



Revista de Lingüística Informática,  
Modelización e Ingeniería Lingüística

**Nro 6 - Diciembre 2012**

**ISSN 1851-1996**

Publicación Anual del

**Grupo INFOSUR**

**Universidad Nacional de Rosario  
Argentina**

**Editorial Responsable**

Grupo INFOSUR, Universidad Nacional de Rosario  
Pueyrredón 1175, 4º F - 2000 Rosario, Santa Fe, Argentina  
Tel. + 54 341 4211284  
mail: [zsolana@arnet.com.ar](mailto:zsolana@arnet.com.ar)  
Web: [www.infosurrevista.com.ar](http://www.infosurrevista.com.ar)

**Directora Editorial**

Dra. Zulema Solana (Universidad Nacional de Rosario)

**Comité Editorial**

Celina Beltrán - Universidad Nacional de Rosario/Indec  
Cristina Bender - Universidad Nacional de Rosario  
Claudia Deco - Universidad Nacional de Rosario  
Silvia Rivero - Universidad Nacional de Rosario

**Comité de Lectura**

Gabriel G. Bès (asesor) Universidad Blaise-Pascal (GRIL) Clermont Fd., Francia  
Cristina Bender - Universidad Nacional de Rosario, Argentina  
Víctor M. Castel - InCiHuSA, CONICET, y FFyL, UNCUIYO, Mendoza, Argentina  
Claudia Deco - Universidad Nacional de Rosario, Argentina  
Daniel Guillot - CEDIA Consultora, Mendoza, Argentina  
Giovanni Parodi - Pontificia Universidad Católica de Valparaíso, Chile  
Zulema Solana - Universidad Nacional de Rosario, Argentina  
Dina Wonsever - Universidad de la República, Montevideo, Uruguay

## Indice

Análisis Automatizado de Sentimiento en Textos Breves de la Plataforma Twitter Fernando Balbachan, Diego Dell'Era	3
Árboles de clasificación y su comparación con análisis de regresión logística aplicado a la clasificación de textos académicos Celina Beltrán	15
Metodologías para la creación colaborativa de libros de texto Claudia Deco, Cristina Bender, Ana Casali, Raúl Kantor	27
Extracción de Candidatos a Términos del Dominio Médico a Partir de la categorización automática de Palabras Walter Koza	35
Los sistemas de ayuda en la navegación hipertextual Bárbara Méndez	45
Una propuesta para la extracción automática del sintagma adverbial Andrea Rodrigo, Rodolfo Bonino	53
Análisis automático de la gramática temprana Zulema Solana	65
Análisis automático de la interlengua de aprendientes de español: la presencia de determinantes indefinidos en el sintagma nominal núcleo Carolina Tramallino	75

## **Análisis Automatizado de Sentimiento en Textos Breves de la Plataforma Twitter**

**Fernando Balbachan**

Facultad de Filosofía y Letras, Universidad de Buenos Aires (UBA)  
Buenos Aires, Argentina  
Socialmetrix - NLP Research  
fbalbachan@socialmetrix.com

**Diego Dell'Era**

Facultad de Filosofía y Letras, Universidad de Buenos Aires (UBA)  
Buenos Aires, Argentina  
Socialmetrix - NLP Research  
ddellera@socialmetrix.com

### **Abstract**

In Computational Linguistics or Natural Language Processing (NLP), Sentiment Analysis in opinionated text is one of the most challenging standard tasks. Although some unsupervised learning approaches based on machine learning use statistical techniques such as Bayesian classifiers, semantic-oriented bigrams, etc. [1], the most successful works in the field lean on lexical-grammar resources under the form of ontologies with sentiment values and several basic syntax rules.

Our solution is based on such symbolic paradigm of lexical-grammar resources and local and global syntax rules. We detect early hints such as emoticons in order to come up with sentiment verdicts. Then, we pre-process the text by standardizing it and segment it into significant units smaller than sentences. Those small units are processed by a lemmatizer (Freeling) with POS-tagging optimized for Spanish. Thus, we get accurate lemmata for each unit of analysis. Those lemmata are valued by our own ontology -similar to SentiWordNet [2]- with more than 4,000 hand-annotated lemmata. Finally, we apply rules for modal changes in negated structures or modal subjunctive and polarity shifters for phrases such as *sin respeto*, *imposible de enojarse*, etc.

**Keywords:** Sentiment Analysis, Opinion mining, Ontologies, Freeling, SentiWordNet.

### **Resumen**

En el área de la Lingüística Computacional o Procesamiento de Lenguaje Natural (PLN), una de las tareas estándares más desafiantes es el análisis de sentimiento (*sentiment analysis*) en texto *opinionado*. Aunque algunos enfoques de aprendizaje no supervisado (*machine learning*) hacen uso de técnicas estadísticas como clasificadores bayesianos, bigramas de orientación semántica, etc. [1], los trabajos más exitosos en el campo recurren a recursos léxico-gramaticales bajo la forma de una ontología con valoración de sentimiento y diversas reglas de sintaxis básica.

Nuestra solución sigue tal enfoque simbólico de recursos léxico-gramaticales y reglas de sintaxis

local y global. En forma temprana detectamos indicios como emoticones para dar veredictos de sentimiento. Luego, pasamos a la etapa de pre-procesamiento de texto: estandarización del texto y segmentación en unidades significativas menores a la oración, las cuales pasan a ser procesadas por un lematizador optimizado para el español (Freeling) con anotación morfosintáctica (*POS-tagging*). De esta manera, obtenemos lemas muy confiables por cada unidad de análisis, los cuales son valorados por nuestra propia ontología -similar a SentiWordNet [2]- con más de 4.000 lemas anotados a mano. Finalmente, aplicamos reglas de cambios de modalidad para estructuras negadas y subjuntivo con modalidad irreal y reglas de polaridad para frases como *sin respeto*, *imposible enojarse*, etc.

**Palabras claves:** Análisis de sentimiento, Minería de texto *opinionado*, Ontologías, Freeling, SentiWordNet.

## 1. INTRODUCCION

En el área de la Lingüística Computacional o Procesamiento de Lenguaje Natural (PLN), una de las tareas estándares es el análisis de sentimiento (*sentiment analysis*) en texto *opinionado*. La dificultad de la tarea se percibe en que la medida de efectividad de estos mecanismos computacionales no llega, en general, al 85%. Incluso para los propios hablantes resulta a veces problemático dar un veredicto de análisis de sentimiento sobre un texto *opinionado*, ya que el soporte lingüístico de nuestras opiniones involucra frecuentemente fenómenos simultáneos pertenecientes a la estructura sintáctica, semántica, pragmática y discursiva del lenguaje, y aun de nuestro conocimiento de mundo [1].

Otro aspecto crucial para encarar esta tarea es la falta de disponibilidad de recursos. Por un lado, antes de la masiva irrupción de Internet y de las plataformas de redes sociales (Facebook, Twitter), el interés en un análisis computacional de opiniones era escaso [3]. Sólo recientemente diversas compañías y organizaciones académicas han incursionado en este tipo de proyectos de investigación como punto de partida para lo que se conoce en marketing como *Social Media Monitoring*, tomando en cuenta el inmenso volumen de información disponible bajo la forma de opiniones de clientes, prospectos de clientes, competidores, etc. Esto explica la notoria falta de recursos léxicos para lenguas que no sean inglés, e incluso la escasez de los mismos para esta lengua.

Finalmente, en lo que hace a este trabajo particular de análisis de sentimiento aplicado a textos breves de Twitter (tweets) en español, debemos destacar el desafío que significa trabajar con texto espontáneo, plagado de errores ortográficos y cierto desapego a las normas gramaticales (especialmente en lo que atañe a puntuación).

## 2. ANTECEDENTES

### 2.1. En los inicios de la tarea de análisis de sentimiento

Los primeros trabajos en el campo de análisis de sentimiento fueron desarrollados a partir de 2000.

Los enfoques en esta tarea apelan tanto al paradigma simbólico de recursos léxicos y reglas de manipulación de símbolos al paradigma estadístico de aprendizaje automático.

Entre los primeros enfoques simbólicos, muchos investigadores demostraron un gran ingenio para lidiar con la escasez de recursos anotados que sirvieran de guía inicial para la valoración de

opiniones. Por ejemplo, Kamps y Marx (2002) utilizan distancia semántica o *Minimal Path Length* MPL (distancia entre nodos) en WordNet<sup>1</sup> (una ontología computacional sin valoración de sentimiento disponible para varios idiomas) entra una palabra blanco, cuya orientación semántica se quiere conocer, y los adjetivos *good* (bueno) y *bad* (malo). Por ejemplo, la MPL entre *good* y *proper* es 2.

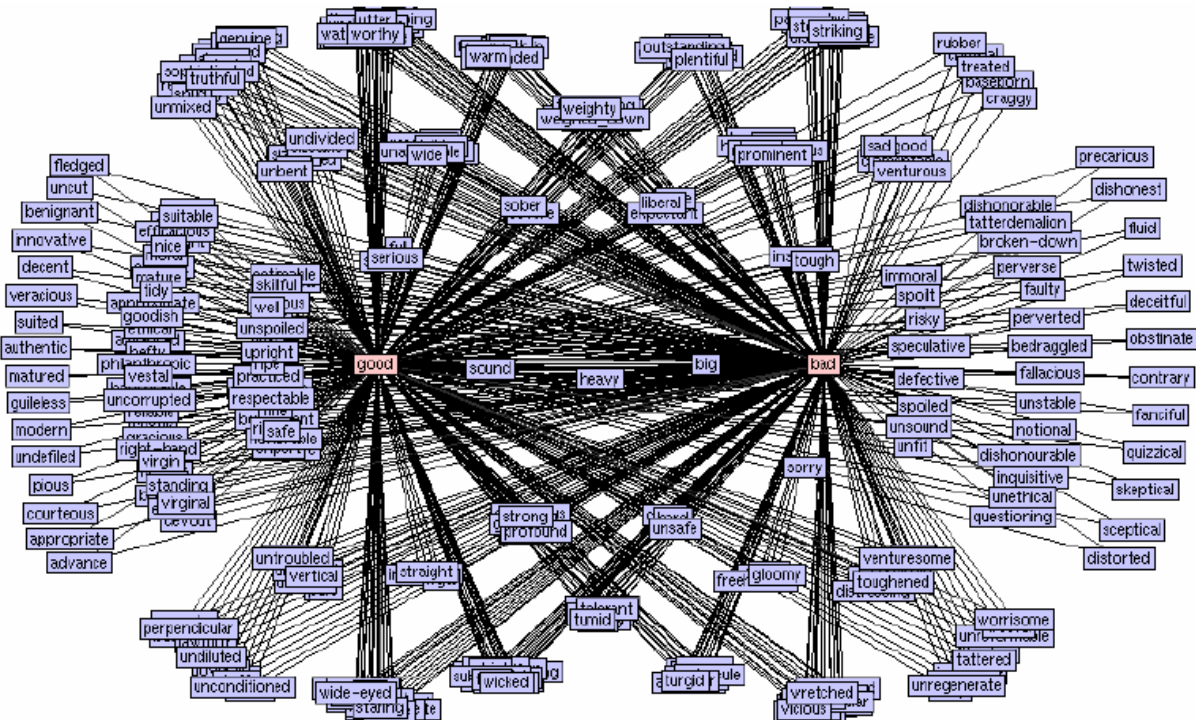


Figura 1: *Minimal Path Lengths* MPLs para algunos adjetivos del inglés, según Kamps y Marx (2002).

Si bien el trabajo de Kamps y Marx (2002) fue pionero en mostrar cómo computar valoraciones subjetivas de texto *opinionado*, la efectividad y el alcance de este enfoque son poco significativos, como los propios autores reconocen [4]. En primer lugar, llega un punto de separación de adjetivos en el que la distancia entre nodos ya no es índice confiable de valoración axiológica entre *bueno* y *malo* (la MPL de *noble* para *good* y para *bad* es 4). En segundo lugar, la ocurrencia de adjetivos axiológicos puede verse afectada por el alcance de un operador modal (negación, cambio de polaridad, etc.). Finalmente, considerar que el soporte lingüístico para opiniones está únicamente basado en adjetivos axiológicos es una concepción demasiado reduccionista: tal como sostiene la teoría de los subjetivemas [5], las palabras que actúan de soporte de la valoración subjetiva del hablante respecto del enunciado pueden ser verbos, sustantivos, adverbios y adjetivos.

Los seguidores del paradigma estadístico también aportaron trabajos iniciáticos para esta tarea. Turney (2002) [6] describe un clasificador no supervisado basado en la orientación semántica de bigramas extraídos mediante simples patrones morfosintácticos en un texto. La orientación semántica es calculada mediante el algoritmo PMI-IR (*Pointwise Mutual Information*). Este enfoque explota la poderosa noción de información mutua (probabilidad de co-aparición de dos términos) en función de dos conjuntos de palabras “semillas” positivas y negativas. Turney [6] reporta una medida  $F = 84\%$ . No obstante, como mencionan Cruz *et al.* [1] esta aparentemente alta efectividad puede deberse a una sobreadaptación al corpus de evaluación, compuesto de reseñas de automóviles en inglés. El género de las reseñas como textos breves presenta un cierto sesgo hacia una opinión polarizada, sea ésta positiva o negativa.

<sup>1</sup> <http://wordnet.princeton.edu/>

## 2.2. Trabajos recientes y análisis de sentimiento en español

Más recientemente, la tarea de análisis de sentimiento atrajo un mayor interés por soluciones más integradoras y abarcativas. Los trabajos más exitosos recurrieron a recursos léxico-gramaticales bajo la forma de una ontología con valoración de sentimiento y diversas reglas de sintaxis básica.

Enmarcado en el paradigma simbólico, Taboada, Brooke *et al.* (2010) [7] desarrollaron el *Semantic Orientation Calculador* (SO-CAL), un conjunto de recursos léxicos con valoración de sentimiento que se ven afectados por intensificadores, índices de modalidad (negación y contrafáctica) y reglas de sintaxis básica. Esta solución integral para el inglés ha sido evaluada en distintos *corpora* de reseñas de dominio específico (hotelería, tecnología, películas, etc.) reportando medidas F que varían entre 74% y 90%, resultados auspiciosos por demás.

Estos enfoques basados en recursos léxicos guiaron nuestra investigación inicial hacia un mecanismo de análisis de sentimiento para el español. Desafortunadamente, nos topamos con el obstáculo de la ausencia total de recursos léxicos para este idioma. Desde 2007, se encuentra disponible un poderoso recurso léxico para el inglés: SentiWordNet <sup>2</sup>, una ontología con valoración de sentimiento en 3 dimensiones (positivo, negativo, neutral) y una amplia cobertura de acepciones. Definitivamente necesitábamos de algo similar para el español. Diversos trabajos en lenguas que no sean el inglés han apelado a enfoques que incluían etapas de traducción automática, pero sin lograr un éxito mayor.

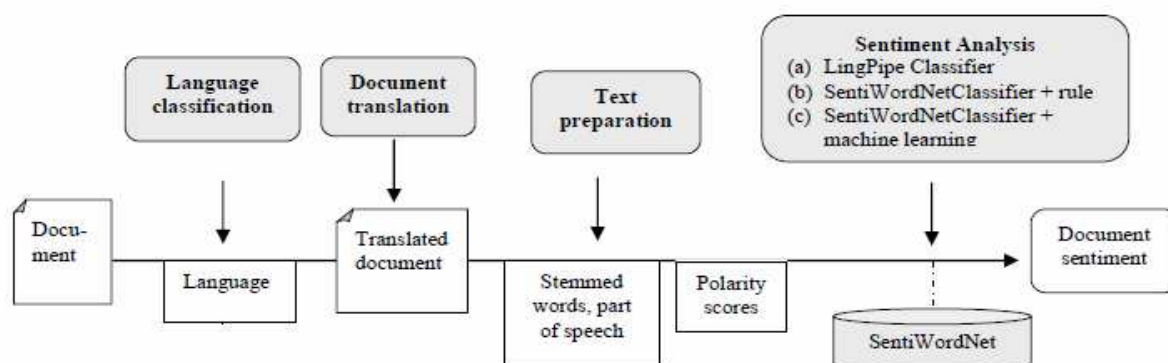


Figura 2: Procesamiento para análisis de sentimiento en idiomas que no sean el inglés (Denecke 2008).

## 3. NUESTRA SOLUCIÓN

### 3.1. Arquitectura general de la solución

Nuestra solución integral, desarrollada para Socialmetrix<sup>3</sup>, sigue el enfoque simbólico de recursos léxico-gramaticales y reglas de sintaxis local y global. En forma temprana detectamos indicios como emoticones para dar veredictos de sentimiento. Luego, pasamos a la etapa de pre-procesamiento de texto: estandarización del texto y segmentación en unidades significativas menores a la oración, las cuales pasan a ser procesadas por un lematizador optimizado para el español (una de las poderosas herramientas de análisis de la suite *open source* y multilenguaje Freeling<sup>4</sup>) con anotación morfosintáctica (*POS-tagging*). De esta manera, obtenemos lemas muy confiables por cada unidad de análisis, los cuales son valorados por nuestra propia ontología de

<sup>2</sup> <http://sentiwordnet.isti.cnr.it>

<sup>3</sup> <http://socialmetrix.com>

<sup>4</sup> <http://nlp.lsi.upc.edu/freeling/> y <http://nlp.lsi.upc.edu/freeling/doc/userman/userman.pdf>

valoración de sentimiento en español -similar a SentiWordNet [2]- con más de 4000 lemas anotados a mano. Finalmente, aplicamos reglas de cambios de modalidad para estructuras negadas y subjuntivo con modalidad irreal y reglas de polaridad para frases como *sin respeto*, *imposible enojarse*, etc.

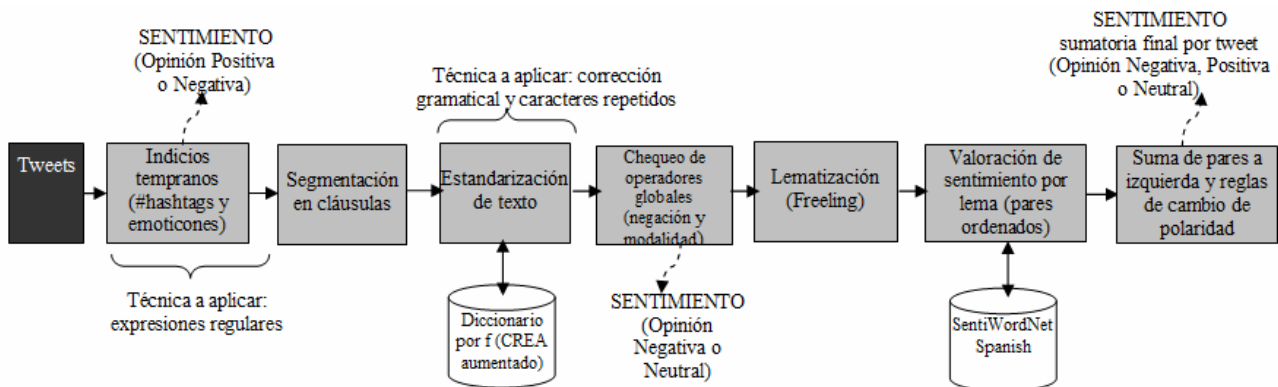


Figura 3: Arquitectura general de la solución de análisis de sentimiento en español para Twitter

## 3.2. Descripción detallada de cada etapa de procesamiento

### 3.2.1. Indicios tempranos (hashtags y emoticones)

En el uso del lenguaje en Internet [8] es una práctica común que los usuarios marquen su propia valoración subjetiva respecto de lo expresado a través de emoticones, como por ejemplo :) o :( , o hashtags (#WTF, #FTW, etc.). Este recurso supralingüístico a menudo no ha sido muy tomado en cuenta por la academia para el procesamiento de lenguaje natural en modelizaciones más abarcativas. No obstante, diversas implementaciones de la industria se basan exclusivamente en este tipo de indicios (por ejemplo, Tweetfeel<sup>5</sup>). En nuestro caso, al estar trabajando con opiniones expresadas en textos breves (tweets) de la plataforma Twitter, estadísticamente el 5% de los datos a analizar en cualquier set de datos hacían uso de este recurso.

Por un lado, la incidencia de estos indicios tempranos de emotividad sobre el set de datos totales no es estadísticamente despreciable. Por otro lado, disponemos de un método trivial para reconocer o identificar estas marcas a través de expresiones regulares, sin hacer uso de ningún recurso léxico o regla adicional. Así pues, logramos un porcentaje de efectividad muy alto (entre 85% y 90%) en la tarea de determinar si el sentimiento del tweet es positivo o negativo, porque en general los indicios positivos y negativos no se superponen ni se contradicen.

