the multivariable technique applied in this work. The performance of the techniques was measured by means of the false classification rate (FCR), the precision rate (PR) and the coverage rate (CO) estimated by a sample text not included in the prediction model neither in the tree construction. The classification tree showed a FCR lower than the logistic model while the scientific text samples showed a major precision.

For the CT, the FCR, the PR and the CO resulted in 4%, 84% and 96% for the scientific texts and 28%, 92% and 72% for the non scientific texts, respectively.

For the LR model, the FCR, the PR and the CO resulted in 14%, 83% and 86% for the scientific texts and 26%, 77% and 74% for the non scientific texts, respectively.

Key words: Multivariable logistic regression – Classification Trees – Automatic text analysis– Text classification.

Resumen

El problema de la clasificación de unidades en grupos o poblaciones conocidas es de gran interés en estadística, por esta razón se han desarrollado varias técnicas para cumplir este propósito. Este trabajo se propone la clasificación de textos científicos y no científicos comparando las técnicas de Árboles de Clasificación (AC) y Regresión logística (RL). 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 como variables explicativas en las técnicas multivariadas aplicadas en este trabajo. El

35. Zhitomirsky-Geffet, Maayan e Ido Dagan. 2009. Bootstrapping distributional feature vector quality. En *Computational Linguistics* (35):435-461.

Anexo - Listado completo de etiquetas morfosintácticas

Nro.	*Tag*	Ejemplos
1	AJ0	adjetivo neutro en número (**bello** en "*lo bello*")
2	AJ1	adjetivo singular (**amable**)
3	AJ2	adjetivo plural (**amables**)
4	AJC	adjetivo comparativo (**peor**)
5	AJS	adjetivo superlativo (**pésimo**)
6	AT0	artículo neutro (**lo**)
7	AT1	artículo singular (**la**)
8	AT2	artículo plural (**los**)
9	AV0	adverbio (**seguidamente**)
10	AVQ	adverbio interrogativo (**cuándo**)
11	CJC	conjunción coordinante (**y**)
12	CJS	conjunción subordinante (excepto *que*) (**cuando**)
13	CJT	conjunción subordinante (**que** en "*dijo que...*")
14	CRD	adjetivo numeral cardinal (**tres**)
15	DPS	determinante posesivo (**su**, **mi**)
16	DT1	determinante definido singular (**aquel** en "*aquel hombre*")
17	DT2	determinante definido plural ("***aquellos*** *hombres*", "***todos*** *los hombres*")
18	EX0	existencial (**hay**)
19	ITJ	interjección (**ah**, **ehmm**)
20	NN0	sustantivo neutro en número (**virus**)
22	NN1	sustantivo singular (**lápiz**)
22	NN2	sustantivo plural (**lápices**)
23	NNP	sustantivo propio (**Rafael**)
24	ORD	adjetivo numeral ordinal (**sexto**)
25	PND	pronombre demostrativo (**éste**, **esto**)
26	PNI	pronombre indefinido (**ninguno**, **todo**)
27	PNP	pronombre personal (**tú**)
28	PNQ	pronombre interrogativo (**quién**)
29	POS	pronombre posesivo (**mío**)
30	PPE	pronombre personal enclítico (*dar*-**lo**, se cuasi-reflejo ("*morir****se***", "*él* ***se*** *cayó*")
31	PRP	preposición (excepto *de*) (**sin**)
32	REL	pronombre relativo (**quien** en "*el presidente, quien avisó...*")
33	SEP	se pasivo ("***se*** *venden casas*") e impersonal ("***se*** *reprimió a los manifestantes*")
34	VBG	gerundio de verbo cópula (**siendo**)
35	VBI	infinitivo de verbo cópula (**ser**)
36	VBN	participio de verbo cópula (**sido**)
37	VBZ	verbo cópula conjugado (**es**)
38	VM0	infinitivo de verbo modal (**soler**)
39	VMZ	verbo modal conjugado (**debía**)
40	VMG	gerundio de verbo modal (**pudiendo**)
41	VMN	participio de verbo modal (**podido**)
42	VVG	gerundio de verbo léxico (**obrando**)
43	VVI	infinitivo de verbo léxico (**vivir**)
44	VVN	participio de verbo léxico (**cifrado**)
45	VVZ	verbo léxico conjugado (**vive**)
46	XX0	adverbio de negación (**no**)
47	$$$	fin de oración

16. Elghamry, Khaled. 2004. *A generalized cue-based approach to the automatic acquisition of subcategorization frames*. Tesis de doctorado. Indiana University.