En esta etapa, no obstante, no podemos dar cuenta del estado NEUTRAL para los tweets, ya que no existen indicios tempranos de tal marcación subjetiva. La gran masa de todos los tweets que no incluyen estos indicios (aproximadamente 95%) debe continuar su camino en el procesamiento de cascada, tal como se describe en la figura 3.

### 3.2.2. Segmentación

La segmentación del texto del tweet involucra decisiones de diseño fundamentales. Las propias convenciones de marcación en Twitter para retweets (RT), menciones, recomendaciones (#followfriday) vínculos (links) y vínculos acortados obligan a tomar ciertas precauciones a la hora de aislar el soporte lingüístico que podría conllevar texto opinado.

<sup>5</sup> <http://www.tweetfeel.com>

En el caso de nuestro algoritmo, decidimos segmentar la unidad de análisis en base a una lista de signos de puntuación y conjunciones coordinantes y subordinantes. Así pues, la unidad de análisis resulta típicamente menor que una oración. Como se verá más adelante, uno de los inconvenientes más difíciles de subsanar a la hora de analizar sentimiento en función de la ocurrencia de subjetivemas es el alcance de negaciones y operadores de modalidad sobre los mismos (véase sección 3.2.6 *Operadores globales y operadores locales de cambios de polaridad*).

- (1) [*Te felicito*] *porque* [*has mejorado*]. (valoración positiva y positiva, respectivamente)
- (2) [*No me alegra*] *que* [*hayas mejorado*]. (valoración neutra o negativa y positiva, respectivamente)
- (3) [*Me alegra*] *que* [*no hayas mejorado*]. (valoración positiva y neutra o negativa, respectivamente)
- (4) [*Mejoraste*] *pero* [*todavía me das asco*]. (valoración positiva y negativa, respectivamente)
- (5) [*No mejoraste*]. (valoración neutra o negativa)

Esta problemática también se da en español con índices de modalidad irreal (subjuntivo irreal) o del enunciado o volitiva de la enunciación (modo imperativo) [5].

- (6) [*Dudo*] *que* [*haya mejorado*]. (valoración neutra y neutra -por subjuntivo irreal-, respectivamente)
- (7) [*Mejorá*]. (valoración negativa o neutra)

Existe bastante bibliografía especializada en el problema específico de detectar el alcance de las negaciones, desde enfoques simbólicos con expresiones regulares y diccionarios de frases negativas hasta métodos estadísticos de *machine learning* como *Support Vector Machine* (SVM) o clasificadores bayesianos ingenuos [9]. El trabajo de Goryachev *et al.* (2008) [9] demuestra que los enfoques simbólicos tienen nuevamente una efectividad mayor que la de los métodos estadísticos para determinar el alcance de las negaciones.

En este sentido, y ante la falta de un mecanismo que entendiera la estructura sintáctica de la oración a alto nivel (*chunking* o *shallow parsing*), nos vimos obligados a minimizar la extensión de las cláusulas de modo tal de garantizar que el alcance de operadores de modalidad no excediera el ámbito de la localidad inmediata de ocurrencia de la partícula negativa hacia la derecha (cambio de polaridad para expresiones privativas como las encabezadas por *sin*) o el ámbito de la globalidad de la cláusula (operadores globales de negación o de modalidad). Para mayor información, véase la sección 3.2.6 *Operadores globales y operadores locales de cambios de polaridad*.

### 3.2.3. Estandarización de texto (distancia mínima de edición y caracteres repetidos)

El uso del lenguaje en plataformas de redes sociales en Internet representa un interesante desafío para la modelización en lingüística computacional. Al tratarse de escritura espontánea y creativa con marcados rasgos de oralidad, “la ortografía resulta sumamente variable y [...] ha sido despojada de su función normalizadora y puesta al servicio de la expresividad de los usuarios” [8]. En este sentido, nuestro proceso de estandarización del texto debe abarcar no sólo una tarea de corrección ortográfica, sino también una tarea de análisis de caracteres repetidos (recurso muy tipificado como intensificador de carga emotiva o expresión enfática).

Con respecto al recurso de intensificación emotiva a partir de la repetición de caracteres, especialmente en el caso de vocales (por ejemplo, “*te amooooo*”), implementamos un control basado en expresiones regulares con *back reference* que reemplaza la ocurrencia de dos o más de estos caracteres por uno solo, a excepción de los grupos *cc, ll, rr*.

Para la tarea de corrección ortográfica, recurrimos al algoritmo de Levenshtein [10] y su noción de “distancia”. Se define como distancia a la medida en que dos términos difieren en cuanto los caracteres que los componen. Así, una distancia de valor 0 indica que dos términos son idénticos,



mientras que valores mayores de distancia indican una creciente disimilitud. La máxima distancia está acotada por la longitud del término más extenso. El algoritmo calcula la distancia entre dos términos como la mínima cantidad de operaciones de edición (inserción, borrado y sustitución) necesarias para convertir a un término en el otro.

En la tabla 1 los términos se representan como secuencias de caracteres alineados, y las operaciones de edición como una matriz con los “costos” acumulados. Si bien es posible asignar un costo diferencial a cada una de estas operaciones, para penalizar ciertos cambios considerados menos frecuentes en el ámbito del que provengan los términos, pruebas preliminares demostraron que se podían obtener iguales resultados con un valor uniforme de 1 para todas las operaciones.

		T	E	C	O	R
		1	2	3	4	5
T	1	0	1	2	3	4
E	2	1	0	1	2	3
M	3	2	1	1	2	3
O	4	3	2	2	1	2
R	5	4	3	3	2	1

Tabla 1: Distancia mínima de edición entre la no-palabra TECOR y la palabra TEMOR con un valor acumulado de 1

Con excepción de los nombres propios, reconocidos por el uso de mayúsculas, una no-palabra es un token alfabético que no ha sido validado contra nuestro diccionario, el cual se compone de la lista completa de formas del *Corpus de Referencia del Español Actual* (CREA)<sup>6</sup> de la Real Academia Española (RAE), ordenada por frecuencia de uso y aumentada con las formas conjugadas del español rioplatense (voseo), totalizando aproximadamente 190.000 formas. La no-palabra genera una serie de posibles candidatos a palabras en base a las sustituciones, inserciones y borrados sucesivos de cada carácter que la componen (distancia de edición igual a 1). Nuestro algoritmo reemplaza la no-palabra con la palabra que surja de esta lista de candidatos a distancia de edición igual a un carácter y que esté en nuestro diccionario con frecuencia más elevada. Obviamente, este método de corrección ortográfica resulta más confiable a medida que la no-palabra es más extensa y, por tanto, existen menos candidatos a distancia de un carácter que puedan resultar palabra.

### 3.2.4. Lematización (Freeling)

La lematización es el proceso de obtención de una forma base (*lema*) a partir de formas flexionadas y derivadas morfológicamente [11]. En lenguas como el español, donde el paradigma verbal y flexivo es muy extenso, este proceso de modelización adquiere vital importancia a la hora de minimizar esfuerzos de anotación de recursos léxicos y maximizar la cobertura de los mismos.

Numerosos trabajos en lingüística computacional recurren a herramientas especializadas como Freeling para este tipo de tareas [1][12]. En nuestro caso aprovechamos la salida morfosintácticamente etiquetada de Freeling como un lematizador optimizado. Freeling resuelve muy bien el escollo de la ambigüedad morfosintáctica del español, mediante un *POS-tagger* (etiquetador morfosintáctico) a partir de reglas simbólicas y cálculo de probabilidades. No obstante, la variedad dialectal del español rioplatense representa un problema, porque las formas verbales pertenecientes al pronombre personal *vos*, cuyo uso se encuentra extendido en toda la población, no son reconocidas por el español de Freeling. Por lo tanto, configuramos el archivo *afixos.dat* de la herramienta para obtener el verbo en infinitivo a partir de las formas conjugadas de voseo.

<sup>6</sup> <http://corpus.rae.es/lfrecuencias.html>

Otra modificación al proceso de lematización de Freeling consistió en derivar por afijación los adverbios terminados en *-mente*, ya que nuestra ontología de valoración de sentimiento (*SentiWordNetSpanish*) incluía los primeros y no los últimos. Cabe recordar que el uso de adverbios de modo terminados en *-mente* (*afortunadamente*, *desgraciadamente*, etc.) está muy difundido como índice de modalidad del enunciado [5] y, por lo tanto, resulta un excelente vehículo para expresar la opinión del hablante.

En el diseño de nuestra etapa de lematización nos topamos con otro obstáculo. El módulo de reconocimiento de entidades nombradas (*Named-Entity Recognition* NER) de Freeling considera como una única unidad varios tokens seguidos que comiencen con mayúscula o estén unidos por partículas unitivas (por ejemplo, *Banco de Bilbao y Vizcaya*). Esta característica interfiere con otra práctica difundida en Internet: la escritura TOTALMENTE EN MAYÚSCULAS como recurso enfático. Decidimos, por lo tanto, deshabilitar en Freeling el módulo NER para multipalabras.

### 3.2.5. Valoración de carga subjetiva (*SentiWordNetSpanish*)

A la hora de planificar la actividad de anotación manual para generar recursos léxicos es importante contar con criterios que optimicen los esfuerzos. Nuestra inspiración original provino del importante recurso léxico *SentiWordNet*<sup>7</sup>.

Dado el procesamiento en cascada descrito en la figura 3, sabíamos de antemano que sólo las palabras que estuvieran en nuestro diccionario de estandarización y en el diccionario que usa Freeling para su configuración (archivo *dicc.src*) serían tomadas en cuenta para la asignación de sentimiento. Comenzamos por obtener los lemas de Freeling y ordenarlos en función de la frecuencia de uso de nuestro diccionario, basado en el CREA. De ese modo, nos asegurábamos comenzar por los lemas más frecuentes. En la lista resultante de 40.000 palabras desechamos las primeras 300 palabras por ser consideradas palabras funcionales con altísima frecuencia y sin valoración de sentimiento alguno.

En cuanto a los criterios de asignación de puntaje positivo y negativo (en una escala de 0 a 100), recurrimos a nuestros juicios como hablantes nativos y nuestro conocimiento de la teoría de los subjetivemas [5]. De ese modo, obtuvimos 4.000 lemas muy confiables con valoración (un 10% aproximadamente de todos los lemas), los cuales representan una cobertura de aproximadamente 40.000 formas no lematizadas del español, en virtud del extenso paradigma verbal y de la rica morfología para lemas nominales del español, todas ellas conllevando valoración de sentimiento.

Freeling dispone de un diccionario de aproximadamente 5.000 giros lingüísticos del español (archivo *locutions\_extended.dat*), el cual puede ser ampliado manualmente. Ésta es una interesante característica que permite identificar frases por sobre la ocurrencia de las palabras que las componen (*multiword recognition*). Así pues, podemos refinar los criterios de anotación, tal como muestra la tabla 2:

Lema o frase	Valoración positiva	Valoración negativa
<i>bueno</i>	50	0
<i>buenos_días</i>	0	0
<i>bueno_para_nada</i>	0	40

Tabla 2: Anotación diferenciada para lemas o frases

<sup>7</sup> <http://sentiwordnet.isti.cnr.it>

### 3.2.6. Operadores globales y operadores locales de cambios de polaridad

A esta altura estamos en condiciones de entender el procesamiento algorítmico de valoración de sentimiento: simplemente nos hacemos pasar cada token de cada unidad de análisis (cláusula) a través de nuestra SentoWordNetSpanish, reemplazando cada token por pares ordenados con las dimensiones (positivo,negativo) que correspondan. Si un token no es encontrado en la SentiWordNetSpanish, se le asigna el par (0,0). Los pares ordenados de una cláusula se van reduciendo (sumando) a izquierda. No obstante, existen dos tipos de ocurrencias de palabras que generan sus propias reglas de procesamiento: operadores globales para negación y modalidades y operadores locales de cambio de polaridad (por ejemplo, expresiones privativas como *sin* o *falta de*).

Los operadores globales son una lista de expresiones negativas (*no*, *ni*, *nadie*, *negar*, etc.) cuya ocurrencia provoca que la cláusula sea considerada negativa -asignando valor (0,60) a la cláusula-, independientemente de cualquier subtotal de arrastre de pares ordenados, constituyendo un fuerte sesgo hacia las opiniones negativas. Evaluamos distintas versiones de sesgo de los operadores globales, por ejemplo hacia la neutralidad -asignando valor (0,0) a la cláusula-; pero las pruebas sugirieron la conveniencia del sesgo hacia opiniones negativas ante la ocurrencia de dichos operadores globales (véase tabla 5). Los ejemplos (8) a (10) ilustran esta disyuntiva entre las opiniones neutrales y negativas para la ocurrencia de expresiones de negación:

- (8) *Mejoraste.* (valoración positiva)
- (9) *No mejoraste.* (valoración neutra o negativa)
- (10) *Nadie me ayudó.* (valoración negativa)

Los operadores locales producen cambios en la polaridad de los subtotales de arrastre de pares ordenados a derecha. Se trata de listas de palabras y expresiones como *sin*, *falta de*, *difícil de*, *deber*, etc. cuyo funcionamiento queda ilustrado en los ejemplos (11) y (12):

- (11) *Se manejó con respeto.* (valoración positiva)
- (12) *Se manejó sin respeto por las normas.* (valoración negativa)

En nuestro algoritmo, simplemente invertimos la polaridad del subtotal de pares ordenados a derecha de la ocurrencia de un operador local dentro de una cláusula.

## 4. EVALUACIÓN

Nuestra solución fue inicialmente evaluada en un corpus de 800 tweets de la industria de la telefonía celular, llegando a tener una efectividad de 76%, respecto del *gold standard* anotado a mano. Como grupo de control evaluamos el desempeño de los propios humanos en esta tarea. Por ejemplo, en los papers en donde se comparan clasificadores humanos entre sí, la tasa de *inter-confiabilidad* (*interreliability ratio*) no supera el 70%, incluso para desacuerdos no fatales (cualquier discrepancia que no sea positivo por negativo y viceversa) [13].

Para nuestro set de datos, los clasificadores humanos cometieron errores más o menos significativos de inconsistencias en el 17% de los casos en promedio, por lo que la comparación final entre humanos y el algoritmo podría ser enunciada como un mecanismo muy costoso y muy lento (clasificación manual con humanos) con un éxito del 83% *versus* un mecanismo algorítmico sin costo alguno, instantáneo y escalable con un éxito del 76%.

1	TEXT (para ver analisis cuantitativo y confusion matrix ir al final de la tabla) STEP 2... SENTI DETECTED BY HINTS	ASSIGNED	ERROR POS	ERROR NEG	ERROR NEUTR
2					
3	ya se estan vendiendo las entradas para #RedHotChiliPeppers en Chile por TicketMaster y 20% Desc @Entel http://bit.ly/9qDQGm:D	POSITIVE			
4	RT @entel: ¡Ya tenemos el Sony Ericsson Xperia Arc! Ven a conocerlo :) http://on.fb.me/q5qKwm	POSITIVE			
5	@entel Tres meses con un iphone y se echa a perder?! Pff!! Cambio de equipo para @claupinov !!	NEGATIVE			
6	@entel hola! Una pregunta: a mi hermana le robaron hoy su Iphone con plan. Hay como inhabilitarlo o seguirlo? Que pueo hacer? Gracias :)	NEGATIVE			1
7	@entel ¿les cuesta mucho avisar cuando se esta acabando la bolsa d navegacion? DEPREDADORES D MIERDA!	NEGATIVE			
8	Y si el comercial de @entel es cierto ?? Se les lesiona Messi y Nos encontramos con Argentina en 2a Ronda ??... Ahí los quiero ver :D	POSITIVE			1
9	Mierda esta que no me deja conectarme digan algo en entel @entel x favor... Entel no está ni ahí	NEGATIVE			
10	@hlfx Pero yo ni siquiera soy cliente de @entel y me llaman para ofrecer productos! #FAIL	NEGATIVE			
11	Esto es velocidad mierda!!! 3 y + en cargar en mi super conexión de internet móvil de @entel @entel_ayuda	NEGATIVE	1		
12	RT @entel: ¡Ya tenemos el Sony Ericsson Xperia Arc! Ven a conocerlo :) http://on.fb.me/q5qKwm	POSITIVE			
13	@danielexhuevo @TWITCARTV @SoledadOnetto @entel @coseche Buena huevo. hermosa ella :-)	POSITIVE			
14	RT @BancoGalicia: #TipQuiero! ¿Sabías que podés pagar el servicio de cable con puntos Quiero! Con solo 225 puntos ahorras un 40% Ht	NEGATIVE	1		
15	@BancoGalicia Buenos días! Ayer me contestaron el primer mail con mis datos..y les contesté lo de la sucursal, pero no recibí respuesta :(	NEGATIVE			
16	@BancoGalicia les mandé mi mail con los datos :) Cuanto demorarán en contestar?	POSITIVE			1
17	Excelente el servicio de atención de @entel_ayuda... rapido, claro y efectivo :)	POSITIVE			
18	Hasta que me llevo la Galaxy Ace, pero no gracias a @entel_ayuda, nada que decir en todo caso del call center, ellos la llevan :) cc @entel	POSITIVE			
19	@entel_ayuda ok, mensaje enviado :)	POSITIVE			
20	@entel_ayuda enviado :)	POSITIVE			
21	@entel_ayuda no.. mi plan se activa mañana! :(	NEGATIVE			
22	gracias chicos de @entel_ayuda por la solución ultra flash!! :)	POSITIVE			
23	@entel_ayuda muchas gracias!!!! :)	POSITIVE			
24	@MovistarArg Muchas gracias ^MS, ahora me quedo mas claro :)	POSITIVE			
25	Será un Motorola Spice ??? #motorola #spice #motorolaspice ?? #loquehayenlacaja ?! OJALAAA!!! LO QUIERDOO! @PersonalAr	NEGATIVE	1		
26	@PersonalAr tengo un iPhone4 y ya probé de todo.. Tengo 3 amigos c iPhone y les pasa lo mismo.. Son Uds. El problema, saludos! =)	POSITIVE		1	
27	@PersonalAr hice todo! ya tengo 4 reclamos hechos! ya nose que hacer... si me compro un equipo nuevo y sigue el problema que pasa?!	NEGATIVE			
28	@PersonalAr Hoy funciona. Qué significa red ocupada?! Y por qué en Constitución, plena CABA, JAMÁS (de los jamases) hay señal?!?!	NEGATIVE			
29	SUBTOTAL POR OPINION POSITIVA, NEGATIVA, NEUTRO		3	1	5
30					
31			TOTAL	9	
32	Distribucion: 38 POSITIVE 17 NEGATIVE DISTRIBUCION REAL 37 POSITIVE 13NEGATIVE 5 NEUTRAL				
33	Error Positive 3/37 Negative 1/13 Neutro 5/5				
34	Efectividad 92% 92,4% 0%				
35	Efectividad total 100-(9/55) = 83.7%				
36	Tráfico de 6.9% de 800 tweets				

Tabla 3: Efectividad total de etapa de indicios tempranos sobre 55 tweets en un corpus inicial de 800 tweets

1	TEXT SENTI DETECTED BY SENTIWORDNET + BASIC SYNTAX	SENTI ASSIGNED	ERROR POS	ERROR NEG	ERROR NEUTR	SENTI CORRECTED
2						
3	RT @cristianruiz: @zonaEntel @entel y que paso con el 2x1 con nevados de Chillan?	POSITIVE				POSITIVE
4	RT @zonaEntel: Mañana los clientes @entel tienen 2x1 en Valle Nevado. http://awe.sm/5OH6R,	POSITIVE				POSITIVE
5	@entel Muy buena la serie #EICr@ck l,	POSITIVE				POSITIVE
6	@entel como se puede hacer valido el descuento para Red Hot Chili Peppers ??,	NEUTRAL				NEUTRAL
7	Alerta!! molestia de Messi, @entel lo predijo ajajajajaj.	NEGATIVE			1	NEUTRAL
8	en la segunda se les lesiona Messi, NO EXISTEN... que cracks los de @entel ajajajajajaj.	NEUTRAL				NEUTRAL
9	Hasta ahora va a pasar exactamente lo que le dijo el pelado del spot de @entel al argentino, camino a Mendoza.,	NEUTRAL				NEUTRAL
10	Fome el comercial de @entel para el partido contra Uruguay. Se ve que ya se les acabaron las ideas al grupo creativo xd,	POSITIVE		1		NEGATIVE
11	Odio el comercial de @Entel ... Garra Charrúa es lo que van a ver el Viernes!!!,	NEGATIVE				NEGATIVE
12	Los del comercial de @entel le están achuntando, casi se lesiona Messi xD,	NEUTRAL				NEUTRAL
13	@personalAR seguramente es un mejor servicio 3G para la gente con #Black #loquehayenlacaja,	POSITIVE		1		NEGATIVE
14	@PersonalAR adentro de la caja hay menos redes 3G colapsadas? #loquehayenlacaja,	NEUTRAL		1		NEGATIVE
15	@PersonalAR buen dial estoy llamando hace una hora para hacer un reclamo y no me atienden.. como puedo solucionarlo?,	NEGATIVE				NEGATIVE
16	@PersonalAR Ok voy a estar esperando su respuesta,	NEGATIVE			1	NEUTRAL
17	@PersonalAR Por unos minutos mi celular dejo de tener el menu : PERSONAL :!,	NEUTRAL				NEUTRAL
18	@PersonalAR me deben un Iphone,	NEUTRAL		1		NEGATIVE
19						
20						
21	muestreo hay#debe	6W48	43W58	50W48		
22	ERROR SUBTOTAL POR OPINION POSITIVA, NEGATIVA, NEUTRO		10	127	73	
23						
24	Distribucion total 192 POSITIVE 238 NEGATIVE 315 NEUTRAL	TOTAL		210		
25	Distribucion corregida a mano extrapolada de muestreo en total 240 POSITIVE 290 NEGATIVE 240 NEUTRAL	TOTAL EXTRAPOLADO = 31*5 = 155				
26	Error extrapolado Positive 10/240 Negative 95/290 Neutro 50/240	(muestreo de 154 sobre 745)				
27						
28	Efectividad total step 3 = 100-(31/154) = 80%					
29	tráfico de 93,1% de 800 tweets					

Tabla 4: Muestra sobre un corpus inicial de 745 tweets (exceptuando los 55 tweets detectados por indicios tempranos)

input total: 745
input total_positive: 120
input total_negative: 333
input total_neutral: 292
output_true_positives: 108
output_true_negative: 280
output_true_neutral: 169
output_false_positive: 62
output_false_negative: 95
output_false_neutral: 31
recall for POSITIVE: 90%
precision for POSITIVE: 63%
recall for NEGATIVE: 84%
precision for NEGATIVE: 74%
recall for NEUTRAL: 57%
precision for NEUTRAL: 84%
F-score for POSITIVE: 76.7647058824 %
F-score for NEGATIVE: 79.3753753754 %
F-score for NEUTRAL: 71.1883561644 %
<b>F-score total (ponderando incidencia de los 3 sentimientos en total): 75.76473076 %</b>

Tabla 5: Métricas totales para efectividad (confusion matrix) sobre 745 tweets en un corpus de 800 tweets (exceptuando los 55 tweets detectados por indicios tempranos, véase tabla 3)

## 5. CONCLUSIONES

El análisis automatizado de sentimiento es aún un campo fértil y vasto para ser explorado. Si bien el estado del arte para esta tarea ha mostrado recientes avances en los resultados, la misma diversidad del soporte lingüístico de la opinión humana torna complejo el problema. Uno de los aspectos más difíciles de modelizar es los que se conoce como análisis de sentimiento basado en entidades (*entity-based sentiment analysis*) [3]. Es decir, muchas veces las opiniones, incluso en textos breves, conllevan polaridades contrapuestas respecto de diversas entidades, como en el siguiente ejemplo:

(13) *Obama hizo un excelente trabajo con esa maldita pérdida de petróleo en New Orleans.*

Este tipo de opiniones representan un gran escollo para los algoritmos de análisis de sentimiento. La asignación de polaridad positiva y negativa a distintas entidades en la misma cláusula (*Obama y pérdida de petróleo*) hacen que sea imprescindible recurrir a una estrategia de *full parsing* complejo para entender el enunciado y a robustos algoritmos de extracción de entidades (*Named-Entity Recognition* NER). Estos enfoques más avanzados resultan un gran atractivo para investigaciones de marketing en empresas que pueden así refinar las opiniones positivas y negativas de clientes respecto de características en particular de productos.

Nuestra solución basada en reglas y recursos léxicos ha demostrado una razonable eficacia para el análisis de textos breves. No obstante, la evaluación del mismo algoritmo en textos más extensos y con dinámicas textuales complejas ha resultado en una drástica reducción de la efectividad. Esto puede deberse a la complejidad propia de los textos elaborados en base a los recursos retóricos y discursivos que se ponen en juego, como argumentaciones concesivas, polifonía [5], etc.

Aun así, una de las principales fortalezas de nuestro enfoque es que brinda la posibilidad de ampliar la cobertura en función de anotaciones manuales modulares en la valoración de SentiWordNetSpanish, en el agregado de palabras o frases nuevas a nuestro diccionario y a Freeling, etc. Incluso, existe la posibilidad de adaptar los recursos a diferentes dominios específicos o temáticos. Por ejemplo, en la industria de la telefonía móvil, palabras como *interferencia*, *cobertura* o frases como *caída\_de\_señal* seguramente serán valoradas de distinta manera que en sus

usos coloquiales generales.

En definitiva, estamos convencidos de que el análisis de sentimiento es un muy reciente campo de aplicación práctica de la investigación académica en lingüística computacional con una clara articulación con la industria. Y en esta convergencia de intereses puede hallarse una muy fecunda interacción para nuevos hallazgos y avances en NLP.

## Referencias

- [1] Cruz, F., Troyano J., Enríquez, F. y Ortega J. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del Lenguaje Natural*, (41):73-80, 2008.
- [2] Denecke, K. Using SentiWordNet for multilingual sentiment analysis. En *Proceedings of the IEEE 24th International Conference on Data Engineering Workshop*: 507-512, 2008.
- [3] Liu, B. Sentiment Analysis: a multi-faceted problem. *IEEE Intelligent Systems* 25(3):76-80, 2010.
- [4] Kamps, J. y Marx, M. Words with attitude. En *1<sup>st</sup> International WordNet Conference*: 332-341, 2002.
- [5] Kerbrat Orecchioni, C. *La enunciación*. Buenos Aires, Edicial, 1993.
- [6] Turney, P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*: 417-424, 2002.
- [7] Taboada, M., Brooke J., Tofiloski, M., Voll, K. y Stede, M. Lexicon-based methods for sentiment analysis. *Computational Linguistics* (1):1-41, 2010.
- [8] Giammatteo, M. y Albano, H. El español en Internet: una mirada a su evolución en los fotologs. *Revista Digital Universitaria UNAM* 10(3):1-17, 2009.
- [9] Goryachev S., Sordo M., Zeng Q.T. y Ngo L. Implementation and evaluation of four different methods of negation detection. DSG reporte técnico, 2008.
- [10] Jurafsky, D. y Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey, Prentice-Hall, 2000.
- [11] Manning, C., Raghavan P. y Schütze, H. *Introduction to Information Retrieval*. Cambridge (Reino Unido), Cambridge University Press, 2008.
- [12] Atserias, J., Casas, B. , Comelles, E., González, M., Padró L. y Padró M. Freeling 1.3: Syntactic and semantic services in an open-source NLP library. En *Proceedings of the 5<sup>th</sup> International Conference on Language (LREC'06)*:48-55, 2006.
- [13] Gryc, W. y Moilanen, K. Leveraging textual sentiment analysis with social network modeling: Sentiment Analysis of political blogs in the 2008 U.S. presidential election. En *Proceedings of the "From Text to Political Positions" workshop (T2PP 2010)*, Vrije Universiteit, Amsterdam, 2010.

# **Árboles de clasificación y su comparación con análisis de regresión logística aplicado a la clasificación de textos académicos**

## **Classification trees and a comparison with an analysis of logistic regression applied to the classification of academic texts**

**Celina Beltrán**

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

### **Abstract**

The problem of unit classification into known groups or populations is of great interest in statistics and several techniques have been developed to serve this classification purpose. This work presents a comparison of the classification tree and a logistic tree technique for text classification according to the field to which these texts belong (BIOMETRY and PHYLOSOPHY).

The development of the techniques has been measured with a Wrong Classification Rate (WCR) calculated over a sample of texts not included in the model estimation and tree construction. The classification tree showed a WCR lower than that of the logistic model, and humanistic texts have been classified with more accuracy.

The resulting WCR generated by the classification tree was 10% (17 percentage points within the Biometrics corpus and 3 percentage points within the Philosophy one). As for the logistic regression model, there was a global result of 20%, and 17 percentage points and 23 percentage points within the Biometrics and Philosophy corpuses.

**Key words:** Multivariate logistic regression, classification trees, automatic analysis of texts.

### **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. En este trabajo se presenta la comparación de la técnica de Árboles de Clasificación y Regresión logística para la clasificación de textos según la disciplina a la que pertenecen (BIOMETRIA y FILOSOFIA).

El desempeño de las técnicas fue medido con la Tasa de Mala Clasificación calculada 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 humanísticos.

La TMC obtenida con el árbol de clasificación fue de 10% (17% dentro del corpus de Biometría y 3% en Filosofía) mientras que con el modelo de regresión logística fue de 20% en forma global y 17% y 23% respectivamente dentro de los corpus de Biometría y Filosofía.

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

## 1. INTRODUCCION

El problema de la clasificación de unidades en grupos o categorías conocidas es de gran interés en estadística. Esto ha hecho que se desarrollaran diversidad de técnicas para cumplir este propósito. Entre ellas podemos citar al análisis discriminante, la regresión logística y los árboles de clasificación. Si bien el análisis discriminante es una de las técnicas más utilizadas para clasificación, el no cumplimiento del requerimiento de normalidad multivariada hace necesario utilizar técnicas alternativas, como la regresión logística y los árboles de clasificación, que no requieran dicho supuesto.

Este trabajo sigue la línea de investigación iniciada en Beltrán (2010) donde se busca evaluar las técnicas multivariadas aplicadas a la caracterización y clasificación de textos académicos. Este trabajo utiliza al analizador morfológico Smorph, implementado como etiquetador, para asignar una categoría morfológica a cada una de las ocurrencias lingüísticas.

Se utiliza la información resultante del análisis automático de textos académicos provenientes de dos áreas científicas (Biometría y Filosofía) para conformar una base de datos sobre la cual se aplica las técnicas de Árboles de Clasificación y Regresión Logística. Esta aplicación presenta diferencias respecto al análisis discriminante aplicado en trabajos previos.

Mediante la interpretación de los nodos del árbol y de los coeficientes del modelo logístico estimado se busca hallar las características, considerándolas simultáneamente a todas ellas, provenientes del análisis automático de los textos que son más discriminatorias de las áreas científicas de las cuales provienen.

## 2. MATERIAL Y METODOS

### 2.1. Diseño de la muestra

El marco muestral para la selección de la muestra es el utilizado en trabajos anteriores. Está compuesto por textos académicos, resúmenes de trabajos presentados a congresos y revistas, extraídos de internet pertenecientes a dos 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”.

Las muestras de cada área disciplinar se compararon respecto al número medio de palabras por texto. Esta comparación se requiere para evitar que el tamaño de los textos pueda afectar la discriminación entre las disciplinas, aunque la información incluida en los modelos estadísticos corresponde a las proporciones de cada categoría gramatical y no a la frecuencia de ellas.

Se seleccionaron 60 textos de cada disciplina y cada muestra fue particionada aleatoriamente en dos submuestras. En cada corpus, una submuestra se utiliza para el entrenamiento o estimación de los modelos y la otra para su validación.

### 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.

El módulo post-smorph 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.

La información resultante del análisis morfológico se dispone en una matriz de dimensión: tantas filas como cantidad de objetos lingüísticos tenga el texto (palabras) y tantas columnas como ocurrencia+lema+valores. De esta manera se obtiene una matriz con la estructura que se muestra en la tabla 1.

Tabla 1. Fragmento de la matriz 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 matriz, donde cada fila es una palabra analizada, se confecciona la base de datos por documento que será analizada estadísticamente. Esta nueva matriz, donde cada fila o unidad experimental es el texto, retiene la información de las variables presentadas en la tabla 2 con la estructura definida en la tabla 3.

Tabla 2. Variables de la base de datos por documento

<b>CORPUS</b>	Corpus al que pertenece el texto
<b>TEXTO</b>	Identificador del texto dentro del corpus
<b>Prop_adj</b>	proporción de adjetivos del texto
<b>Prop_adv</b>	proporción de adverbios del texto
<b>Prop_cl</b>	proporción de clíticos del texto
<b>Prop_cop</b>	proporción de copulativos del texto
<b>Prop_det</b>	proporción de determinantes del texto
<b>Prop_nom</b>	proporción de nombres (sustantivos) del texto
<b>Prop_prep</b>	proporción de preposiciones del texto
<b>Prop_v</b>	proporción de verbos del texto

Tabla 3. Fragmento de la base de datos para análisis estadístico

<b>CORPUS</b>	<b>TEXTO</b>	<b>adj</b>	<b>Adv</b>	<b>cl</b>	<b>cop</b>	<b>det</b>	<b>nom</b>	<b>prep</b>	<b>v</b>	<b>OTRO</b>	<b>TOTAL_PAL</b>
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. Metodología Estadística

Uno de los problemas que concentran gran interés en estadística es la clasificación de objetos o unidades en grupos o poblaciones.

Es posible distinguir dos enfoques del problema de clasificación:

- El primero 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, para este caso las técnicas más utilizadas son el Análisis Discriminante y la Regresión Logística.
- 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, dentro de estas técnicas se encuentra el Análisis de Clusters.

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 y en la construcción del modelo no hay ningún supuesto en cuanto a la distribución de probabilidad de las variables por lo que puede incluirse cualquier tipo de variable. Este modelo puede considerarse como una fórmula para calcular la probabilidad de pertenencia a uno de los grupos, de manera que se asigna cada unidad al grupo cuya probabilidad de pertenencia estimada sea mayor.

Los árboles de clasificación son una técnica de análisis discriminante no paramétrica que permite predecir la asignación de unidades u objetos a grupos predefinidos en función de un conjunto de variables predictoras. Esto es, dada una variable respuesta categórica, los AC 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 área disciplinar a la que corresponde: BIOMETRIA-FILOSOFIA 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 recursivas logran formar grupos homogéneos respecto a la variable respuesta (en este caso la disciplina 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. El proceso termina cuando se hayan clasificado todas las observaciones correctamente en su grupo.

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

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

donde  $j = 1, \dots, k$  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

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

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”.

Los AC son más flexibles que otras técnicas de clasificación porque permiten incorporar predictores medidos virtualmente en cualquier escala: continua, ordinal o mezclas de ambas escalas. Por su condición "no paramétrica", constituyen una opción favorable entre otras técnicas alternativas como el análisis discriminante o la regresión logística.

A modo de ejemplo, supóngase una variable respuesta  $Y$  que se pretende discriminar en función de tres predictores  $X_1$ ,  $X_2$  y  $X_3$ . Asumir además que la variable respuesta  $Y$  puede asumir dos valores posibles o categorías (SI/NO), las variables explicativas  $X_1$  y  $X_2$  son cuantitativas continuas y el tercer regresor  $X_3$  es una variable categórica nominal que puede asumir sólo dos valores o categorías posibles (a/b). El árbol de la figura 1 representa los resultados de la aplicación de la técnica de AC a este ejemplo.

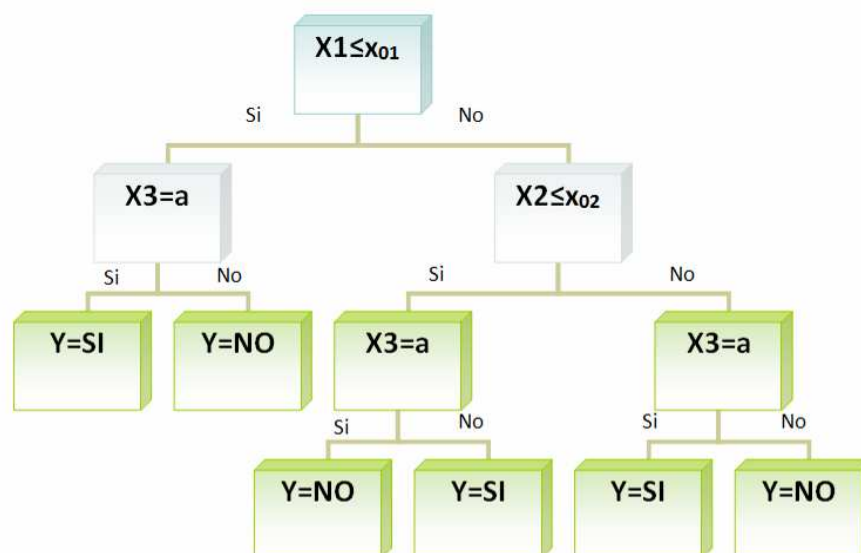


Figura 1: AC

Del árbol de la figura 1, a partir de los seis nodos terminales, se concluye que los regresores  $X_1$ ,  $X_2$  y  $X_3$  discriminan las categorías de la variable respuesta  $Y$  (SI/NO) de la siguiente manera:

- Si  $X_1$  es menor o igual que  $x_{01}$  y  $X_3 = a$  entonces resulta  $Y = SI$
- Si  $X_1$  es menor o igual que  $x_{01}$  y  $X_3 = b$  entonces resulta  $Y = NO$
- Si  $X_1$  es mayor que  $x_{01}$ ,  $X_2$  es menor o igual a  $x_{02}$  y  $X_3 = a$  entonces resulta  $Y = NO$
- Si  $X_1$  es mayor que  $x_{01}$ ,  $X_2$  es menor o igual a  $x_{02}$  y  $X_3 = b$  entonces resulta  $Y = SI$
- Si  $X_1$  es mayor que  $x_{01}$ ,  $X_2$  es mayor a  $x_{02}$  y  $X_3 = a$  entonces resulta  $Y = SI$
- Si  $X_1$  es mayor que  $x_{01}$ ,  $X_2$  es mayor a  $x_{02}$  y  $X_3 = b$  entonces resulta  $Y = NO$

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

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

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

$\beta_0$  es la constante del modelo o término independiente

$p$  el número de covariables

$\beta_i$  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:

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

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

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

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 (en este caso las disciplinas). 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.

### 3. RESULTADOS

#### 3.1. Análisis preliminar.

Previamente se hizo referencia a la comparación de los corpus respecto a la cantidad 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 la proporción de clíticos, adverbios, determinantes, nombres y preposiciones en los documentos analizados (Tabla 4). La proporción de clíticos, nombres y preposiciones es mayor en los textos de biometría y la proporción de adverbios y determinantes es superior en los textos de filosofía.

Tabla 4. Comparación mediante test de Wilcoxon

Variable	BIOMETRIA	FILOSOFIA	Valor de p
prop_adj	0.08968	0.10098	0.13536
prop_adv	0.01763	0.03136	0.00440
prop_cl	0.02603	0.01572	0.00130
prop_cop	0.02788	0.03376	0.10864
prop_det	0.16037	0.17645	0.01246
prop_nom	0.25971	0.24251	0.04514
prop_prep	0.17819	0.16219	0.02760
prop_v	0.12730	0.12493	0.88246

#### 3.2. Árboles de clasificación

Se aplicó la técnica de árboles de clasificación para obtener reglas de clasificación que permitan asignar los textos en estas dos poblaciones, definidas por el área científica a la que pertenecen: BIOMETRIA y FILOSOFIA. 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, clíticos, preposición y adjetivos. El árbol final presenta 7 nodos terminales.

La figura 2 muestra el árbol resultante de esta aplicación. La variable regresora más fuertemente asociada con el área disciplinar 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. Este primer subgrupo es clasificado como FILOSOFIA y constituye un nodo terminal, mientras que el grupo con menos del 2.9% de adverbios es dividido nuevamente. El predictor más relevante en esta segunda división fue la proporción de clíticos en el texto, categorizada en inferior y superior a 0.011 (1.1%). El subgrupo con un porcentaje de clíticos inferior a 1.1% (que ya tenían definido un porcentaje de adverbios inferior a 2.9%) pertenecen al área de BIOMETRIA y constituyen el segundo nodo terminal. Por otro lado, el grupo con un porcentaje de clíticos superior al 1.1% (y con menos del 2.9% de adverbios) vuelve a subdividirse respecto al mismo predictor, definiendo un nuevo nodo terminal para aquellos que presentan un porcentaje de clíticos superior a 2.8%, clasificándolos en el área de BIOMETRIA. Continuando con la subdivisión del otro grupo, la proporción de preposiciones es el regresor determinante. Textos con un porcentaje de preposiciones superior a 20% (y que ya tenían bajo porcentaje de adverbios y un porcentaje de clíticos entre 1.1% y 2.8%) se clasifican en BIOMETRIA y constituye el cuarto nodo terminal. El resto de los textos se vuelven a separar en dos grupos en función de la proporción de adjetivos. Textos con un porcentaje de preposiciones inferior al 20% (y que ya tenían bajo porcentaje de adverbios y un porcentaje de clíticos entre 1.1% y 2.8%) se dividen teniendo en cuenta los adjetivos. Un porcentaje de adjetivos menor a 7,5% pertenecen a FILOSOFIA y constituyen el quinto nodo terminal. Cuando el porcentaje de adjetivos mayor a 7,5% y menor a 11% pertenecen a BIOMETRIA y constituyen el sexto nodo terminal, mientras que si el porcentaje de adjetivos mayor a 11% pertenecen a FILOSOFIA y constituyen el séptimo nodo terminal.