17. Graça, João, Kuzman Ganchev, Luísa Coheur, Fernando Pereira y Ben Taskar. 2011. Controlling Complexity in Part-of-Speech Induction. En *Journal of Artificial Intelligence Research* (41):527-551.

18. Johnson, Kent. 2004. Gold's theorem and cognitive scoences. En *Philosophy of Science* (71):571-592.

19. Jusczyk, Peter, Derek Houston y Mary Newsome. 1999. The beginnings of word segmentation in English-learning infants. En *Cognitive Psychology* (39):159-207.

20. Klein, Dan y Christopher Manning. 2004. Corpus based induction of syntactic structure: models of dependency and constituency. En *Proceedings of ACL 2004*, pp.478-485. Barcelona.

21. Lappin, Shalom y Stuart Shieber. 2007. Machine learning theory and practice as a source of insight into universal grammar. En *Linguistics* (43):393-427.

22. Levy, Yonata. 1985. It's frogs all the way down. En *Cognition* (15):75-93.

23. Manning, Christopher y Hinrich Schütze. 1999. *Foundations of statistical Natural Language Processing*. Cambridge, Massachusetts. MIT Press.

24. Martin, Sven, Jörg Liermann y Hermann Ney. 1998. Algorithms for bigram and trigram word clustering. En *Speech Communication* (24):19-37.

25. Mehler, Jacques, Anne Christophe y Franck Ramus. 1998. What we know about the initial state of language. En *Proceedings of the 1st mind-brain articulation project symposium*, pp.51-75. Tokio.

26. Mintz, Toben. 2003. Frequent frames as a cue for grammatical categories in child directed speech. En *Cognition* 90(1):91-117.

27. Nath, Joydeep, Monojit Choudhury, Animesh Mukherjee, Chris Biemann y Niloy Ganguly. 2008. Unsupervised Parts-of-Speech induction for Bengali. En *Proceedings of LREC'08, European Language Resources Association (ELRA)*, pp.1220-1227. Marrakesh.

28. Pinker, Steven. 1979. Formal models of language learning. En *Cognition* (7):217-282.

29. Popova, Maria. 1973. Grammatical elements of language in the speech of pre-school children. En Ferguson, Charles y Dan Slobin (eds.). *Studies of child language developments*. Nueva York. Holt, Rinehart & Winston.

30. Pullum, Geoffrey. 1996. Learnability, hyperlearning and the argument from the poverty of the stimulus. En *Parasession on learnability, 22nd annual meeting of the Berkeley Linguistics Society*, pp.498-513. Berkeley, California.

31. Redington, Martin, Nick Charter y Steven Finch. 1998. Distributional information: a powerful cue for acquiring syntactic categories. En *Cognitive Science* 22(4):425-469.

32. Schütze, Hinrich. 1993. Part-of-speech induction from scratch. En *Proceedings of the 31st annual conference of the Association for Computational Linguistics,* pp.251-258. Columbus.

33. Shi, Rushen, Janet Werker y James Morgan. 1999. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. En *Cognition* 72(2):11-21.

34. Wang, Hao. 2012. *Acquisition of functional categories*. Tesis de doctorado. University of Southern California.

del lenguaje; en especial, si consideramos que este tipo de enfoques para el español –un idioma particularmente desafiante por el orden libre de sus constituyentes sintácticos– ha venido escaseando durante la última década en el panorama global del estado del arte dentro del paradigma estadístico de la lingüística computacional. En última instancia, la evidencia psicolingüística debería ser refrendada por la neurología, las ciencias cognitivas o incluso la biolingüística, pero la plausibilidad de dicha evidencia mediante una modelización efectiva es claramente un asunto para la agenda actual de la lingüística computacional.

Referencias bibliográficas

1. Berg-Kirkpatrick, Taylor, Alexandre Côté, John Denero y Dan Klein. 2010. Painless unsupervised learning with features. En *Proceedings of NAACL 2010*, pp.582-590. Los Angeles.

2. Böhm, Christian, Christos Faloutsos, JiaYu Pan y Claudia Plant. 2006. Robust information-theoretic clustering. En *Proceedings of the 12th ACM SIGKDD International Conference knowledge discovery and data mining*, pp.65-75. Philadelphia.

3. Brown, Peter, Vincent Della Pietra, Peter Desouza, Jennifer Lai y Robert Mercer. 1992. Class-based n-gram models of natural language. En *Computational Linguistics* 18(4):467-479.

4. Chomsky, Noam. 1957. *Estructuras sintácticas*. México. Siglo XXI.

5. ----------.1959. A review of B.F. Skinner's verbal behavior. En *Language* (35):26-58.

6. ----------. 1975. *Reflexiones sobre el lenguaje*. Buenos Aires. Sudamericana.

7. Christophe, Anne, Séverine Milotte, Savita Bernal y Jeffrey Lidz. 2008. Bootstrapping lexical and syntactic acquisition. En *Language and Speech* (51):61-75.

8. Christodoulopoulos, Christos, Sharon Goldwater y Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? En *Proceedings of the 2010 conference on empirical methods in Natural Language Processing*: 575–584. Cambridge, Massachusetts.

9. Clark, Alexander. 2000. *Inducing syntactic categories by context distribution clustering*. En Proceeding of the CoNLL-2000 and LLL-2000, pp.91-94. Lisboa

10. ----------. 2002. *Unsupervised language acquisition: theory and practice*. Tesis de doctorado. University of Sussex.

11. ----------. 2003. Combining distributional and morphological information for part of speech induction. En *Proceedings of EACL 2003*, pp.59-66. Morristown.

12. Clark, Alexander y Shalom Lappin. 2011. Computational learning theory and language acquisition. En Ruth Kempson, Tim Fernando, y Nicholas Asher (eds.). *Handbook of the philosophy of science*. Volumen 14: Philosophy of Linguistics, pp.1-34. Oxford. Elsevier.

13. Clark, Alexander y Shalom Lappin. 2013. Complexity in language acquisition. En *Topics in Cognitive Science* (5):89-110.

14. Deerwester, Scott, Susan Dumais, George Furnas, Thomas Landauer y Richard Harshman. 1990. Indexing by Latent Semantic Analysis. En *Journal of American Society of Information Sciences* 1(6):391-407.

15. Dromi, Esther. 1987. *Early lexical development*. Nueva York. Cambridge University Press.

> to learn is bounded by the same computational limitations that restrict human abilities in other cognitive domains. The interaction of this condition with the complexity of inducing certain types of representations from available data constitutes a fruitful object of study." [Clark y Lappin 2013:90-91]

El progreso de las técnicas estadísticas y el avance de las investigaciones sobre corpora abarcativos revelan que incluso los más simples mecanismos estadísticos pueden contribuir al esclarecimiento del proceso de adquisición del lenguaje. En particular, el conjunto de marcas e indicios provistos por la información distribucional constituye una herramienta válida para la inducción de juicios acerca de la pertenencia de palabras a categorías morfosintácticas. Hemos demostrado empíricamente la estrecha correlación entre palabras cue vs. palabras target, distinción operativamente homologable a las nociones lingüísticas de palabras funcionales vs. palabras de contenido, y hemos señalado el importante papel que podrían desempeñar dichas palabras funcionales en la adquisición del lenguaje, aunando las respectivas agendas de investigación de la lingüística computacional y de la psicolingüística. Justamente, una deuda pendiente en el campo de la psicolingüística es la necesidad de compatibilizar evidencia contradictoria acerca del momento ontogenético de la adquisición de las palabras funcionales en producción y en comprensión, lo cual contribuirá a la mayor adecuación explicativa de los enfoques computacionales, en función de los diversos pre-requisitos de modelización (el pre-requisito son las cues, no la categorización de las cue).