El árbol final fue evaluado utilizando la muestra que no fue utilizada en la construcción del mismo hallando una tasa de mala clasificación del 10%, siendo 17% para biometría y 3% para filosofía (Figura 3).

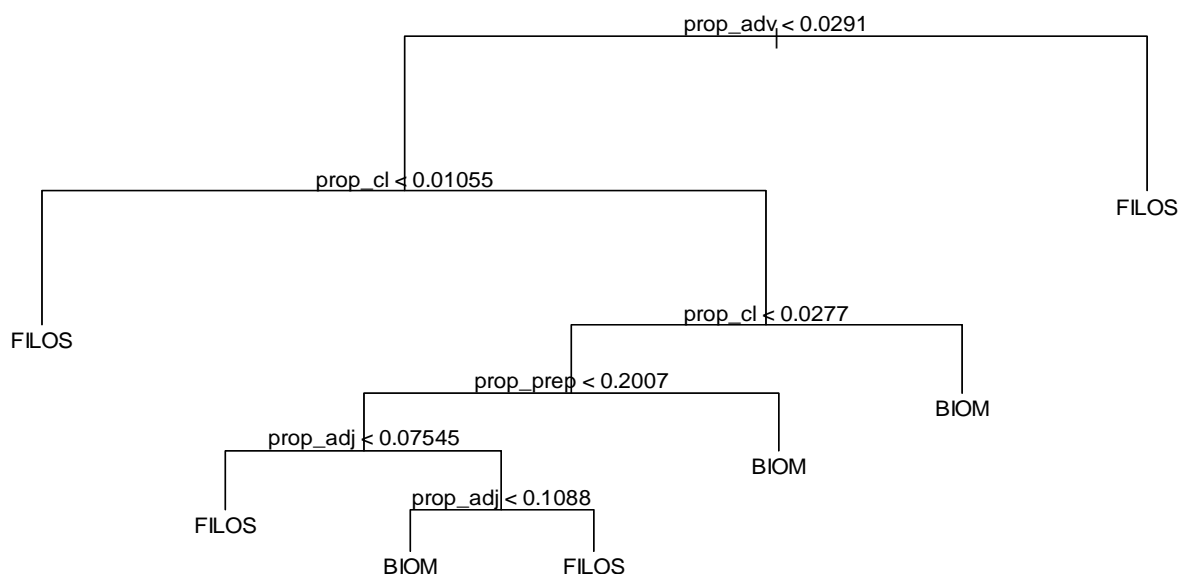


Figura 2: Árbol de clasificación de textos según área disciplinar.

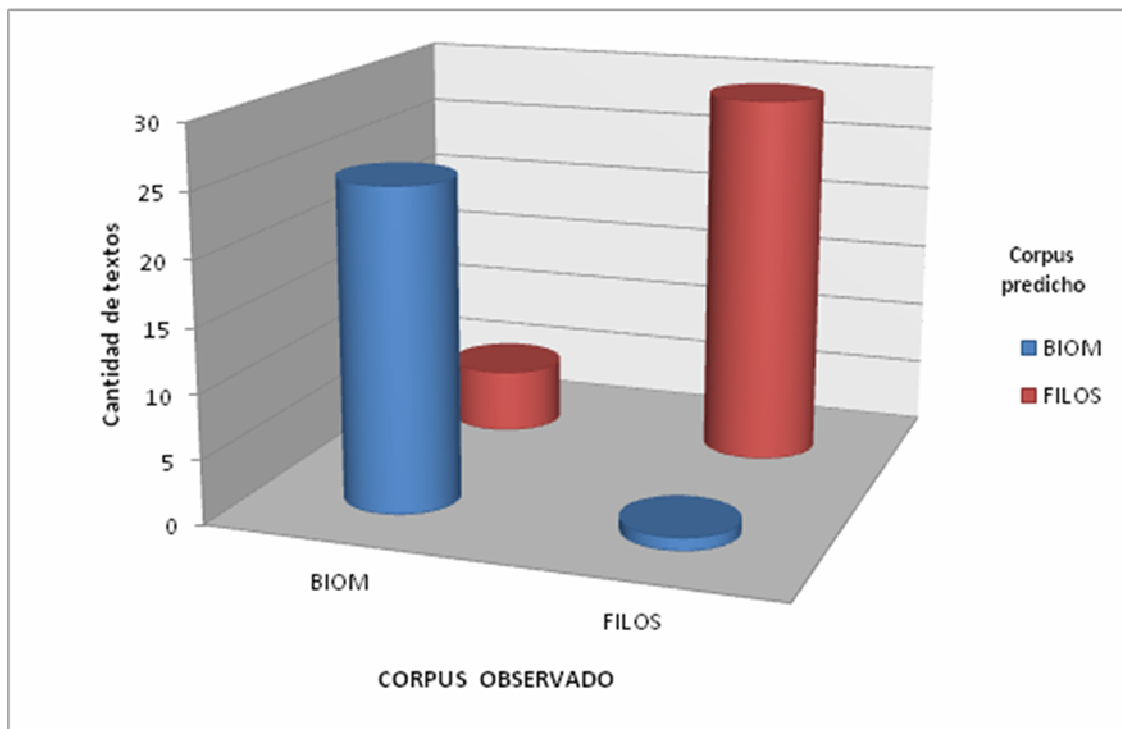


Figura 3: Clasificación de los textos de prueba mediante el árbol final.

### 3.3. 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 área científica a la que pertenecen, 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 la proporción de adverbios y de clíticos.

El modelo final estimado luego de la selección de variables se muestra en la tabla 5.

Tabla 5: Coeficientes del modelo de regresión logística final

Estimación máximo verosímil					
Coeficiente	gl	Estimador	Error estándar	Est. Chi-cuadrado	Prob. asociada
Intercepto	1	-0.0362	0.7293	0.0025	0.9604
Prop_adv	1	-78.7323	27.6806	8.0902	0.0045
Prop_cl	1	90.0310	30.0669	8.9662	0.0028

La bondad del ajuste se evalúa 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 obtiene una tasa de error global del 17% mediante validación cruzada y la probabilidad asociada en el test de bondad de ajuste es  $p=0.2062$  evidenciando lo adecuado del modelo. Respecto a la tasa de mala clasificación, ésta resultó de un 20%, siendo 17% dentro del área de biometría y 23% dentro de filosofía (Figura 4).



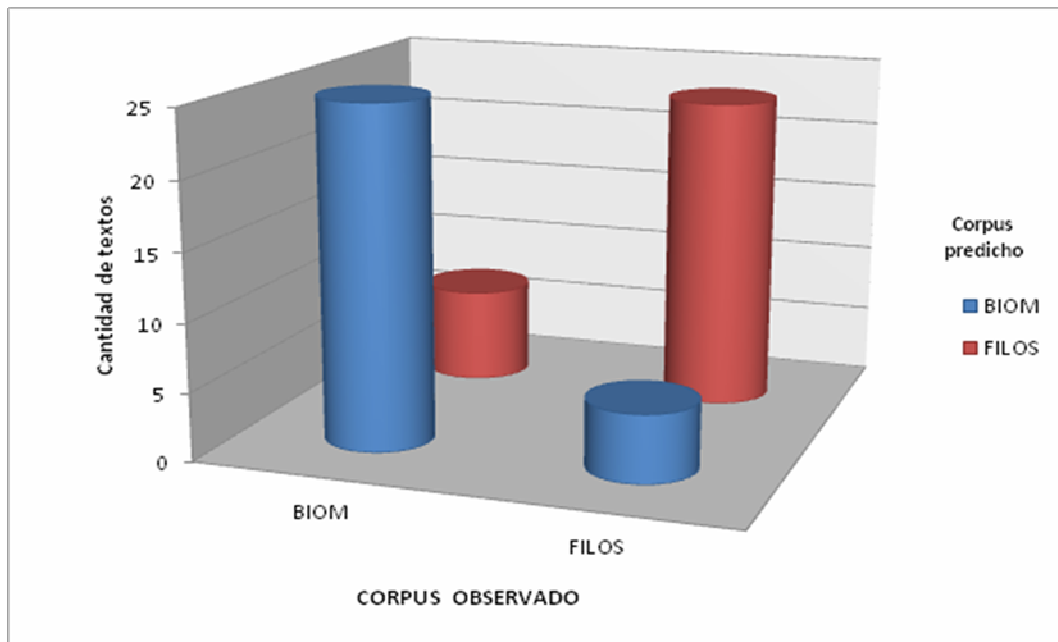


Figura 4: Clasificación de los textos de prueba mediante modelo logístico.

## 4. CONCLUSIONES

Los resultados del análisis morfológico de los textos se analizaron teniendo en cuenta simultáneamente todas las mediciones realizadas sobre ellos sin aplicar ninguna transformación a las variables.

Si bien el número de unidades utilizadas en el entrenamiento y evaluación no era elevado, el árbol de clasificación obtenido mostró un buen desempeño frente al modelo de regresión logística, 10% y 20% respectivamente. La diferencia en la tasa de mala clasificación sólo se diferenció en el área de Filosofía para la cual con el árbol se obtuvo un 3% de mala clasificación versus un 23% para el modelo de regresión logística.

En ambos tipos de análisis, las diferencias entre los dos tipos de textos está centrada principalmente en el porcentaje de clíticos y de adverbios presentes. Sin embargo, en esta nueva aplicación de los árboles de clasificación han intervenido otras variables en la discriminación como el porcentaje de preposiciones y adjetivos. Estas variables intervienen determinando una interacción entre las variables que no se alcanza a observar en el modelo de regresión logística.

Similares resultados se hallaron en Beltrán (2010) utilizando un análisis discriminante sobre las variables transformadas, debido al requerimiento de distribución Normal de las mismas.

Esta particularidad de los textos analizados de estas disciplinas puede deberse a que, en los textos de biometría/estadística hay más clíticos que en los humanísticos por la frecuencia de expresiones impersonales o pasivas con el clítico “se” del tipo:

“se ajusta un modelo cuadrático”

“se estima la variancia poblacional”

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

Respecto a la metodología estadística planteada, se puede afirmar que entre las ventajas de los árboles de clasificación está la adaptación para recoger el comportamiento no aditivo de las variables predictoras, de manera que las interacciones se incluyen de manera automática. Por el

contrario, entre las desventajas se evidencia que las variables predictoras continuas se tratan como variables dicotómicas perdiendo información.

Esta metodología puede ser generalizada a un número mayor de disciplinas. Asimismo, se continuará trabajando en la comparación de las técnicas estadísticas mediante simulación.

## 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. 2010 *Análisis discriminante aplicado a textos académicos: Biometría y Filosofía*. Revista INFOSUR. Grupo INFOSUR. 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.
- Hosmer, D.W.; Lemeshow, S. 1989 *Applied Logistic Regression*. John Wiley & Sons. New York.
- Johnson R.A. y Wichern D.W. 1992 *Applied Multivariate Statistical Analysis*. Prentice-Hall International Inc.
- Khattre R. y Naik D. 1999 *Applied Multivariate Statistics with SAS Software*. SAS. Institute Inc. Cary, NC. USA.
- Khattre R. y Naik D. 2000 *Multivariate Data Reduction and Discriminatio with SAS Software*. SAS Institute Inc. Cary, NC. USA
- Maindonald, J.; Braun, J. 2004. *Data Analysis and Graphics Using R.- an example-based approach*. Cambridge University Press.
- 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.
- Verzani, J. 2005 *Using R for Introductory Statistics*. CHAPMAN & HALL/CRC. Boca Raton London New York Washington, D.C.

# **Metodologías para la creación colaborativa de libros de texto**

## **Methodologies for the collaborative creation of textbooks**

Claudia Deco<sup>1</sup>, Cristina Bender<sup>1</sup>, Ana Casali<sup>1,2</sup>, Raúl Kantor<sup>1,3</sup>

<sup>1</sup> Facultad de Ciencias Exactas, Ingeniería y Agrimensura  
Universidad Nacional de Rosario

{deco,bender,acasali,kantor}@fceia.unr.edu.ar

<sup>2</sup> Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas - CIFASIS,  
Rosario, Argentina.

<sup>3</sup> Facultad de Arquitectura, Planeamiento y Diseño  
Universidad Nacional de Rosario

### **Abstract**

A problem faced by higher education students is the high cost of textbooks. One reason for this cost is that a large percentage is created by non-Latin American authors and editors. An additional problem is that most academic books are not adapted to the cultural context of Latin America. One solution to improve access to textbooks is to promote the creation of open textbooks produced by academic authors in the region. That is, textbooks that might be freely copied, printed, modified, adapted and distributed to students. This paper discusses some methodologies for the collaborative creation of open textbooks, and the requirements for a technological platform for collaborative work. This work is part of the LATIn Initiative (Latin American Open Textbook Initiative), which is under development, and its main objectives are the proposal of a support architecture, the design of a methodology for the collaborative creation of free open-access textbooks, and the definition of strategies for their implementation, adoption and dissemination throughout the region.

**KeyWords:** Open Textbooks, Technological platforms, Collaborative methodologies.

### **Resumen**

Un problema que enfrentan los estudiantes de educación superior para llevar adelante sus estudios, es el alto costo de los libros de texto. Un motivo de este costo es que en un gran porcentaje son creados fuera de la región por autores y editores no latinoamericanos. Un problema adicional es que la mayoría de los libros académicos no está adaptada al contexto de la educación superior en América Latina. Una solución para mejorar el acceso a los libros de texto de este nivel educativo es promover la creación de libros de texto abiertos académicos generados por autores de la región. Esto es, libros de texto que pueden ser libremente copiados, impresos, modificados, adaptados y distribuidos a los estudiantes. En este trabajo se analizan metodologías para la creación colaborativa de libros de texto abiertos, así como los requerimientos a cumplir por la plataforma tecnológica para este trabajo colaborativo. Este trabajo se enmarca dentro del Proyecto LATIn Initiative (Latin American open Textbook Initiative) que se encuentra en desarrollo y sus principales objetivos son la creación de una arquitectura de soporte, el diseño de una metodología para la creación

colaborativa de libros de texto gratuitos y de acceso abierto, y la definición de las estrategias para la implementación, adopción y disseminación de libros de textos abiertos para la educación superior de toda la región.

**Palabras claves:** Libros de texto abiertos, Metodologías para trabajo colaborativo, Plataformas tecnológicas.

## 1. Introducción

La importancia de la colaboración en la escritura de libros de texto electrónicos en el contexto de recursos abiertos para la educación ha ido creciendo en todo el mundo. En este trabajo se presenta un análisis del estado del arte de metodologías para la creación colaborativa de libros, con especial atención a los libros de texto académico abiertos.

Un problema que enfrentan los estudiantes de América Latina para llevar adelante sus estudios, es el alto costo de los libros de texto. Este problema se potencia en familias de ingresos bajos y medios. Las bibliotecas solucionan parcialmente este problema al proveer ejemplares de los libros de texto de forma gratuita a los estudiantes. Sin embargo, estas bibliotecas no tienen presupuesto para satisfacer en su totalidad la demanda. Inclusive en el caso de que el alumno emplee fotocopias de los libros que, además de no ser legales, en muchos casos degradan la calidad de imágenes, gráficos, etc. Por otro lado, los apuntes de clase, pueden ser acotados o incompletos en el desarrollo de los temas. De esta forma se crea una diferencia de posibilidades entre los estudiantes que pueden tener el libro de texto y los que no pueden (Ochoa et al., 2011).

En los últimos años, varios gobiernos de América Latina han atacado el problema del costo de los libros de texto ofreciendo alternativas creadas a nivel local. Estos libros producidos por los respectivos gobiernos se han entregado a los estudiantes con excelentes resultados. Esta estrategia, sin embargo, se ha limitado a la educación primaria y secundaria, donde el plan de estudios es administrado en forma centralizada. La diversidad, especialización y libertad académica hacen muy difícil poner en práctica un plan similar para libros de texto en la educación superior. En particular, en la educación superior, uno de los motivos de los altos costos de los libros de texto es que en un gran porcentaje son creados fuera de la región por autores no latinoamericanos. Esto no está relacionado, en general, con la falta de capacidad de producción, sino con la dificultad que los profesores o autores locales tienen para publicar y distribuir sus libros. El origen externo de estos libros de texto tiene consecuencias adicionales, ya que la mayoría no está adaptada al contexto de la educación superior en América Latina, y, en muchos casos, las versiones más recientes no están disponibles en un idioma local. Además, se fomenta entre los estudiantes la percepción, muchas veces errónea, de que el conocimiento sólo puede tener origen fuera de la región.

Una solución alternativa para mejorar el acceso a los libros de texto de este nivel, preservando la libertad académica de cada profesor, ha ido surgiendo en el ámbito de las Tecnologías de Aprendizaje, a partir del concepto de recurso educativo abierto. Estos son materiales educativos que pueden ser libremente copiados, modificados, compartidos, impresos y distribuidos, y han llevado a la creación del concepto de libros de texto abiertos.

Los libros electrónicos pueden ser considerados como los libros de texto tradicionales pero disponibles en formato electrónico. Se plantea tener una visión más amplia, como la que presenta Faquet et al. en (2001) que los define como documentos virtuales compuestos de fragmentos que pueden ser ensamblados para constituir documentos reales que pueden ser leídos o impresos, o hipertextos que pueden leerse mediante la navegación. Estos libros además, pueden tener contenido multimedia (vídeos, audio, animaciones, contenido y aplicaciones interactivas 3D) y la posibilidad de ser actualizados con la frecuencia necesaria. De este modo, además de la reducción del costo de los libros de texto, se puede mejorar tanto el formato de entrega como la calidad del contenido.

Hay varias propuestas relacionadas con el concepto de libro de texto abierto como por ejemplo: Wikilibros ([www.wikibooks.org/](http://www.wikibooks.org/)), Connexions ([cnx.org/](http://cnx.org/)), y el Proyecto California Open Source Textbooks ([www.opensourcetext.org/](http://www.opensourcetext.org/)), entre otras. Estas iniciativas han logrado reducir significativamente el costo de los libros de texto para los estudiantes. Algunos de estos libros incluyen pequeñas secciones de materiales en español o portugués, pero la participación de autores latinoamericanos es bastante limitada. En la Iniciativa LATIn (Latin American open Textbook Initiative)<sup>1</sup> se propone como principal objetivo la creación y diseminación de libros de texto abiertos y colaborativos para la educación superior para América Latina. En el marco de esta iniciativa, en este artículo se analizan algunas propuestas de metodologías colaborativas para la creación de libros de texto abiertos y se presentan algunos aspectos referidos a las necesidades que deberían contemplar las plataformas tecnológicas para ayudar a la implementación de estas metodologías.

## 2. Metodologías para la escritura colaborativa de libros

Los avances en las Tecnologías de la Información y Comunicaciones (TICs) que facilitan la comunicación y el intercambio de información en formatos digitales parecen tener un impacto positivo en la tendencia a trabajar en forma colaborativa. Además, a partir de las iniciativas de software libre y la Web 2.0, hoy en día las personas se reúnen en espacios digitales y utilizan herramientas y medios digitales para crear cosas que se puedan compartir entre ellos y con los demás.

Existen muchas definiciones para *colaboración*. Patel, Pettitt y Wilson (2012) describen la colaboración como una actividad comunitaria en la que las personas se coordinan para comunicarse y lograr objetivos comunes. El Oxford Dictionaries Online la define como la acción de trabajar con alguien para producir algo (Oxford University Press, 2012), y Wikipedia la define como “trabajar juntos para lograr un objetivo” (Wikipedia, 2012). Podemos ver que *colaboración* tiene que ver con personas que trabajan juntas (es decir, coordinan ellos mismos para comunicarse e interactuar) con el fin de lograr objetivos comunes, especialmente cuando las metas implican la producción de algo, y más específicamente cuando ese algo es un libro. Esta definición distingue la colaboración de fenómenos estrechamente relacionados, como cuando las personas trabajan por un objetivo común, pero sin ningún tipo de coordinación, la comunicación o la interacción, a través de contribuciones individuales y aisladas. También distingue colaboración de los casos en que las partes trabajen juntas hacia metas independientes.