En este sentido, y sin menoscabo de otros mecanismos de aprendizaje que podrían actuar simultáneamente, se puede concluir que la información distribucional se perfila como un enfoque enriquecedor. El paradigma estadístico se propone como un promisorio marco epistemológico de investigación que requerirá una amplia gama de herramientas y experimentos para explorar cabalmente todo su potencial. Valga, pues, la aclaración de que el experimento delineado en este trabajo representa una mera prueba de concepto que debe ser exhaustivamente mejorada a futuro.

Finalmente, resulta imperioso situar este tipo de investigaciones en el marco más general de un proyecto de inducción integral de sintaxis (Clark 2002; Klein y Manning 2004). El aprendizaje no supervisado de sintaxis o, en otras palabras, el problema de la inducción de una gramática a partir de un corpus sin anotaciones, todavía presenta interesantes desafíos desde el punto de vista de la lingüística teórica y de sus aplicaciones prácticas.

Por otro lado, los investigadores del campo reconocen que es necesaria una mayor evidencia translingüística que apoye la plausibilidad psicolingüística de un aprendizaje general no supervisado de una gramática formal a partir de técnicas estadísticas. En la actualidad no existen trabajos que se hayan propuesto probar tales enfoques para la inducción integral de sintaxis en lenguas flexivas y con orden libre de constituyentes como el español. Así pues, en última instancia el objetivo final de nuestro trabajo a futuro será aportar dicha evidencia translingüística, estudiando la factibilidad de inducir fenómenos sintácticos del español mediante técnicas estadísticas a partir de corpus no estructurado y modelos formales de aprendizaje no supervisado.

Aunque no demostramos necesariamente que el mecanismo por el cual se adquiere una gramática de un lenguaje natural involucre técnicas de clustering, sí demostramos la invalidez del APS en cuanto a que los PLD son suficientemente ricos para inducir una gramática formal (al menos, las categorías POS-tags) únicamente a partir de la información distribucional. Asimismo, dirigimos nuestra atención al debate epistemológico en torno del APS, tratando de arrojar cierta luz sobre confusiones generalizadas en cuanto a los mecanismos lógicos inductivos que podrían actuar como el sustrato cognitivo de los mecanismos generales de aprendizaje que modelizamos en nuestra investigación.

Consideramos entonces que el mérito de la presente investigación es abarcar modelos de inducción de fenómenos sintácticos que puedan aportar renovada evidencia al debate acerca de la adquisición

4.2 Plausibilidad psicolingüística de la modelización

Recapitulando todo lo expuesto hasta ahora, podemos consignar que nuestro experimento reporta exitosamente la viabilidad de inducir categorías morfosintácticas a partir de la información distribucional de los PLD mediante un mecanismo general de aprendizaje, bajo las siguientes dos premisas:

1) Habilidad temprana para reconocer palabras y segmentar oraciones y frases fonológicas (Mehler *et al*. 1998; Jusczyk *et al.* 1999). Evidencia de disponibilidad a partir de los 10 meses.
2) Identificación de las cues (mayormente palabras funcionales) sin necesidad de una tipología diferenciada (no importa si son preposiciones, pronombres o incluso palabras de contenido). Aunque Wang (2012) sostiene que las palabras funcionales pueden estar representadas en forma temprana en el léxico de un modo abstracto, identificadas a partir de indicios prosódicos pero sin acceso a su significado o tipología, en nuestro experimento basta con su reconocimiento como marcas muy frecuentes en los PLD y sus propiedades articulatorias (*pivot*) respecto de las palabras target. (Elghamry 2004). Evidencia de disponibilidad a partir de los 14 meses.

Estas condiciones están plausiblemente dadas incluso bastante antes de la explosión léxica (*vovabulary spurt*) (Dromi 1987) que se da alrededor de los dos años y ciertamente para los 15 meses en donde se verifican los primeros juicios de categorización (Shi *et al.* 1999), por lo que nuestro algoritmo resulta compatible con la evidencia empírica psicolingüística. Lo que demuestra nuestro algoritmo, entonces, es la suficiencia de los PLD mismos para aportar la información necesaria en el proceso de categorización de palabras, sin necesidad de postular conocimiento innato específico de dominio.