Respecto a la escritura de libros, este proceso no es simple. Es responsabilidad del autor definir el tema, llevar a cabo investigaciones sobre los temas, organizar las ideas y puntos de vista que se

<sup>1</sup> Proyecto Alfa III: DCI-ALA/19.09.01/11/21526/279-155/ALFA III(2011)-52

manejan, decidir cómo estructurar el texto, entre otros pasos. Este proceso se vuelve aún más complejo al tratar de realizarlo en forma colaborativa, ya que es necesario atender dos frentes: las dificultades comunes a cualquier proceso que implique la colaboración y los pasos para crear un texto. Con un grupo de personas dedicadas a esta tarea, existe el reto de escribir un texto de alta calidad con estilos de escritura diferentes, ideas, vocabulario e incluso cultura. Es esencial que el objetivo del texto esté bien definido y exista un cierto nivel de compromiso con él, roles identificados entre los miembros del grupo, una fuerte división de tareas, discusiones sobre el contenido y la definición del crédito correspondiente a cada uno de los autores.

Posner y Baeker (1993) crearon una taxonomía para analizar el proceso de escritura colaborativa. Esta taxonomía se divide en cuatro categorías: funciones, actividades, métodos de control de documentos y estrategias de escritura. Así, los diferentes roles que la gente puede tomar son: *Escritor*, que es el responsable de transformar ideas abstractas en un texto coherente y organizado; *Consultor*, que es quien trabaja en estrecha colaboración con los escritores, pero no toma parte en la redacción del texto; *Editor*, es quien hace cambios a los documentos que fueron escritos por otra persona; y *Revisor*, que es quien ofrece comentarios sobre el documento, los que pueden ser aceptados o ignorados por el escritor.

Adkins et al. (1999) añaden dos funciones adicionales a esta taxonomía: *Lider facilitador*, quien estructura y controla el proyecto; y *Editor de textos*, quien es el encargado de pulir el borrador final para su publicación. Estos autores mencionan que una persona puede tomar varios de estos roles durante la ejecución del proceso de escritura. Mencionan también que otras actividades que normalmente tienen lugar durante el proceso de escritura son: lluvia de ideas, planificación, investigación, redacción, edición, revisión. El proceso de escritura no incluye necesariamente todas las actividades y el orden en el que se ejecutan no es secuencial y depende de la organización de cada grupo.

También hay diferentes maneras en que los miembros del grupo colaboran juntos en el proceso de escritura. Pueden nombrarse cuatro tipos de estrategias: un solo escritor, escritores independientes (donde cada persona trabaja en una parte diferente de la obra), redacción conjunta (donde los autores trabajan juntos en estrecha colaboración sincrónica en el texto), y un escritor que con base en las discusiones del grupo, escribe el documento. Los patrones y la taxonomía permiten comparar fácilmente las metodologías para la escritura colaborativa utilizada por los diferentes grupos.

Los principales objetivos a tener en cuenta para una metodología que guíe el proceso de creación colaborativa de contenidos son garantizar la calidad de los materiales, facilitar el proceso de colaboración y promover la reutilización de los componentes individuales.

En este sentido, en la iniciativa LATIn se intenta determinar el modo de colaboración en la autoría de los libros, las herramientas a utilizarse y el tipo de licencia para los autores y para los profesores que reutilicen este material para editarlo. Se establece la forma en que estos libros pueden ser personalizados permitiendo la selección de módulos, compaginado y reeditado de sus partes. Esta metodología debe tener en cuenta las características culturales de los diferentes países involucrados y en particular, cómo se realizarán las traducciones y adaptaciones necesarias. Se planean traducciones del material al portugués y al español hablado en los distintos países de Latinoamérica.

Así, el proceso de creación de un libro debería tener en consideración las siguientes fases: construir un equipo de trabajo, planificar las tareas y cómo distribuirlas en el equipo, diseñar el libro (producto a obtener), realizar las tareas de escritura, las revisiones, la edición final de la obra, y finalmente la publicación del libro.

Entre los aspectos a considerar es importante tener en cuenta que la escritura colaborativa todavía es un proceso que es re-descubierto por cada grupo involucrado en esta tarea. Una buena interacción y

una comunicación fluida entre participantes es uno de los factores principales para su éxito. También los diferentes roles y responsabilidades deben estar claramente definidos. Además, es importante tener en vista aspectos de normalización tal como usar una guía de estilo.

No hay una única metodología para todos sino que depende de cada caso y situación. El abanico de estrategias varía de acuerdo a las necesidades y al contexto de cada grupo.

### 3. Plataformas tecnológicas

El uso de computadoras y redes informáticas es uno de los sistemas de soporte más reciente y estudiado para el trabajo colaborativo y las actividades de coordinación, y se conoce como *Computer-Supported Cooperative Work* (CSCW). Esto es, una actividad coordinada asistida por computadora y llevada a cabo por grupos de personas (Baecker, 1995). En este sentido, una aplicación *Groupware* es un tipo de aplicación que está destinada a apoyar el trabajo colaborativo. Los *Groupware* son sistemas basados en computadoras que apoyan a grupos de personas que trabajan en una tarea común, y que proporcionan una interfaz para un entorno compartido. Sosa, Zarco y Postiglioni (2006) dicen que es software y hardware que apoya y ayuda a que el trabajo se realice en grupo. Hoy en día, existen numerosas posibilidades proporcionadas por herramientas basadas en web para editar, publicar y compartir contenido. Así, en lugar de planear un producto hipermedial como una amalgama de contenido estático previsto, nuevas piezas de información pueden ser constantemente añadidas, modificadas, actualizadas, descartadas, sustituidas o incluso reubicadas.

Para la escritura colaborativa de libros de texto abiertos, se necesitan tipos especiales de plataformas tecnológicas. Hay que trabajar en el diseño de una plataforma tecnológica basada en la Web que brinde las funcionalidades necesarias para soportar la metodología para la creación colaborativa de los libros. Estas deben proporcionar las funcionalidades necesarias para apoyar la metodología para la creación colaborativa de secciones y capítulos de libros; proporcionar herramientas para mezclar estas secciones y capítulos en libros de texto personalizados para ser utilizados en un curso específico; proporcionar herramientas para que los usuarios puedan leer los libros en línea, descargarlos en algún formato electrónico como PDF, o imprimirlos para la lectura fuera de línea. También debería facilitar la creación de nuevas versiones, adaptaciones de versiones existentes, o traducciones a otros idiomas (por ejemplo, portugués o español de distintos países), así como proveer herramientas para la edición de nuevos módulos, de forma de poder reutilizar los libros, y proporcionar herramientas de recomendación para la creación de nuevas comunidades y nuevos libros de texto relevantes.

Se realizó un análisis de algunas plataformas existentes, para su posible adopción y adaptación a los objetivos planteados. Se analizaron, entre otros aspectos, si son plataformas abiertas, si permiten la colaboración y la modularización de contenidos, y cómo tratan el control de calidad y la autoría. Entre las plataformas que se analizaron pueden nombrarse: Wikibooks<sup>2</sup>, Connexions<sup>3</sup>, Flat World Knowledge<sup>4</sup>, The Global Text Project<sup>5</sup> y Textbook Media<sup>6</sup>. A continuación se presentan las características más relevantes de estas plataformas.

---

<sup>2</sup> [www.wikibooks.org](http://www.wikibooks.org)

<sup>3</sup> [cnx.org](http://cnx.org)

<sup>4</sup> [www.flatworldknowledge.com](http://www.flatworldknowledge.com)

<sup>5</sup> [globaltext.terry.uga.edu/](http://globaltext.terry.uga.edu/)

Wikibooks (Wikibooks, 2012) es una plataforma para el desarrollo de libros completamente colaborativos y usa la misma interfase que Wikipedia. La plataforma tiene limitaciones para atribuir autoría y para controlar la calidad de los textos, pero es una valiosa herramienta para una clase constructivista. Los Wikibooks son publicados bajo licencia libre GNU Free Documentation License (GDFL).

Connexions (Dholakia et al. 2006; Baker et al. 2009) es una plataforma y repositorio dedicado exclusivamente al desarrollo de libros abiertos. Usa un modelo de contribución abierta en el cual cualquier usuario registrado puede aportar contenidos, los cuales están modularizados por lo cual permite aumentar la flexibilidad de los libros. El formato modular hace necesaria la compilación de los contenidos por parte de los profesores, pero les brinda una gran flexibilidad. Los libros son publicados bajo licencia Creative Commons.

Flat World Knowledge (FWK) (Shelstad 2008; Hilton and Wiley 2011) presenta una alternativa de negocio para la generación de libros abiertos. Esta plataforma no tiene la flexibilidad de Connexions pero hace más fácil el trabajo al profesor porque el libro se genera en su forma completa y hay control editorial. FWK provee versiones electrónicas completamente libres y la ganancia la obtienen de la impresión a demanda de las versiones de los libros a un precio razonable.

The Global Text Project<sup>5</sup> (de la fundación Jacobs Foundation) fue desarrollado por las Universidades de Georgia y de Denver con el propósito de proveer 1000 libros abiertos a los alumnos de países en vías de desarrollo, pero pueden ser accedidos por cualquier persona vía Web. Actualmente los libros están disponibles en formato PDF y no tienen la opción de impresión a demanda. Al igual que Connexions y FWK, en este proyecto los libros son publicados bajo licencia Creative Commons.

Textbook Media<sup>6</sup> implementa un modelo de negocios diferente al de FWK, usando propaganda para solventar las versiones libres de sus libros. Las versiones sin propaganda están disponibles a un precio razonable. Todos los libros son provistos en un formato final y tienen control de calidad editorial. Este modelo es interesante para solventar libros libres pero la interfase de esta versión es difícil de usar y todo el trabajo es publicado bajo el tradicional copyright editorial.

Como consideración global de las plataformas para que puedan cumplir adecuadamente el objetivo de ser un soporte a la escritura colaborativa, éstas deberían estar basadas en Web, soportar el trabajo colaborativo, modularidad, reusabilidad de módulos y libros, facilitar la creación de nuevas versiones, poseer manejo de autoría módulos/libros, poseer herramientas para uso de los libros en línea, permitir la descarga (por ejemplo en formato PDF) e impresión, ser fácil de usar, proveer soporte para cada paso del proceso, permitir el manejo de formatos multimedia, tener herramientas de comunicación embebida, distribución y personalización, y proveer herramientas para trayectorias de autoría.

## 4. Conclusiones

En este trabajo se presentaron diversas metodologías dedicadas a la escritura colaborativa de libros de texto abiertos, que es el objetivo de la iniciativa LATIn en la cual los autores están involucrados. Existen algunas directrices (Ede y Lunsford, 1990; Posner y Baecker, 1993), las que, aunque su carácter es general, son útiles para analizar los procesos de colaboración que guían el desarrollo de

---

<sup>6</sup> [www.textbookmedia.com/](http://www.textbookmedia.com/)



una estrategia metodológica. Las nuevas formas de comunicarse y relacionarse, a partir del advenimiento de las tecnologías Web 2.0, vuelve a abrir el debate sobre cómo organizar con éxito grupos de escritura colaborativa. Las principales conclusiones que pueden extraerse de este trabajo es que para poner en práctica una metodología para la creación colaborativa de libros de texto abiertos académicos, el contexto y la composición de cada grupo deben ser tenidos en cuenta. El abanico de estrategias varía de acuerdo a las necesidades y al contexto de cada una de las iniciativas. Sin embargo, el papel de una comunicación fluida entre los participantes parece ser factor fundamental. No hay ningún tipo de metodología que se aplique a todo caso y situación. Cualquier posible propuesta de metodología para libros de texto de cualquier iniciativa de colaboración abierta debe ser adaptable a las diferentes actividades de cada grupo y debe incorporar los últimos tipos de colaboración derivados de nuevas tecnologías de Internet. La metodología propuesta para la creación de libros abiertos que guíe el proceso de trabajo colaborativo entre profesores deberá garantizar la calidad de los materiales, facilitar el proceso de colaboración y asegurar la usabilidad de los componentes individuales.

## Referencias

- Adkins, M., Reinig, J. Q., Kruse, J., & Mittleman, D. (1999). "GSS collaboration in document development: Using GroupWriter to improve the process." Thirty-Second Annual Hawaii International Conference on System Sciences. p 11-21.
- Baecker, R. M., Grudin, J., Buxton, W. A. S., Greenberg, S. (1995) "Readings in Human-Computer Interaction: Towards the Year 2000" (Second Edition) Morgan Kaufmann Publishers, Inc.
- Baker J., Thierstein J., Fletcher K., Faur M., Emmons J. (2009) OpenTextbook Proof of Concept via Connexions. The International Review of Research in Open and Distance Learning, Volume 10, Number 5. ISSN: 1492-3831.
- Dholakia U., King W. J. , Baraniuk R. (2006) What makes an open education program sustainable? The case of Connexions. Open Education 2006: Community, Culture, and Content, COSL, Utah State University, Logan UT.
- Ede L. and Lunsford A. (1992). Singular Text/Plural Authors: Perspectives on Collaborative Writing. Southern Illinois University Press (January 14, 1992)
- Falquet G., Hurni J., Guyot F. and Nerima L. (2001) Learning by creating multipoint of view scientific hyperbooks, Proc. of European Perspectives on Computer-Supported Collaborative Learning (CSCL 2001), Maastricht, March 22-24, pp. 222-9.
- Henderson S., Nelson D., (2011) The Promise of Open Access Textbooks. A Model for Success. Florida Distance Learning Consortium. Tallahassee, Florida. pp. 1-82. Disponible en [www.openaccesstextbooks.org/pdf/ModelDraft.pdf](http://www.openaccesstextbooks.org/pdf/ModelDraft.pdf).
- Hilton J., Wiley D. (2011) Open-Access Textbooks and Financial Sustainability: A case study on Flat World Knowledge. The International Review of Research in Open and Distance Learning, Vol. 12 Nro. 5. North America. En [www.irrodl.org/index.php/irrodl/article/view/960/1860](http://www.irrodl.org/index.php/irrodl/article/view/960/1860).
- Ochoa, X., Silva Sprock A., Frango Silveira, I. (2011) Collaborative Open Textbooks for Latin America – the LATIn Project. Information Society (i-Society), International Conference on, pp. 398-403.

- Oxford University Press. (2012). Oxford Dictionaries Online. <http://oxforddictionaries.com/>.
- Patel, H., Pettitt, M., and Wilson, J. R. (2012). "Factors of collaborative working: A framework for a collaboration model", *Applied Ergonomics*, 43 (1), pp. 1–26.  
doi:10.1016/j.apergo.2011.04.009.
- Posner, I. & Baecker, R. (1993). "How people write together". *Proceedings of the International Conference on System Sciences*, 25, 127-137.
- Shelstad J. (2008) *Flat World Knowledge: Creating a Global Revolution in College Textbooks!*
- Sosa M., Zarco R., Postiglioni A. (2006) Modelando aspectos de grupo en entornos colaborativos para proyectos de investigación. *Revista de Informática Educativa y Medios Audiovisuales* Vol. 3(7), págs. 22-31.
- Textbooks\Media. [www.textbookmedia.com/HowItWorks.aspx](http://www.textbookmedia.com/HowItWorks.aspx)
- The Global Text Project. Jacobs Foundation. [globaltext.terry.uga.edu/](http://globaltext.terry.uga.edu/)
- Wikibooks. Disponible en [es.wikibooks.org/wiki/Wikilibros:Libros\\_nuevos](http://es.wikibooks.org/wiki/Wikilibros:Libros_nuevos)
- Wikipedia Contributors. (2012). Collaboration. Wikipedia, The Free Encyclopedia. Retrieved from [en.wikipedia.org/w/index.php?title=Collaboration](http://en.wikipedia.org/w/index.php?title=Collaboration)

# **Extracción de Candidatos a Términos del Dominio Médico a Partir de la categorización automática de Palabras<sup>1</sup>**

## **Extraction of Medical Domain Candidate Terms through the Automatic Word Categorization**

**Walter Koza**

ILCL – Pontificia Universidad Católica de Valparaíso  
Viña del Mar, Chile  
Grupo INFOSUR – UNR  
Walter.koza@ucv.cl

### **Abstract**

A set of experiments is presented with the objective of developing a method for the automatic extraction of candidate terms for the medical domain. One of the main problems is the constant change of medical terminology, which does not allow for a fast manual updating of terminologies. So, it is hypothesized that, given that words in medical texts but not in the dictionary software (UW) are mostly medical term candidates, their adequate automatic categorization could facilitate extraction tasks. Firstly, we attempt to deduce which grammatical categories the UW belong to, through word formation rules (1<sup>st</sup>-level analysis) and syntactic rules (2<sup>nd</sup>-level analysis). Secondly, noun phrases (NPs) with UW are formed and these NPs are extracted as domain term candidates. Finally, an evaluation of accuracy is conducted and, through the assessment by medical domain experts, candidates extracted are evaluated as potential term candidates. Smorph [1] and MPS [2] are software used in the computational work. Smorph conducts the morphological analysis and MPS works on local grammars.

**Keywords:** Medical Terminology, Automatic Extraction, Term Candidates, Smorph, MPS

### **Resumen**

Se presenta una serie de experimentos con el objetivo de desarrollar un método automático de extracción terminológica del dominio de la medicina. Uno de los inconvenientes típicos es el cambio constante de la terminología médica, que imposibilita mantener terminologías actualizadas inmediatamente por medios manuales. Se parte de la hipótesis de que las palabras incluidas en los textos médicos que no aparecen en el diccionario fuente del software analizador, denominadas PD, son, en su mayoría, expresiones específicas del dominio médico, por lo cual, una categorización automática de estas ayudaría en la extracción. Primeramente, se intenta deducir a qué categoría pertenecen las PD mediante reglas de formación de palabras (1er. Nivel de análisis) y sintácticas

---

<sup>1</sup> Este trabajo fue presentado y publicado en el Congreso Argentino de Informática y Salud 2012.

(2do. Nivel de análisis). Luego, se procede a la conformación de sintagmas nominales que involucren PD, para extraerlos como candidatos a términos del dominio. Finalmente, se evalúa la precisión de las categorizaciones y, posteriormente, con el asesoramiento de profesionales del área de medicina, se verifica la posibilidad que tienen los candidatos a términos extraídos de ser promovidos a términos. En trabajo computacional, se utilizan las herramientas Smorph [1] y Módulo Post-Smorph (MPS) [2]. Smorph realiza el análisis morfológico y MPS trabaja sobre gramáticas locales.

**Palabras claves:** Terminología Médica, Extracción Automática, Candidatos a Término, Smorph, MPS.

## 1. INTRODUCCION

Se presenta una serie de experimentos realizados con el objetivo de deducir la categoría gramatical de aquellas palabras que no se encuentran en el diccionario fuente de los softwares de análisis automático de textos. Este trabajo se enmarca en el ámbito la lingüística informática, por un lado, y, por otro, en las tareas de minería textual, y forma parte de una serie de tareas tendientes a desarrollar un método de extracción terminológica del dominio de la medicina. A tales efectos, se tomaron en consideración dos tipos de antecedentes: los estudios sobre terminología y extracción de términos, y los de utilización de formalismos y softwares declarativos, en los que la máquina algorítmica está dissociada de los datos a utilizar.

Las tareas de extracción de términos poseen un lugar destacado en actividades de extracción y organización del conocimiento. Un término es una unidad léxica caracterizada por una referencia especial dentro de una disciplina [3], y puede estar conformado por una sola palabra (unigrama), por ejemplo ‘asma’, ‘hormona’; o una combinación de ellas (n-gramas), como ser ‘tuberculosis pulmonar’, ‘sistema cardiovascular’ (bigramas); ‘esquema de tratamiento’, ‘estado de enfermedad’ (trigramas), etcétera. Un conjunto de términos constituye la terminología.

Las tareas de extracción de términos suelen enfocarse en dominios específicos, y uno de ellos es el de la medicina. En este caso, la extracción de términos representativos suele destinarse a la elaboración de listas de entradas para diccionarios electrónicos específicos, la creación de base de datos o de ontologías y taxonomías que organizan y especifican el dominio de conocimiento, etcétera. Otra de las aplicaciones apunta a la clasificación textual, es decir que, a partir de la extracción realizada sobre varios textos o informes de medicina, los especialistas (médicos, enfermeros, técnicos del área e incluso pacientes) puedan reconocer a qué subáreas pertenecen dichos textos.

Uno de los inconvenientes en la identificación automática de términos en este ámbito es el cambio constante de la terminología médica. Esto dificulta de sobremanera mantener terminologías actualizadas inmediatamente por medios manuales, por lo que se hace necesario contar con herramientas que puedan considerar en el análisis aquellos términos nuevos.

Parto de la hipótesis de que las palabras incluidas en los textos médicos que no aparecen en el diccionario fuente del software analizador (se trata de un diccionario estándar), y que son

etiquetadas como palabras desconocidas (PD), son, en su mayoría, expresiones específicas del dominio médico, por lo cual, una categorización automática de estas ayudaría a la extracción terminológica. A partir de allí, se elaboran reglas para deducir las PD y las que son categorizadas como nombres, pasan a formar parte de la lista de términos del dominio médico establecida automáticamente. En este caso la lista se divide en unigramas ('osteoporosis'), bigramas ('insuficiencia ovárica') y trigramas ('síndrome de Cushing').