En resumen, tomando el trabajo de Redington *et al.* (1998) como punto de partida, nos propusimos encarar un experimento que incorpore sustanciales mejoras en el diseño del algoritmo. A su vez, también éramos concientes de los casi inexistentes intentos previos de llevar a cabo procedimientos sistemáticos de clustering sobre corpora en español. El objetivo del experimento fue demostrar que la información distribucional es una poderosa herramienta suficiente para la inducción de juicios de pertenencia de ítems lexicales a categorías sintácticas. Como se remarcó a lo largo de todo este artículo, el diseño general del experimento respondió a una necesidad de compatibilizar la modelización algorítmica con la plausibilidad psicolingüística del proceso ontogenético de la categorización temprana de palabras.

5. Trabajo a futuro para el experimento de categorización

Los experimentos aludidos en este artículo son una versión resumida de nuestra tesis de doctorado y nos revelan una importante veta de indagación científica que obliga a replantearse cuestiones tan sensibles para la lingüística como la naturaleza del lenguaje y los mecanismos de adquisición del mismo, a la luz de las promisorias técnicas de aprendizaje de máquina y de los procesos de inducción de gramáticas.

> "This problem does not entail that formal learning theory has nothing to offer the study of language acquisition. On the contrary, it is highly relevant. However, we argue that the crucial problems are not information theoretic, as suggested in the Gold results. Instead, they are complexity theoretic. By modeling the computational complexity of the learning process, we can, under standard assumptions, derive interesting result concerning the types of representations (or grammars) that are efficiently learnable. It is uncontroversial that the human capacity

Esta convergencia en las distribuciones de los hiperclusters otorgaría una mayor robustez a nuestro enfoque, ya que no sería necesario postular un parámetro inicial de K clusters, para incializar el modelo, en virtud de la iteración convergente a partir de los ciclos medios. Este punto de consolidación de los ciclos de agrupamiento dependería exclusivamente de la cantidad de cues identifacadas en el corpus. Esto reforzaría la plausibilidad algorítmica del modelo, en tanto no demandaría de un mecanismo de evaluación basado en mínimos o máximos locales sino que la mera iteración convergería a distribuciones consolidadas.

	CICLO 87											
Hipercluster	n		TP	FP	TN	FN	**Precision**	**Recall**	**Fscore**			
AJ1	106		41	37	xxxxxx	65	0,525641026	0,386792453	**0,44565217**	AJ1	0,05185415	
AJ2	38		18	34	xxxxxx	20	0,346153846	0,473684211	**0,4**	AJ2	0,016684962	
AV0	55		32	70	xxxxxx	23	0,31372549	0,581818182	**0,40764331**	AV0	0,024610738	
CRD	14		10	1	xxxxxx	4	0,909090909	0,714285714	**0,8**	CRD	0,012294182	
DPS	7				xxxxxx	7	0	0	**0**	DPS	0	
DT1	7		3	2	xxxxxx	4	0,6	0,428571429	**0,5**	DT1	0,003841932	
DT2	7		4	9	xxxxxx	3	0,307692308	0,571428571	**0,4**	DT2	0,003073546	
NN1	342		304	49	xxxxxx	38	0,861189802	0,888888889	**0,87482014**	NN1	0,328417661	
NN2	92		64	9	xxxxxx	28	0,876712329	0,695652174	**0,77575758**	NN2	0,078342148	
NNP	43		19	33	xxxxxx	24	0,365384615	0,441860465	**0,4**	NNP	0,018880351	
PND	5				xxxxxx	5	0	0	**0**	PND	0	
PRP	8		4	1	xxxxxx	4	0,8	0,5	**0,61538462**	PRP	0,005404036	
VMZ	14		3	2	xxxxxx	11	0,6	0,214285714	**0,31578947**	VMZ	0,004852967	
VVI	42		32	32	xxxxxx	10	0,5	0,761904762	**0,60377358**	VVI	0,027835884	
VVN	14		9	3	xxxxxx	5	0,75	0,642857143	**0,69230769**	VVN	0,010639196	
VVZ	117		103	38	xxxxxx	14	0,730496454	0,88034188	**0,79844961**	VVZ	0,10254512	
INDECIDIBLES	16 clusters con 29 miembros								**0,50184864**	PROMEDIO	**0,68927687**	PONDERADO

Tabla 4: Detalle de evaluación de ciclo 87

4. Discusión de los resultados y conclusiones

4.1 Consideraciones cuantitativas y cualitativas

1) Todas las categorías sintácticas mayores fueron inducidas con un alto grado de pureza. Se observan refinamientos granulares en rasgos de género y número (para sustantivos) y en otras caracterizaciones morfosintácticas (verbos modales VMZ vs. verbos léxicos VVZ).
2) Al igual que en Redington *et al*. (1998), las categorías sintácticas mayores, coincidentes con palabras de contenido (verbos y sustantivos), reportan medidas F altísimas, del orden del 80% y hasta 90%.
3) En el otro extremo, uno de los hiperclusters con menor medida F (40,7%) son los adverbios (AV0). Este grupo quedó confinado a un cluster único y masivo de 95 miembros muy heterogéneos, con objetos claramente marginales (caracteres únicos como '*d*', '*p*', '*v*', etc.). Como reporta Nath *et al*. (2008), es normal que en el clustering partitivo quede en cada ciclo uno o dos clusters masivos que actúan como receptáculo indiferenciado de objetos del espacio vectorial. Posiblemente éste sea el caso.
4) Si bien los adjetivos presentan medidas F bajas, en muchos casos el refinamiento por cluster es sumamente interesante. En uno de los cluster aparecen adjetivos que en general son usados con una proposición ("*es preciso que*...", "*es necesario que*...", etc.).
5) En todos los casos, es notable la consolidación de los agrupamientos a partir de los ciclos medios (ciclo 52 en adelante).

por la ubicación en el espacio vectorial de los centroides de los clusters que lo conforman, lo cual, a su vez, refleja particularidades morfosintácticas propias del dominio lingüístico al que pertenecen los datos.

3.3 Evaluación iterativa de todos los ciclos de clustering con la métrica many-to-1

Ahora que explicamos en detalle en qué consistió nuestro experimento de clustering para inducción de categorías sintácticas en español, su plausibilidad de modelización, sus lineamientos de diseño y sus métricas de evaluación, llegó el momento de analizar la salida completa de los 106 ciclos. Recordemos que el experimento corre iterativamente en ciclos que van desde K=2 clusters hasta K = 106 clusters. Si bien el corte inicial era de 1000 palabras target, 89 de esas palabras correspondía a categorías morfosintácticas marginales: categorías funcionales de poquísimos miembros y de prevalencia intermitente (en muy asialdas ocasiones) en los clusters (REL, AJC, CJC, CJS, etc.). Las restantes 911 palabras target, entonces, se distribuyeron entre 16 categorías de inducción casi permanente a lo largo de todo el experimento, con elevados valores de pureza consolidados a partir de los ciclos medios.

TOTALES	*n*	**Baseline = n/1000**	Probabilidad de acertar el POS-tag por azar
AJ1	106	0,106	
AJ2	38	0,038	
AV0	55	0,055	
CRD	14	0,014	
DPS	7	0,007	Si no se pondera el promedio, la probabilidad de acertar el POS-tag es 1/16, lo cual sigue siendo muy bajo (0,0625 = 6,25%)
DT1	7	0,007	
DT2	7	0,007	
NN1	342	0,342	
NN2	92	0,092	
NNP	43	0,043	
PND	5	0,005	
PRP	8	0,008	
VMZ	14	0,014	
VVI	42	0,042	
VVN	14	0,014	
VVZ	117	0,117	
	Total = 911	**0,0569 = 5,7%**	**Baseline ponderado**

Tabla 3: Palabras target a ser clusterizadas según POS-tag de corpus de referencia y baseline de cada POS-tag

En cada ciclo calculamos Precisión, Cobertura y medida F para cada uno de los 16 POS-tags, prevalezcan o no como el *cluster_tag*, en cada uno de los hiperclusters inducidos. Sobre estas 16 medidas F calculamos el promedio común y el promedio ponderado (según el peso de cada POS-tag en la distribución de 911 palabras target).