El artículo se organiza de la siguiente manera. En primer lugar, se presentan los antecedentes en esta área. En segundo lugar, la metodología. Finalmente, en tercer lugar, la descripción de la tareas a llevar a cabo.

## 2. ACERCA DE LA TERMINOLOGÍA MÉDICA Y LAS TAREAS DE EXTRACCIÓN AUTOMÁTICA

Se denomina “términos” a los elementos léxicos utilizados en un ámbito temáticamente restringido para denominar un concepto [4], con lo cual, una correcta identificación de estos es clave para acceder al conocimiento que se transmite en los textos. Generalmente, son los sintagmas nominales los que se corresponden con los términos [5], a tales efectos, en el presente proyecto, la extracción estará focalizada en ellos.

En lo que atañe al área de la medicina, se observa que la terminología es extremadamente cambiante y compleja, por lo que la identificación de términos en este dominio se ha convertido en uno de los principales tópicos de investigación, tanto en el procesamiento del lenguaje natural, como así también en las comunidades biomédicas [6]. Términos tales como nombres de genes, proteínas, organismos, drogas, componentes químicos, etcétera, son los medios que se utilizan en medicina para identificar conceptos del dominio.

Existen diversos trabajos en la literatura que realizan la extracción de términos [7,8,9,10]. En relación con el dominio de medicina, de acuerdo con Castro [11], para el caso del inglés, hay varias investigaciones orientadas al procesamiento de textos y de datos de ese dominio [12,13], sin embargo, se encuentran pocas iniciativas para el español [14,15,16,17,18].

El principal inconveniente para una identificación automática exitosa es que las palabras y los términos comparten la misma estructura superficial [4], otros problemas se derivan de las variaciones léxicas, los casos de sinonimia (cuando un concepto está representado por varios términos) o de homonimia (cuando un término tiene varios significados). Por otro lado, cabe mencionar el cambio constante de la terminología médica; algunos términos aparecen en un período muy corto y, a la vez, se crean nuevos casi a diario. Esto hace imposible mantener terminologías actualizadas inmediatamente por medios manuales. Otro de los problemas, señalado por Krauthammer y Nenadić [6], remite a la falta de convenciones firmes en la nomenclatura. Si bien existen algunas directrices para ciertos tipos de entidades médicas, estas no imponen restricciones a los términos del dominio. Tuason et al. [19], por ejemplo, mencionan que las causas de las falencias en sus experimentos de extracción se debieron principalmente a variaciones de “puntuación” ('bmp-4', 'bmp4'), uso de diferentes tipos de numerales ('synt4', 'synt iv') y diferentes transcripciones de las letras del alfabeto griego ('igα', 'ig alpha').

Ante la imposibilidad de mantener diccionarios totalmente completos, Aït-Mokhtar y Rodrigo Mateos [20] señalan, en la descripción de la herramienta Smorph [1], que, en algunos casos, se puede describir la categoría de una palabra desconocida (PD) a partir de su terminación

morfológica. Por ejemplo, toda cadena de caracteres terminadas en ‘-ción’ es un nombre femenino singular, o toda cadena terminada en ‘-ó’ es un verbo flexivo en pretérito perfecto simple.

A tales efectos, en primer lugar, se intentará deducir a qué categoría pertenecen las PD mediante reglas de formación de palabras (1er. Nivel de análisis) y reglas sintácticas (2do. Nivel de análisis). En segundo lugar, se procederá a la conformación de sintagmas nominales que involucren PD, para luego extraerlos. Finalmente, en tercer lugar, se evaluará la precisión de las categorizaciones y, posteriormente, con el asesoramiento de profesionales del área de medicina, se verificará la posibilidad que tienen los candidatos a términos extraídos de ser promovidos a términos.

### 3. METODOLOGÍA

La presente investigación se basa, principalmente, en “deducir” la categoría de las palabras que no se encuentran en el diccionario fuente de los softwares de análisis lingüístico. Para ello, se tomarán en consideración los estudios de formación de palabras [21] y la relación entre morfología y terminología [22], como así también, los análisis de conformación de sintagmas.

Para el trabajo informático, se recurrirá a las herramientas Smorph [1] y Módulo Post Smorph (MPS) [2]. El primero permite analizar morfológicamente la cadena de caracteres, dando como salida la asignación categorial y morfológica correspondiente a cada ocurrencia de acuerdo con los rasgos declarados. MPS, por su parte, tiene como input la salida de Smorph y, a partir de reglas de recomposición, descomposición y correspondencia declaradas por el usuario, analiza la cadena de lemas resultante del análisis morfológico.

Las fuentes declarativas de Smorph están constituidas por 5 archivos: (i) ascii.txt: contiene los códigos ascii específicos tales como los separadores de oración y de párrafo; (ii) rasgos.txt: incluye etiquetas de rasgos morfológicos a aplicar en el análisis de las cadenas de caracteres con sus posibles valores (ej.: EMS ‘nombre’, ‘verbo’; Género: ‘masculino’, ‘femenino’, etcétera); (iii) term.txt: carga las diferentes terminaciones que cada lema puede presentar en su derivación morfológica (ej.: -o, -a, -os, -as); (iv) entradas.txt: es el listado de lemas y modelos correspondientes de derivación (ej. casar v1), y (v) modelos.txt: define las clases de acuerdo con los parámetros de concatenación regular de cadenas a partir de las entradas y las terminaciones (ej.: modelo v1: raíz + terminaciones de la 1ª conjugación regular + rasgos).

Las fuentes declarativas de MPS, en cambio, están constituidas por un único tipo de archivo, rcm.txt, que incluye un listado de reglas que especifican cadenas posibles de lemas con una sintaxis informatizada. Las reglas pueden ser de tres tipos: (i) de reagrupamiento:  $D + N = SN$ ; (ii) de descomposición:  $Contracc = P + D$ , y (iii) de correspondencia:  $Art = D$ .

Se utilizará el archivo entradas.txt elaborado por Infosur, y la correspondiente modelización desarrollada por el mismo equipo para las formas flexivas.

El proceso de reconocimiento de PD y posterior extracción de candidatos a términos se compone de las siguientes etapas:

- **Etapas I:** Análisis morfológico y reconocimiento de los signos de puntuación por medio de Smorph. Aquí se les asignará a las palabras desconocidas la etiqueta ‘PD’;
- **Etapas II:** Modificación del archivo term.txt mediante la asignación de terminaciones distinguidas con su correspondiente clasificación morfológica (ej.: ción pd/nom/fem/sg’, ‘-

ciones pd/nom/fem/pl'). Teniendo en cuenta el género textual del corpus, solo se incluirán las terminaciones distinguidas de los verbos de la tercera persona singular y plural, formados a partir de los sufijos '-ar', '-ear', '-ecer', '-ificar', '-izar', del presente y el pretérito del indicativo y el subjuntivo, el participio y el gerundio. No se cargará la terminación del infinitivo para evitar etiquetados erróneos (ej.: 'vascular' puede ser etiquetado como verbo). Posteriormente, se volverá a pasar el corpus por Smorph a fin de obtener las categorías que se ajusten a dichas terminaciones. También en esta etapa se considerará la posibilidad de que la PD sea un nombre propio o una sigla a partir de si presenta o no caracteres en mayúscula;

- **Etapa III:** Creación y aplicación de reglas sintácticas que permitan deducir la categoría de las PD. Aquí se hará hincapié en la estructura del sintagma nominal (SN) (Ej.: Det + PD + Adj = SN/ART+NOM+ADJ);
- **Etapa IV:** Extracción de los SN que involucran PD, en calidad de candidatos a términos. Aquí los términos serán simplificados con la técnica de stemming [23], que consiste en reducir las palabras a sus formas no flexivas y no derivativas;
- **Etapa V:** Evaluación de las categorizaciones y de los candidatos a términos extraídos.

He aquí un esquema de las etapas de trabajo.

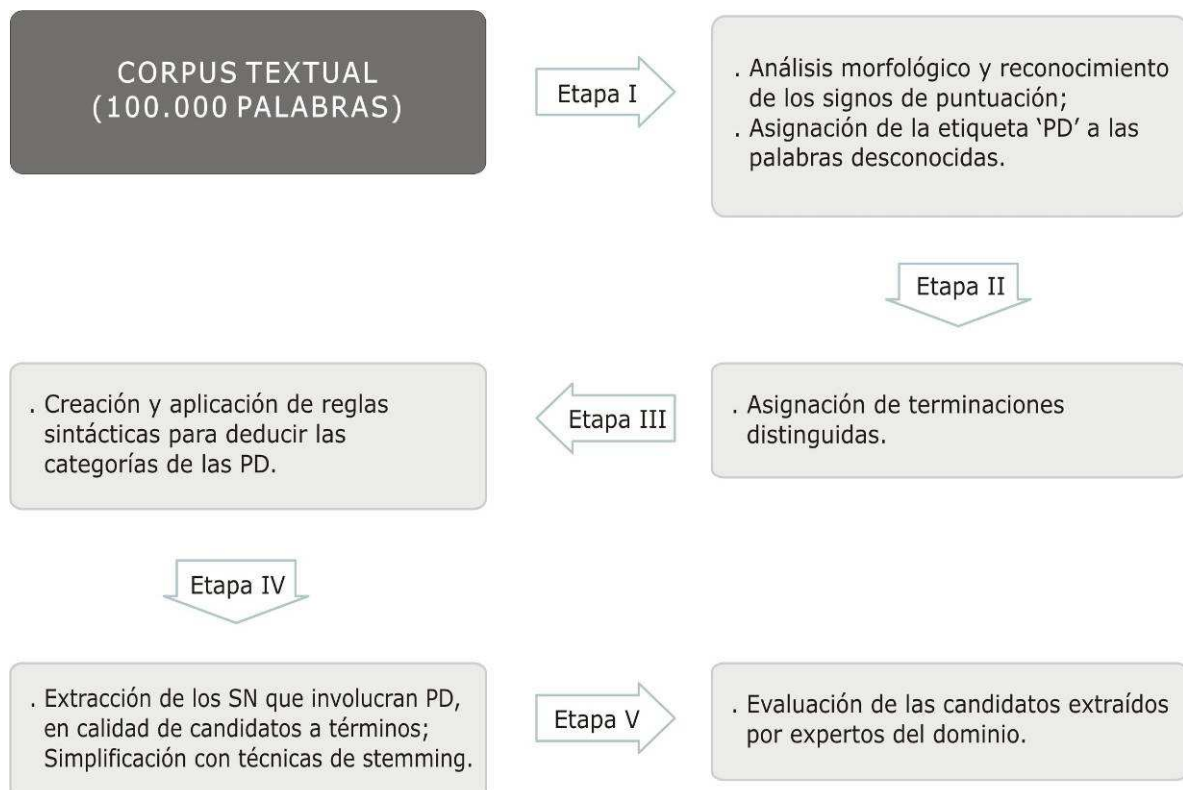


Figura 1

A continuación se ejemplificará con un breve texto del dominio médico.

## 4. EXPERIMENTACIÓN

En las etapas preliminares, se tomó un texto base para desarrollar las reglas iniciales. Aquí se presenta un fragmento de este.

### Enfermedades Vasculares

En esta categoría se encuentran patologías como el accidente cerebrovascular hemorrágico, el accidente cerebrovascular isquémico, la hemorragia subaracnoidea, las malformaciones arteriovenosas, etc., también pueden incluirse, dentro de este grupo, los casos de aneurisma. Todos ellos tienen en común que son eventos de gran severidad y que las consecuencias neurológicas del retraso en el diagnóstico y el tratamiento pueden ser hasta fatales. Dentro de los métodos de diagnóstico que se utilizan en estas patologías están: el TAC, la resonancia magnética, la angiografía cerebral y otros. El tratamiento de estos problemas depende de cada caso, aunque últimamente la mayoría de estos puede ser diagnosticado y tratado a través de la terapia endovascular. La figura de la izquierda muestra una hemorragia intraparenquimatosa.

(Mora, “Enfermedades del cerebro”, texto adaptado)

En la primera pasada, Smorph señaló como palabras desconocidas las siguientes cadenas: ‘isquémico’, ‘subaracnoidea’, ‘arteriovenosas’, ‘aneurisma’, ‘TAC’, ‘angiografía’, ‘endovascular’ e ‘intraparenquimatosa’. Luego, se adicionó en el archivo term.txt, las terminaciones distinguidas. A continuación, ejemplo de algunas de ellas:

grafía	pd/nom/fem/sg .
oso	pd/adj/masc/sg .
mente	pd/adv .
izó	pd/v/pret/ind/3p/sg .

Una vez cargadas, el texto se volvió a pasar por Smorph y se logró clasificar: ‘subaracnoidea’ (pd/adj/fem/sg), ‘arteriovenosas’ (pd/adj/fem/pl), ‘angiografía’ (pd/nom/fem/sg) e ‘intraparenquimatosa’ (pd/adj/fem/sg). Se ilustra con el etiquetado de ‘angiografía’:

```
'angiografía'.  
[ 'angiografía', 'EMS', 'pd', 'EMS', 'nom', 'GEN', 'fem', 'NUM', 'sg' ].
```

En segundo lugar, se estableció que todas las palabras que estuvieran completamente en mayúsculas fueran etiquetadas como siglas y las que comenzaran en mayúscula sin estar al inicio de la cláusula fueran nombres propios. Así, se logró clasificar correctamente ‘TAC’ (pd/abrev).

Vale aclarar que las PD que quedaron no eran verbos (conjugados, participios o gerundios), adverbios (los que no están terminados en ‘-mente’, fueron cargados en el diccionario de Smorph), preposiciones, artículos, pronombres, etcétera (todos estos fueron cargados previamente).



Asimismo, cabe destacar que, con este procedimiento, se reduce significativamente la ambigüedad, por ejemplo, los artículos seguidos de una PD no se confunden con pronombres ('la', 'las', 'lo', 'los') ya que dicha PD no puede ser un verbo conjugado (por su terminación), pero tampoco un infinitivo, ya que solo permite un pronombre clítico. Quizá el único riesgo que se podría correr sería confundir un verbo en infinitivo con un nombre en construcciones como 'el cantar de Rolando', por ejemplo. No obstante, a los efectos del presente trabajo, que consiste en la extracción de SN, la opción seguiría siendo válida, puesto que se trata de un SN con núcleo en infinitivo.

Posteriormente, se procedió a la clasificación de PD a partir del contexto sintáctico. Se tomaron en consideración a las palabras y a los signos de puntuación que rodeaban a las PD para la creación de reglas de reagrupamiento con MPS. He aquí unos ejemplos:

- Artículo + PD + Adjetivo = SN\_PD/ART+NOM+ADJ;
- Nombre + Preposición + PD + Signo de puntuación = SN\_PD/NOM+ PREP+NOM;
- Preposición + Artículo + PD = SP\_PD/PRE+ART+NOM.

Por medio de dichas reglas, se pudo reconocer 'aneurisma' e 'isquémico'. En 'aneurisma', se observa que el término tiene a su izquierda un artículo más un nombre ('los casos') seguido de una preposición ('de') y, a su derecha un punto. Por ende, se estableció que se trataba del término del SP adjunto del SN. Con respecto a 'isquémico', este está precedido de un nombre y un adjetivo; debido a que una expresión del tipo 'nombre + adjetivo + nombre' es agramatical, la única categoría posible para 'isquémico' es la de adjetivo.

Quedó por resolver el caso de 'endovascular', que no se pudo clasificar dado que no poseía una terminación distinguida y, a su vez, los elementos que la rodeaban no eran suficientes. La expresión del tipo "artículo + nombre + PD" no permitió deducir la categoría de la palabra, ya que la PD se ajustaba a cualquiera de las siguientes tres estructuras:

- Artículo + Nombre + Adjetivo (el caso del ejemplo, 'la terapia endovascular');
- Artículo + Nombre + Nombre ('las células madres');
- Artículo + Nombre + Infinitivo (—vio a— 'la paciente mejorar').

Se plantea, en este tipo de problemas, la posibilidad de postergar secuencias como esas y continuar con el reconocimiento de otras, para luego retomar las primeras con reglas de ejecución secundaria, en donde se aprovechen las secuencias previamente analizadas.

Lo que resulta evidente, no obstante, es la estrecha relación entre la PD y el nombre que la precede, sea esta de nombre-complemento, palabra compuesta o sujeto-verbo. A tales efectos, y para utilizarla momentáneamente, se creó una regla de reagrupamiento en la que las expresiones correspondientes a la estructura "artículo + nombre + PD" fueran etiquetadas como 'SN\_PD'.

Una vez realizada la categorización, se necesitó constituir reglas de reconocimientos de SN que incluyeran PD. Por el momento, he tomado las que elaboré en mi tesis doctoral para el reconocimiento de SN, en las que introduje algunas modificaciones pertinentes (cambiar la regla

‘artículo + nombre = SN’ por ‘artículo + PD = SN\_PD’, por ejemplo). No obstante, van a requerirse reglas más específicas, que contemplen fenómenos que podrían dificultar la clasificación, como por ejemplo un inciso.

Los SN\_PD detectados fueron simplificados mediante técnicas de steeming y de esta forma se obtuvieron los siguientes candidatos a términos: ‘accidente cerebrovascular isquémico’, ‘hemorragia subaracnoidea’, ‘malformación arteriovenosa’, ‘caso de aneurisma’, ‘TAC’, ‘angiografía cerebral’, ‘terapia endovascular’ y ‘hemorragia intraparenquimatosa’. De todos ellos, 7 fueron corroborados como términos específicos del dominio de medicina, por profesionales del área. El inconveniente surgió con ‘caso de aneurisma’, que no fue considerado término, pero, en dicha expresión, señalaron elementos que sí eran términos por separado: ‘caso’ y ‘aneurisma’.

## 5. ORGANIZACIÓN DEL TRABAJO Y DIFICULTADES

La experimentación se llevará a cabo como se ha ejemplificado, sobre un corpus que actualmente se está construyendo. Hasta ahora, se reunieron textos de medicina que suman un total de 1.000.000 de palabras y que están siendo revisados manualmente por médicos a fin de obtener listas de referencias de los términos allí incluidos.

Las dificultades que se han presentado hasta el momento tienen que ver con PD que se encuentran solas, es decir, que no están rodeadas por otros elementos que permitan decir su categoría, o bien, cuando hay una combinación de ellas. Otro de los inconvenientes son los nombres propios que, en algunos casos, pueden ser términos (‘Alzheimer’). Por último, hay que mencionar los errores de ortografía cometidos por los autores de los textos.

A tales efectos, será necesario recurrir a técnicas de evaluación automática de la calidad de los candidatos a términos extraídos, para ello, no se descarta para recurrir a métodos estadísticos de evaluación, entre otros.

## Referencias

- [1] AÏT MOKTHAR, S. (1998) SMORPH: Guide d’utilisation. Rapport technique. Universidad Blaise Pascal/GRIL. Clermont-Fd.
- [2] Abbaci, F. (1999) Développement du Module Post-Smorph. Memoria del DEA de Linguistique et Informatique. Universidad Blaise-Pascal/GRIL. Clermont-Fd.
- [3] Sager, J. (1993) Curso práctico sobre el procesamiento de la terminología. Madrid: Fundación Sánchez Ruíz Pérez.
- [4] Vivaldi, J. (2011) “Terminología y Wikipedia”. Seminario IULATerm. Barcelona: Universitat Pomeu Fabra.
- [5] Moreno-Sandoval, A. Terminología y sociedad del conocimiento. 2009.

- [6] Krauthammer, M. y Nenadić, G. (2004) “Term identification in the biomedical literature”. En J. of Biomedical Informatics.
- [7] Barrón-Cedeño et al. (2009) “An improved automatic term recognition method for spanish”. In A. Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 5449 of Lecture Notes in Computer Science, pages 125-136. Springer Berlin / Heidelberg.
- [8] Bosman, W. and Vossen. P. (2010) Bootstrapping language neutral term extraction. In (N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta, may 2010. European Language Resources Association.
- [9] Bonin, F. et al. (2010) “A contrastive approach to multi-word extraction from domain-specific corpora”. In (N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta. European Language Resources Association.
- [10] Gelbukh, G. et al. (2010) “Automatic term extraction using log-likelihood based comparison with general reference corpus”. In Proceedings of the Natural language processing and information systems, and 15th international conference on Applications of natural language to information systems, NLDB'10, pages 248-255, Berlin, Heidelberg. Springer-Verlag.
- [11] Castro, E. (2010) “Automatic identification of biomedical concepts in spanish-language unstructured clinical texts”. In Proceedings of the 1st ACM International Health Informatics Symposium, IHI '10, pages 751-757, New York, NY, USA. ACM.
- [12] Lacoste, C. et al. Medical-image retrieval based on knowledge-assisted text and image indexing. IEEE Trans. Circuits Syst. Video Techn., 17(7):889-900, 2007.
- [13] D. Sánchez, et al. Web-based semantic similarity: An evaluation in the biomedical domain. Int. J. Software and Informatics, 4(1):39-52, 2010.
- [14] López Rodríguez, C. et al. Gestión terminológica basada en el conocimiento y generación de recursos de información sobre el cáncer: el proyecto Oncoterm. Revista E Salud, 2(8), 2006.
- [15] López Rodríguez, C. et al. Terminología basada en el conocimiento para la traducción y la divulgación médicas: el caso de Oncoterm. Panace, VII (24):228-240. 2006.
- [16] Vivaldi, J. and Rodríguez, H. Using wikipedia for term extraction in the biomedical domain: First experiences. Procesamiento del Lenguaje Natural, 45:251-254, 2010.
- [17] Vivaldi, J. et al. Automatic summarization using terminological and semantic resources. In (N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, Proceedings of the Seventh conference on International Language Resources and Evaluation, Valletta, Malta, may 2010. European Language Resources Association.
- [18] Alarcón, R. Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios. In [CD-ROM], editor, Serie Tesis, number 26. Barcelona: IULA, 2010. ISBN: 13: 978-84-89782-46-4.
- [19] Tuason et al. (2004) “Biological Nomenclature: A source of Lexical Knowledge and Ambiguity”, en: Proceedings of Pac Symp Biocomput.
- [20] Ait-Mokthar, S. y Rodrigo Mateos, J. (1995) “Segmentación y análisis morfológico en español utilizando el sistema Smorph”, en SEPLN Revista/'95. Jaén: SEPL.
- [21] Lang, M. (2002) Formación de palabras en español. Madrid: Cátedra.

- [22] Cabré Castelví, M. (2006) “Morfología y terminología”. En Feliú Arquiola (Ed.), La morfología a debate. Jaén: Universidad de Jaén.
- [23] Manning, C. et al. Language models for information retrieval. In An Introduction to Information Retrieval, chapter 12. Cambridge University Press, 2008.

## **Los sistemas de ayuda en la navegación hipertextual**

### **The support systems in hypertext navigation**

**Bárbara Méndez**

Universidad Nacional de Rosario  
Facultad de Humanidades y Artes  
Grupo Infosur  
barbaramendez555@hotmail.com

### **Abstract**

Hypertext navigation comprises a dynamic and interactive activity which allows the reader to move not necessarily in a sequential manner through the information contained in the digital document. However, the possible readings offered by a hypertext document may not get fully exploited when the document does not have “assisted navigation”, that is to say, tools that guide the reader through different alternative routes.

This paper studies the elements and components of the hypertext navigation systems. I start from the assumption that these tools have the function of helping to improve navigation and to facilitate access to information.

First, I study the basic structure of the hypertext. Then, following Codina's description, I conduct a brief analysis of aid tools in hyperdocuments. Finally, I describe the aid system in the doctoral thesis “Hipertexto: El nuevo concepto de documento en la cultura de la imagen” realizada por María Jesús Lamarca Lapuente. (“Hypertext: The new concept of document in an image culture” by María Jesús Lamarca Lapuente).

**Keywords:** Hypertext, navigation systems of hypertext.

### **Resumen**

La navegación hipertextual comprende una actividad dinámica e interactiva permitiendo al lector realizar trayectos no necesariamente secuenciales a través de la información que contiene el documento digital. Sin embargo, las posibles lecturas que ofrece el hipertexto pueden desaprovecharse cuando el documento no presenta una navegación asistida, esto es, herramientas que orienten al lector en los diferentes recorridos posibles.

En este trabajo se estudian los elementos y componentes de los sistemas de ayuda a la navegación de un hipertexto. Parto de la hipótesis de que estas herramientas tienen como función ayudar a mejorar la navegación y favorecer el acceso a la información.

En primer lugar, estudio la estructura básica que presenta el hipertexto, luego realizo, de acuerdo a Codina una breve descripción de las herramientas de ayuda en los hiperdocumentos. Por último,

analizo el sistema de ayuda en la tesis doctoral “Hipertexto: El nuevo concepto de documento en la cultura de la imagen” realizada por María Jesús Lamarca Lapuente.

**Palabras Claves:** hipertexto, sistema de ayuda en la navegación hipertextual.

## 1. INTRODUCCIÓN

Un hipertexto es un documento digital cuya información se encuentra organizada en una red de nodos enlazados a través de los cuales los lectores pueden navegar libremente en forma no lineal. La estructura hipertextual posibilita diferentes formas de acceso que trascienden la secuencialidad propia de los documentos impresos. De esta manera, el lector de hipertextos puede realizar diferentes itinerarios de lectura de acuerdo a sus conocimientos e intereses.

Codina [1] plantea que la libertad de los lectores de un hipertexto de elegir su propio recorrido de lectura tiene un inconveniente: “el desbordamiento cognitivo”, éste se produce cuando el lector no puede controlar todas las bifurcaciones del sistema y se ve incapaz de explorar todos los caminos que le ofrece. En otras palabras, las posibles lecturas que posibilita el hipertexto pueden desaprovecharse cuando el documento no presenta una navegación asistida, esto es herramientas que orienten al lector en los diferentes recorridos posibles.

En este trabajo me propongo estudiar los sistemas de ayuda a la navegación hipertextual. Parto de la hipótesis de que estas herramientas tienen como función ayudar a mejorar la navegación y favorecer el acceso a la información por parte del usuario.

En primer lugar, siguiendo a Lamarca Lapuente [2] estudio la estructura básica que presenta el hipertexto, luego me centro en dos tipos de navegación hipertextual y realizo, de acuerdo a Codina, una breve descripción de las herramientas de ayuda que contiene el hipertexto. Por último, analizo el sistema de ayuda en la tesis doctoral “Hipertexto: El nuevo concepto de documento en la cultura de la imagen” realizada por María Jesús Lamarca Lapuente.

## 2. ELEMENTOS BÁSICOS DE UN HIPERTEXTO

Lamarca Lapuente sostiene que el concepto de estructura hipertextual incluye tres conceptos distintos que hacen referencia a diferentes aspectos del hipertexto:

La **Arquitectura estructural** que se compone de nodos (unidades básicas que contienen la información), enlaces (interconectan los nodos) y anclajes (marcan el inicio y el destino de cada enlace). Estos elementos básicos y simples dan lugar al desarrollo de estructuras muy heterogéneas y complejas que permiten acceder a la información mediante el recorrido a través de los nodos, mediante los enlaces.

Los nodos son los elementos que contienen la información y son las unidades básicas del hipertexto. Se trata de las porciones de información (palabras, frases, imágenes, etc.) que entran en relación con otros nodos a los que proporcionan acceso. Cada nodo pertenece únicamente a un documento. Los Enlaces interconectan nodos, se trata de las conexiones o vínculos que se establecen entre segmentos de información, es decir, entre los nodos que relacionan los documentos.

**La Arquitectura navegacional** que contiene las herramientas de acceso a la información y navegación por los nodos de información contenida en los documentos, generalmente atendiendo a una estructura conceptual o temática.

**La Arquitectura funcional** que comprende los componentes, mecanismos y herramientas esenciales que hacen posible el establecimiento de la propia arquitectura estructural y navegacional del hipertexto.

### 3. LA NAVEGACIÓN HIPERTEXTUAL

La navegación hipertextual comprende una actividad dinámica e interactiva en la que lector a partir de la información que le parece pertinente realiza una lectura no lineal. Un hipertexto brinda dos tipos de navegación: la navegación superpuesta y la navegación implicada.

La navegación implicada se efectúa con los enlaces incrustados o implicados en los propios nodos. La navegación superpuesta se realiza en los desplazamientos entre un elemento de representación (menú, sumario, mapa, índice) y un nodo o viceversa.

Las herramientas de ayuda permiten visualizar y navegar en un texto a través de las unidades textuales y por tanto: a) facilitan el acceso a la información ya que funcionan a modo de brújula para ayudar al usuario a moverse por el hipertexto, b) evitan la sensación de pérdida tan habitual en algunas producciones digitales interactivas, c) brindan representaciones superestructurales de los contenidos y d) permiten la búsqueda y recuperación de información.

A continuación se presentan las principales las principales herramientas de ayuda a la navegación:

Tabla n° 1: Herramientas de navegación

<b>ELEMENTOS DE UN SISTEMA DE NAVEGACIÓN</b>	
<b>Sumarios</b>	<ul style="list-style-type: none"> <li>• Globales</li> <li>• Locales</li> </ul>
<b>Índices</b>	<p>Índices analíticos:</p> <ul style="list-style-type: none"> <li>• general</li> <li>• de nombres</li> <li>• de temas</li> <li>• de lugares</li> </ul>
<b>Orientaciones</b>	<p>Informes de situación:</p> <ul style="list-style-type: none"> <li>• Contexto</li> <li>• Historia</li> <li>• Visión global</li> </ul>

<b>Acciones de navegación y lectura</b>	<ul style="list-style-type: none"> <li>• Enlaces desde: <ul style="list-style-type: none"> <li>○ cualquier nodo al sumario global del hiperdocumento.</li> <li>○ cualquier nodo a los índices del hiperdocumento</li> <li>○ un nodo a nodos adyacentes (anterior, siguiente), si es el caso</li> <li>○ cualquier nodo al sumario local, si es el caso</li> <li>○ histórico de lectura</li> </ul> </li> <li>• Recorridos posibles que debe facilitar la navegación: <ul style="list-style-type: none"> <li>○ estructurales (jerárquicos y verticales)</li> <li>○ semánticos (horizontales)</li> <li>○ aleatorios</li> </ul> </li> <li>• Anotaciones: <ul style="list-style-type: none"> <li>○ notas añadidas por el lector</li> <li>○ marcas de lectura</li> </ul> </li> </ul>
-----------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

#### 4. EL SISTEMA DE AYUDA EN LA TESIS DOCTORAL DE MARÍA JESÚS LAMARCA LAPUENTE

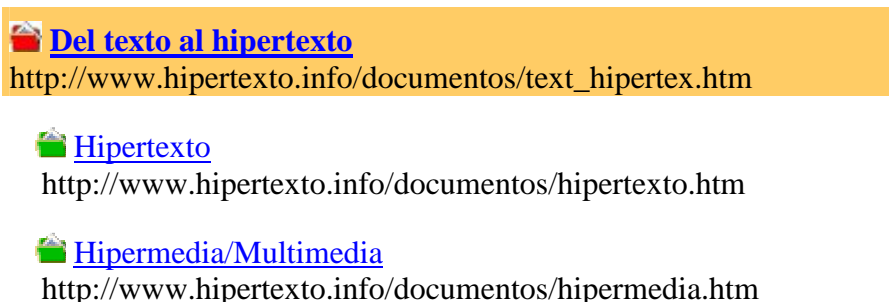
La tesis doctoral “*Hipertexto: El nuevo concepto de documento en la cultura de la imagen*” realizada por María Jesús Lamarca Lapuente posibilita una lectura secuencial y no secuencial de la tesis. El recorrido secuencial también es guiado a través del ícono *siguiente* que aparece al inicio y al final de las páginas. En cada página, se presentan diferentes íconos de navegación que dan cuenta de las herramientas de ayuda a las que se puede acceder

Además, ofrece una sección de ayuda a la navegación de hipertexto en la que se explica el sistema de herramientas que posibilitan la orientación del usuario, la búsqueda y la recuperación de información. Este conjunto de herramientas están disponibles y accesibles desde todas y en cada una de las páginas del hipertexto.

Las herramientas utilizadas son las siguientes:

- **Mapas de navegación:** la tesis presenta dos mapas. El primero muestra las herramientas disponibles en el hipertexto. El segundo contiene los temas y subtemas principales del hipertexto.

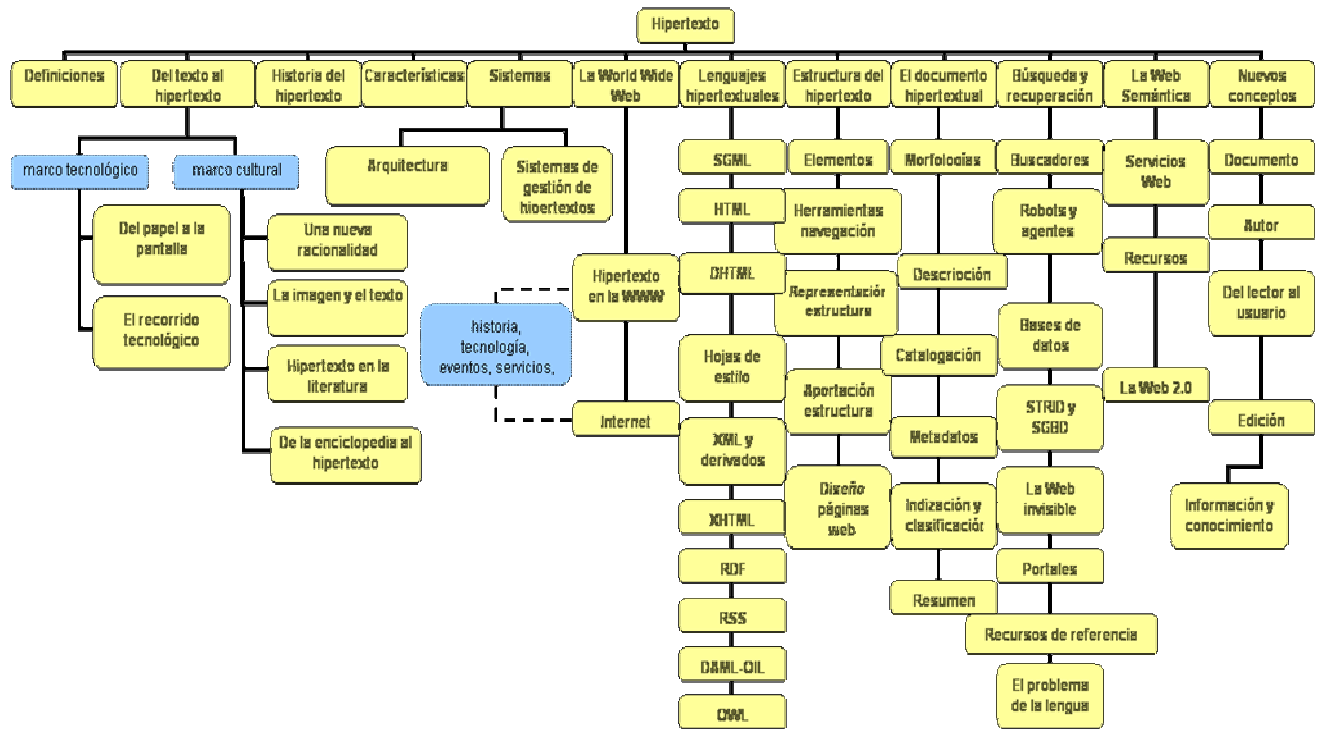
Figura nº 1: Mapa de navegación





- **Mapas conceptuales:** La tesis presenta dos mapas conceptuales. El primero presenta los contenidos del hipertexto, estableciendo una red de relaciones jerárquicas y semánticas sobre los temas que se desarrollan.

Figura 2: Mapa conceptual jerárquico y semántico



El siguiente mapa muestra las herramientas y secciones principales que contiene este hipertexto.

Figura n ° 4: Mapa conceptual sobre las herramientas de navegación del hipertexto





- **Tabla de contenidos:** Esta tabla es similar al mapa conceptual nº1, contiene todos y cada uno de los temas y subtemas principales. Esta tabla es dinámica y se puede plegar y desplegar las distintas ramas.

Figura nº 5: Tabla de contenido de la tesis

  Introducción


 Metodología

 Objetivos



 Ayuda a la navegación por este hipertexto

  Definiciones/Hiperdefiniciones


 Hipertexto


 Hipermedia/Multimedia

 Documento/Hiperdocumento

  Del texto al hipertexto

  Del papel a la pantalla

 La escritura

 La lectura

 La interfaz gráfica

- **Índice temático:** Se accede al índice de temas y subtemas de la tesis.

Figura n° 6: Índice temático

Hipertexto  
Arquitectura de un sistema hipertextual  
Características del hipertexto  
Cronología de los SGH  
Cronología del hipertexto  
Definiciones/Hiperdefiniciones  
Del texto al hipertexto  
El documento hipertextual  
Elementos de un hipertexto  
Estructura de un hipertexto  
Hipertexto  
Hipertexto en la literatura  
Historia del hipertexto  
Los sistemas pre-web de gestión de hipertextos  
Principales eventos sobre hipertexto  
Sistemas de gestión de hipertextos  
Sistemas de hipertexto

- **Índice de autores:** está ordenado alfabéticamente, se informa acerca de cada autor y se presenta la sección donde se citó a dicho autor.

- **Tabla de documentos:** se muestra una imagen miniaturizada de todas las páginas que componen este documento.

Figura n° 7: tabla de documentos



## **Buscador**

Para realizar la búsqueda, se siguen estos pasos: el usuario introduce uno o más términos de búsqueda; el sistema busca los términos por medio de índices y otros mecanismos; y el sistema responde mostrando los resultados. Para llevar a cabo este proceso se necesita, además de establecer tanto interfaces de consulta como de respuesta adecuados, establecer los mecanismos que hagan posible esta tarea (creación de bases de datos, uso de indexación manual o automática, diseño y aplicación de herramientas de búsqueda automatizadas, etc.)

La presentación de resultados puede ser muy variada: el sistema devuelve la primera ocurrencia del término, una lista de ocurrencias o puede integrar los resultados de la búsqueda en el mapa del hipertexto, resaltando los nodos en los que aparece el término buscado.

## **5. CONCLUSIÓN**

El objetivo principal de este trabajo ha sido el estudio de los sistemas de ayuda en la navegación de hiperdocumentos. Planteamos que las herramientas guían al usuario a moverse por la red hipertextual, orientando al lector en los diferentes recorridos posibles y facilitando el acceso, la búsqueda y la recuperación de la información.

La tesis doctoral de Lamarca Lapuente dispone de una serie de herramientas que permiten una navegación completa y eficaz. El lector puede acceder directamente a los mapas, índices e indicaciones ya que forman parte de la estructura de navegación.

## **Referencias**

- [1] Codina, Lluís. H de Hipertext, o la teoría de los hipertextos revisada. Cuadernos de Documentación Multimedia. 1997, nº 6-7
- [2] Lamarca Lapuente, María Jesús. Hipertexto: El nuevo concepto de document en la cultura de la imagen, Facultad de Ciencias de la Información. Dpto. de Biblioteconomía y Documentación; Universidad Complutense de Madrid, 2007: <http://www.hipertexto.info>
- [3] Codina, Lluís. El libro digital y la www. Tauro Ediciones, Madrid, 2000.

# Una propuesta para la extracción automática del sintagma adverbial

## A proposal for the automatic extraction of the adverbial syntagm

**Andrea Rodrigo**

Grupo INFOSUR, U.N.R.  
Rosario, Argentina  
andreafrodrigo@yahoo.com.ar

**Rodolfo Bonino**

Grupo INFOSUR, U.N.R.  
Rosario, Argentina  
rodolfobonino@yahoo.com.ar

### Abstract

The aim of this work is to postulate a linguistic formalism for the adverbial syntagm, understood it as a unit constituted by a head adverbial syntagm (SADV<sub>N</sub>) and a complement- to be implemented into Nooj [2]. In order to do so, the differences between the head adverbial syntagm (SADV<sub>N</sub>) and the adverbial syntagm (SADV) are stated and some SADV<sub>N</sub> which frequently occur in the formation of SADV are taken into consideration by observing authentic texts. Thus, we can find sequences such as *más abajo de la mesa* (*further under the table*), *más lejos que lo esperable* (*farther away than expected*), etc., where the SADV<sub>N</sub> takes an SPN (head prepositional syntagm) as a complement, a gerund construction or a comparative complement which, in turn, also contain an SNN or another SPN as complements.

Previous works are taken as a point of departure, in which the description of the SADV<sub>N</sub> was formalized according to the properties of the 5P Paradigm (Bès, G.G. 1999) and implanted into other software tools. Nooj is a computational device developed by **Max Silberztein (2002)**, used for the formalization of linguistic phenomena and text analysis of natural languages. Given that it only makes concatenating operations, Nooj does not allow to implant 5P's requirement and exclusion properties but is highly efficient to operate with linearity properties, thus making it feasible to corroborate the linguistic hypothesis on the SADV previously stated with other computational formalism.