Es de destacar que a partir de los ciclos medios (ciclo 52 en adelante), las medidas F de la mitad de los POS-tag se presentan consolidadas en valores relativamente estables, especialmente para las categorías mayores de sustantivos y verbos (NN1, NN2, VVZ, VMZ, VVI, VVN), lo cual significa que a partir de cierto momento de la "historización" de la inducción, las clases están mayormente consolidadas en cuanto a la pertenencia de sus miembros (con mínimas fluctuaciones).

9) Para la evaluación de nuestro experimento exploraremos diversas alternativas, pero podemos adelantar que nos basaremos principalmente en la métrica *many-to-1* (Christodoulopoulos *et al*. 2010). También seguiremos a Redington *et al*. (1998) en una evaluación discriminada para cada tipo de categoría inducida y postularemos nuestra propia justificación algebraica del agrupamiento de clusters (*cluster merging*) (Böhm *et al*. 2006) en *hiperclusters* a partir del mapeo *many-to-1*.

Básicamente el algoritmo propuesto se muestra en el siguiente esquema:

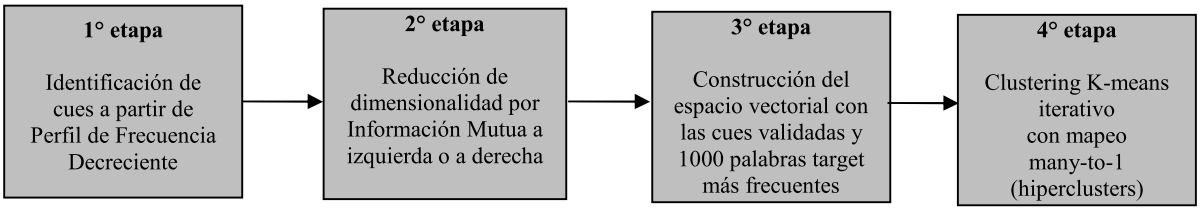

Figura 1: Esquema del algoritmo de categorización de palabras propuesto

3.2 La medida justa: mapeo many-to-1 e hiperclusters

Ante la posibilidad de que algunas categorías del gold standard aparezcan repartidas en varios clusters en función de la granularidad morfosintáctica del tag, la mayor parte de los trabajos de clustering recurren a un mapeo de varios clusters en una única categoría, criterio denominado mapeo *many-to-1*:

> "Many-to-one mapping accuracy (also known as *cluster purity*) maps each cluster to the gold standard tag that is most common for the words in that cluster (henceforth, the *preferred tag*), and then computes the proportion of words tagged correctly. More than one cluster may be mapped to the same gold standard tag. This is the most commonly used metric across the literature as it is intuitive and creates a meaningful POS sequence out of the cluster identifiers. However, it tends to yield higher scores as $|C|$ [number of clusters] increases, making comparisons difficult when $|C|$ can vary." [Christodoulopoulos et al. 2010:577]

En nuestro experimento adoptamos esta decisión de diseño. Más allá de la justificación metodológica, existe una inituición gramatical en adoptar este criterio de evaluación general de la distribución de un ciclo de clustering. Es de esperar que la ubicación de los clusters en el espacio vectorial refleje en alguna medida el criterio de agrupamiento de clusters en función de la similitud de los miembros preeminentes en cada uno de ellos. Así, pues a dos o más clusters del mismo tipo (indicado por el valor del *cluster_tag*) corresponde un mismo *hipercluster*.

Si un centroide representa prototípicamente la ubicación espacial de un cluster, al menos en cuanto a la concentración mayoritaria de sus miembros, entonces al computar la distancia euclideana de los centroides entre sí podemos darnos una idea de qué clusters están más cercanos o más alejados entres sí. Nuestra intucición metodológica de los hiperclusters podría verse justificada empíricamente si, por ejemplo, los clusters *sustantivos singulares* NN1 que conforman el hipercluster NN1 aparecen de algún modo más cercanos entre sí, en comparación con, por ejemplo los clusters que conforman el hipercluster *verbos en infinitivo* VVI.

El concepto de hipercluster, tal como denominamos en este trabajo al agrupamiento de clusters, resulta muy significativo. Desde un punto de vista metodológico permite una evaluación que resuelve el problema del mapeo de un número creciente de clusters inducidos en las categorías del gold standard. Desde un punto de vista algebraico el hipercluster se ve justificado en gran medida

especialmente eficaz en agrupar eventos con una cierta ocurrencia frecuente en el espacio vectorial (Martin *et al*. 1998). A su vez, esta decisión de diseño se condice con la plausibilidad de la evidencia empírica psicolingüística y con la robustez de los modelos matemáticos postulados en dichas técnicas de clustering, reduciendo los costos implementativos:

> "Although the child might not have access to 1,000 vocabulary items, if the child applies distributional analysis over its small productive vocabulary, this will work successfully, because this vocabulary consists almost entirely of content words. Moreover, prior to the vocabulary spurt, the child's syntax, and thus, presumably, knowledge of syntactic categories is extremely limited, and hence even modest amounts of distributional information may be sufficient to account for the child's knowledge. By the third year, the child's productive vocabulary will be approaching 1,000 items (*e.g*., Bates *et al*. 1994, found that the median productive vocabulary for 28 month olds was just under 600 words) and hence could in principle exploit the full power of the method.
> **It is also possible that, even when children's productive vocabularies are small, they may have a more extensive knowledge of the word forms in the language. It is possible that the child may be able to segment the speech signal into a large number of identifiable units**, before understanding the meaning of the units (Jusczyk 1997)." [Redington *et al*. 1998:454] (*las negritas y el subrayado son nuestros*)

> "In practical systems, it is usual to not actually calculate n-grams for all words. Rather, the *n*-grams are calculated as usual only for the most common *k* words […] Because of the Zipfian distribution of words, cutting out low frequency items will greatly reduce the parameter space (and the memory requirements of the system being built), while not appreciably affecting the model quality (*hapax legomena* often constitute half of the types, but only a fraction of the tokens)." [Manning y Schütze 1999:199]

7) El inglés es un idioma con orden fijo de constituyentes sintácticos, los cuales mayormente siguen el orden canónico SVO. Este mecanismo actúa para desambiguar morfosintácticamente formas léxicas idénticas, a falta de marcación morfológica enriquecida. Gran parte del vocabulario inglés puede funcionar indistintamente como verbo o sustantivo. Esto justificaba el tratamiento de la ambigüedad del tipo de palabra morfosintáctica que se observa en Schütze (1993) y en Clark (2002) como un problema de *soft clustering* (posibilidad de asignar un miembro a más de una clase) (Manning y Schütze 1999). Sin embargo, éste no es el caso del español, un idioma morfológicamente rico. Si bien existen en español numerosas formas POS-ambiguas, incluso entre las palabras más frecuentes de cualquier corpus (por ejemplo '*como*', '*para*', '*era*', etc.), consideramos que esta problemática no está tan extendida como en inglés (Graça *et al*. 2011). Por eso, al igual que Redington *et al*. (1998), implementaremos un mecanismo de desambigüación morfosintáctica para tales casos, basado en un corpus de referencia. Es decir, nuestro algoritmo trabajará con un *hard clustering* que asignará cada miembro de las palabras *target* a una única clase o cluster.

8) El corpus con el que se trabajará contará con una extensión compatible con los experimentos de Redington *et al*. (1998) del orden de 2 millones de tokens, respetando criterios de balance y plausibilidad de modelización de los PLD (Chomsky 1959; Pullum 1996). Si bien Clark (2002) sostiene que un corpus que modelice los PLD debe ir desde 10 millones de tokens a 100 millones de tokens para los cuatro años de estímulos linigüísticos que abarcan el período de surgimiento de una gramática de un lenguaje natural, preferimos reducir la complejidad combinatoria de nuestro experimento y demostrar que dichos corpus reducidos ya ofrecen las condiciones suficientes para la categorización de palabras mediante la información distribucional. Si nuestro objetivo se verifica, la hipótesis será validada *a fortiori* para un corpus más masivo.