**Key words:** computational linguistics, Spanish, adverbs, head syntagm, adverbial syntagm, Nooj application into Spanish

### Resumen

En este trabajo, se trata de postular un formalismo lingüístico que permita que el sintagma adverbial (SADV), entendido como una unidad constituida por SADV<sub>N</sub> (sintagma adverbial núcleo) y un complemento, pueda ser implantado en Nooj [2]. Para ello, se diferencian las nociones de sintagma adverbial núcleo (SADV<sub>N</sub>) y sintagma adverbial (SADV) y se consideran algunas de las

combinaciones de SADVN que intervienen más frecuentemente en la conformación de los SADV, según la observación de textos reales. Así se pueden encontrar secuencias como: *más abajo de la mesa*, *más lejos que lo esperable*, donde se ve que el SADVN se complementa con un SPN (sintagma preposicional núcleo) o un complemento comparativo que, a su vez, contienen un SNN u otro SPN como complementos.

Se toman como punto de partida trabajos anteriores, donde la descripción del SADVN se formalizó según las propiedades del Paradigma 5P (Bès, G.G. 1999) [4] y se implantó en otras herramientas informáticas. NooJ es un dispositivo computacional desarrollado por Max Silberztein (2002), que se utiliza para la formalización de fenómenos lingüísticos y el análisis de textos en lenguas naturales. Dado que solo efectúa la operación de concatenación, no permite implantar las propiedades de exigencia y de exclusión de 5P; pero tiene gran eficacia para operar con propiedades de linealidad, por ello hace factible corroborar con otro formalismo computacional las hipótesis lingüísticas planteadas previamente en torno al SADV.

**Palabras claves:** linguística computacional, español, adverbios, sintagma núcleo, sintagma adverbial, aplicación Nooj en español

## 1. INTRODUCCIÓN

Teniendo en cuenta la noción de sintagma núcleo<sup>1</sup>, sintagmas que comienzan en el inicio de la construcción y finalizan en el núcleo; en este trabajo, se trata de postular un formalismo lingüístico que permita que el sintagma adverbial (SADV), entendido como una unidad constituida por SADVN (sintagma adverbial núcleo) y un complemento, pueda ser implantado en NooJ. Para ello, se diferencian las nociones de sintagma adverbial núcleo (SADVN) y sintagma adverbial (SADV) y se consideran algunas de las combinaciones de SADVN que intervienen más frecuentemente en la conformación de los SADV, según la observación de textos reales. Se pretende abordar las siguientes combinaciones<sup>2</sup>:

- **SADVA** [SADVN + SPN (SNN)]: Ej: *más abajo de la mesa, arriba de la cama, bien enfrente de tu casa, poco después de las diez*<sup>3</sup>, *más allá del gobierno*.<sup>4</sup>
- **SADVB** [En una comparativa: SADVN + complemento comparativo (CONJ+ SNN) (CONJ+SNN)]: Ej: *más lejos que lo esperable, más aquí que allá*

Se toman como punto de partida trabajos anteriores, donde la descripción del SADVN se formalizó según las propiedades del Paradigma 5P (Bès, G.G. 1999) [4] y se implantó en otras herramientas informáticas. Aquí utilizamos NooJ, que es un dispositivo computacional desarrollado por Max Silberztein (2002) para la formalización de fenómenos lingüísticos y el análisis de textos en lenguas naturales. Dado que solo efectúa la operación de concatenación, no permite implantar las propiedades de exigencia y de exclusión de 5P; pero tiene gran eficacia para operar con propiedades de linealidad, por ello hace factible corroborar con otro formalismo computacional las hipótesis lingüísticas planteadas previamente en torno al SADV.

<sup>1</sup> Según la investigación que lleva a cabo el Grupo Infosur, con los lineamientos trazados por el GRIL (Groupe de Recherche dans les Industries de la Langue) de la Universidad Blaise Pascal de Clermont Ferrand.

<sup>2</sup> Se da cuenta de las estructuras más frecuentes, según el banco de datos del Grupo Infosur que dispone de un corpus de 100.000 palabras de periódicos argentinos.

<sup>3</sup> Ejemplo citado en Rev. Infosur N° 5, Extracción del Sintagma Nominal, (Solana, Rodrigo, Méndez), p. 18

<sup>4</sup> Pag. 12, 4/01/04

Nuestro objetivo es crear diccionarios y gramáticas que, aplicados a secuencias del lenguaje natural, posibiliten la extracción de sintagmas adverbiales (SADV), a partir de reglas establecidas. Para lograrlo se crean diccionarios donde las entradas léxicas se asocian con rasgos que permiten calcular su comportamiento sintáctico. El rasgo categorial (adverbio, nombre, determinante, preposición) en muchos casos debe ser complementado con otros que identifican propiedades específicas de subconjuntos de elementos integrantes de la categoría; de modo que en los diccionarios que elaboramos las palabras no se definen por su contenido semántico sino por los conjuntos (de uno o más elementos) de rasgos que se asocian entre sí y caracterizan la proyección sintáctica de la entrada léxica.

## 2. DESCRIPCIÓN DEL SADV

El SADV (sintagma adverbial) está formado por los siguientes sintagmas núcleos:

### a. El SADV<sub>N</sub> (sintagma adverbial núcleo)

Está conformado por adverbios, lo que son clasificados por sus combinaciones, según Rodrigo (2011) [1].

ADV

ADV1 Ej: *alrededor, actualmente*

ADV 2

ADV2a

ADV2am<sup>5</sup> Ej: *aproximadamente,*

ADV2a0 Ej: *abajo, lejos*

ADV2b

ADV2bm Ej: *absolutamente,*

ADV2b0 Ej: *casi, menos, muy*

ADV2c

ADV2cm Ej: *admirablemente, inmediatamente,*

ADV2c0 Ej: *más*

Los adverbios pueden aparecer en grupos de dos o bien, solos. Entre las combinaciones más frecuentes de adverbios se observa:

[ADV2c0 + ADV2a0] Ej: *más abajo, bien enfrente, más allá*

[ADV2b0 + ADV2c0] Ej: *muy bien*

[cuant + ADV2a0] Ej: *poco después*

A su vez, en algunos SADV, el SADV<sub>N</sub> está seguido por un complemento comparativo.

<sup>5</sup> Se llaman 2am y 2a0 en lugar de 2a1 y 2a2 porque no era posible “repetir el rasgo 1 y 2” según la sintaxis que requiere NooJ. Lo mismo en 2b y 2c.

### b. El SPN (sintagma preposicional núcleo)

Está conformado por una preposición núcleo que se ubica al principio del sintagma<sup>6</sup>, seguido de un SNN (sintagma nominal núcleo) en su interior.

(2) [*de (las diez) SNN*] SPN

El SNN está constituido por<sup>7</sup>:

determinante + nombre. Ej:

(3) *las diez*

En la categoría determinante, se incluye: artículo, posesivo, según Solana/Rodrigo (2005) [6] y Rodrigo (2006) [3].

También puede observarse que el SPN incluye al SADVN en su interior,

(4) [*en (adelante) SADVN*] SPN

Es factible que el SPN se integre en una unidad mayor, el SP (sintagma preposicional).

## 3. LA HERRAMIENTA INFORMÁTICA

NooJ [2] es una herramienta informática para el tratamiento de las lenguas naturales desarrollada por Max Silberztein a partir del año 2002; analiza textos digitalizados mediante la aplicación de diccionarios y gramáticas creadas previamente; es de libre acceso y, actualmente, es utilizado por investigadores de varias universidades del mundo para la modelización de diversas lenguas. Sus usuarios intercambian conocimientos a través de un foro de Internet y realizan congresos anuales. El autor colabora activamente con los proyectos que utilizan el programa, asesorando a los investigadores y efectuando las modificaciones necesarias para la resolución de problemas específicos de cada investigación.

El procedimiento de implantación requiere la creación de varios archivos:

a) **Definición de propiedades (.def):** se declaran los rasgos que se utilizarán para etiquetar las entradas de los diccionarios. Estos rasgos pueden ser utilizados por separado o en forma conjunta en las gramáticas sintácticas. Por ejemplo: si un sintagma requiere la presencia de cualquier adverbio solo se utilizará el rasgo categorial ADV, que incluye a todos los adverbios (ADV1, ADV2am, ADV2a0, ADV2bm, ADV2b0, ADV2cm y ADV2C0), en cambio si requiere la presencia de cualquier adverbio en *-mente* al rasgo categorial se adicionará el rasgo +m, que incluye a los ADV2am, ADV2bm y ADV2b0.

b) **Gramáticas morfológicas (.nof):** se utilizan para obtener automáticamente las variaciones morfológicas flexivas o derivacionales de cada entrada léxica, de modo que en los diccionarios NooJ se declara el lema y el modelo flexivo o derivacional que le corresponde y el sistema genera automáticamente todas las variaciones que se indican en las gramáticas morfológicas que el diccionario utiliza. El tratamiento de la morfología es muy eficiente, porque no solo efectúa las

<sup>6</sup> El SPN es el único sintagma núcleo que comienza con el núcleo, a diferencia de los demás, que terminan en el núcleo.

<sup>7</sup> Nos referimos a los snn's más frecuentes según la observación de textos reales.



operaciones de sustracción y de concatenación al final de una cadena, sino también las de sustracción, cambio y duplicación en lugares que pueden ser determinados por el usuario (por ejemplo, al final de palabra, al principio de palabra, dos caracteres a la izquierda, tres caracteres a la derecha, etc.). En trabajos previos, [10] se propone una morfología flexional del verbo en español.

c) **Diccionarios** (.dic): en los diccionarios se declaran las gramáticas morfológicas que se van a utilizar y las entradas léxicas con la etiqueta categorial, el modelo de flexión o derivación que se le aplica y los demás rasgos sintácticos y semánticos que la caracterizan. Los diccionarios se compilan y el sistema genera automáticamente un nuevo diccionario .nod, con todas las variaciones morfológicas de cada entrada, que es el que utiliza el analizador.

d) **Gramáticas sintácticas** (.nog): se declaran las reglas sintácticas que se aplican en la formación de sintagmas. Tanto las gramáticas morfológicas como las sintácticas se pueden declarar en forma de reglas (rule editor) como en forma de gráficos (graphical editor) y son recursivas en tanto permiten utilizar en la definición el elemento definido. Por ejemplo: el sintagma preposicional núcleo se puede definir como:  $SPN = P + :SN$  y el sintagma nominal que aparece en el SPN como  $SN = :SNN + :SPN$ ; como se ve, en la definición de SN (sintagma nominal) se utiliza la definición de SPN, que lo contiene; esto permite analizar secuencias como *en la casa de su hermano* [ $SPN$ en [ $SNN$  la casa [ $SPN$ de [ $SNN$ su hermano]]]].

e) **Textos**: (.not) se cargan o se importan los textos que se pretende analizar.

Los pasos para efectuar el análisis son los siguientes:

a) Se abre el texto que se quiere analizar: *File* → *Open* → *Text* .

b) Se compilan los diccionarios que se van a utilizar en *Lab* → *Dictionary*. Este proceso se debe aplicar incluso en los diccionarios de categorías invariables porque, como se explicó más arriba, el analizador utiliza el diccionario .nod generado automáticamente por la compilación del archivo.dic

c) En *Info* → *Preferences* se visualizan tres pestañas (*General*, *Lexical Analysis* y *Syntactic Analysis*). En *General* se selecciona el idioma; en *Lexical Analysis* los diccionarios y en *Syntactic Analysis* las gramáticas sintácticas. Las gramáticas morfológicas se aplican indirectamente a través de los diccionarios.

d) En *TEXT* se selecciona *Linguistic Analysis*, que aplica todos los diccionarios y gramáticas sintácticas seleccionadas. El análisis completo se puede ver seleccionando *Show Text Annotation Structure*; si se pretende encontrar una secuencia específica de palabras o etiquetas de palabras, o el resultado de una sola gramática; se debe seleccionar *TEXT* → *Locate* que da dos alternativas: *a NooJ regular expression*, que permite buscar palabras o secuencias de palabras (*abajo*, *más abajo de la mesa*) o etiquetas categoriales (<ADV>, <SADV>) y *a NooJ grammar*, que selecciona las secuencias lingüísticas que genera una gramática determinada. En ambos casos, los resultados de *Locate* se obtienen haciendo clic sobre las etiquetas coloreadas que se encuentran en la parte inferior derecha de la ventana. Los diferentes colores se utilizan únicamente para seleccionar el color de la fuente de salida.

## 4. IMPLANTACIÓN DE NOOJ PARA LA EXTRACCIÓN DEL SINTAGMA ADVERBIAL

Primero es preciso que NooJ reconozca cada uno de los sintagmas núcleos y las categorías que los integran. Por tanto en el archivo correspondiente a diccionario se declaran las etiquetas según se ve a continuación:

a. Diccionario de adverbios, (siempre se adosa un fragmento de cada diccionario)

tampoco,ADV+1  
todavía,ADV+1  
abiertamente,ADV+2+a+m  
abnegadamente,ADV+2+a+m  
abruptamente,ADV+2+a+m

b. Diccionario de cuantificadores

poco,CUANT  
bastante,CUANT  
un poco,CUANT  
demasiado,CUANT  
nada,CUANT  
medio,CUANT  
algo,CUANT

c. Diccionario de preposiciones

a,PREP  
ante,PREP  
bajo,PREP  
con,PREP  
contra,PREP  
de,PREP  
desde,PREP  
en,PREP  
entre,PREP  
hasta,PREP  
hacia,PREP  
para,PREP  
por,PREP  
según,PREP  
sin,PREP  
sobre,PREP  
tras,PREP  
al,CONTR  
del,CONTR

d. Diccionario de nombres

abrigo,N+FLX=ABRIGO  
gobierno,N+FLX=ABRIGO  
mesa,N+FLX=MESA  
casa,N+FLX=MESA  
cena,N+FLX=MESA  
cama,N+FLX=MESA  
altura,N+FLX=MESA

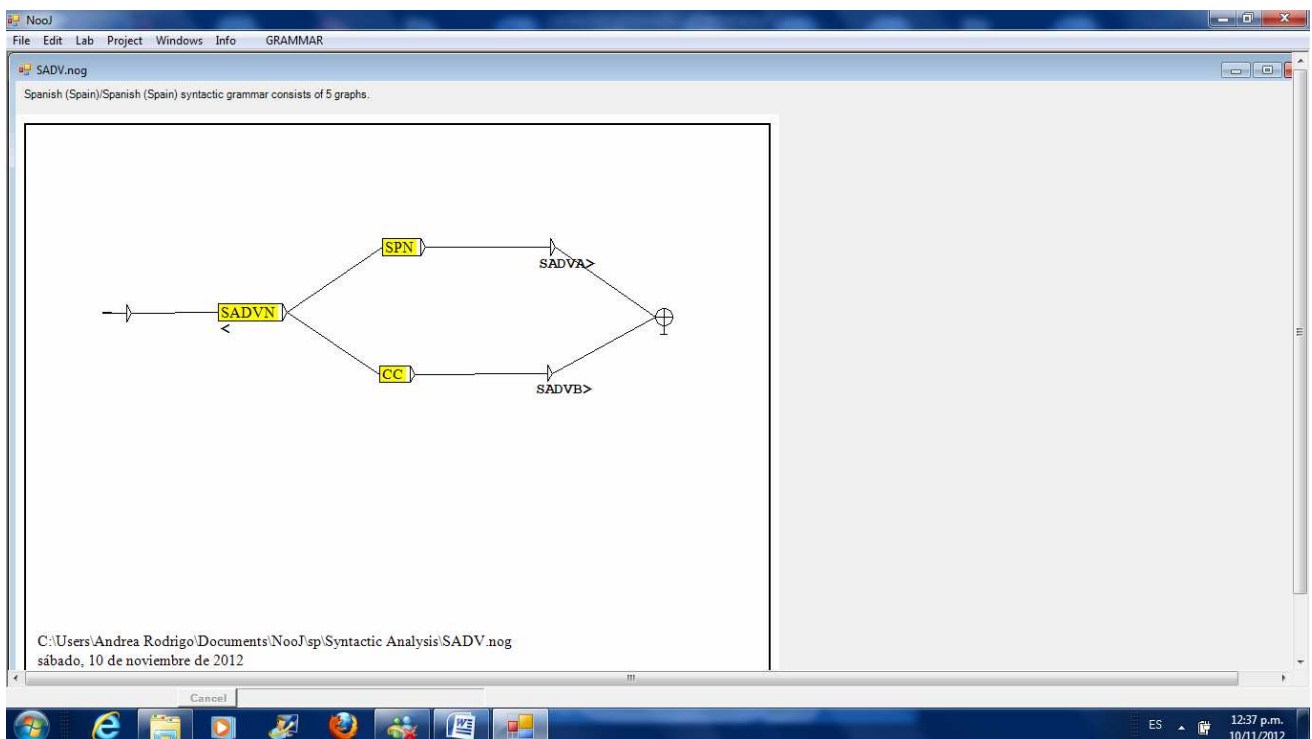
esquina,N+FLX=MESA  
 accionista,N+FLX=ACCIONISTA  
 niño,N+FLX=NIÑO  
 hermano,N+FLX=NIÑO  
 acumulador,N+FLX=ACUMULADOR  
 edad,N+FLX=EDAD  
 vez,N+FLX=VEZ

e. Diccionario de conjunciones

que,CONJ+subord  
 si,CONJ+subord  
 porque,CONJ+subord  
 conque,CONJ+subord  
 y,CONJ+coord  
 ni,CONJ+coord  
 pero,CONJ+coord  
 sino,CONJ+coord  
 e,CONJ+coord  
 u,CONJ+coord  
 o,CONJ+coord.

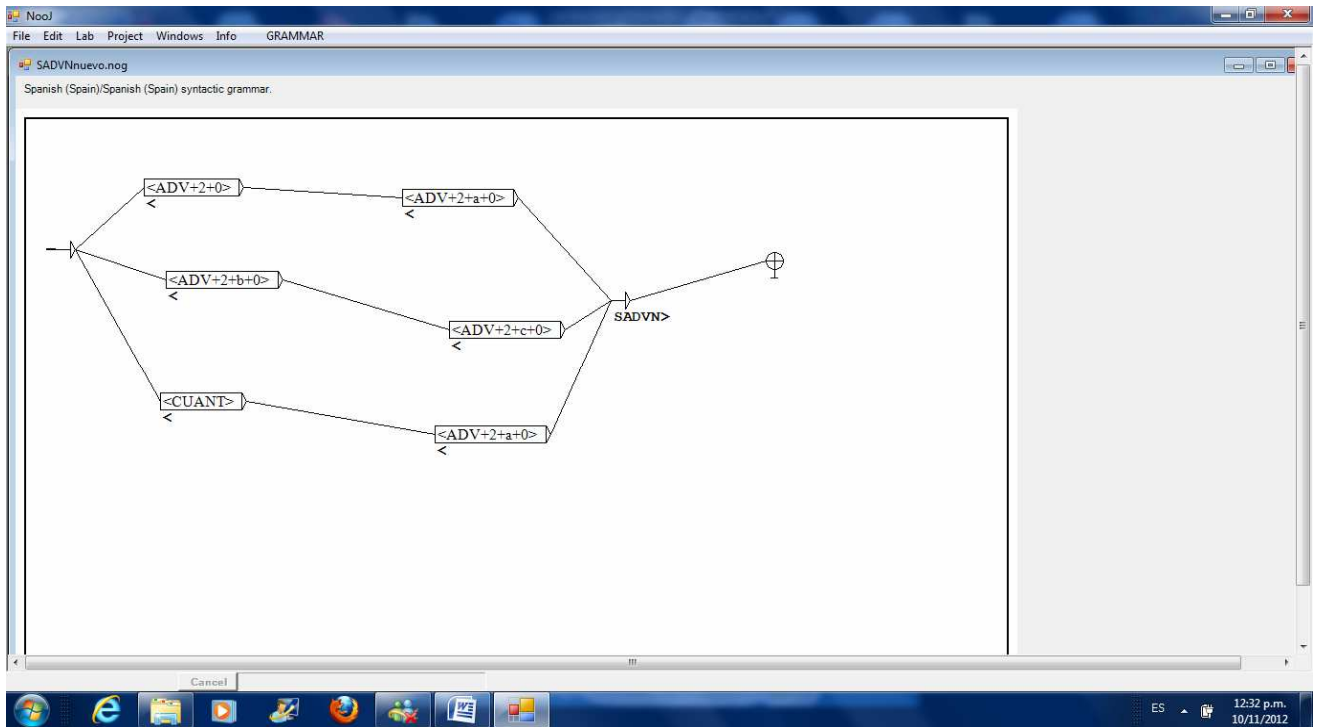
A continuación,

f. la gramática correspondiente al SADV como unidad mayor:

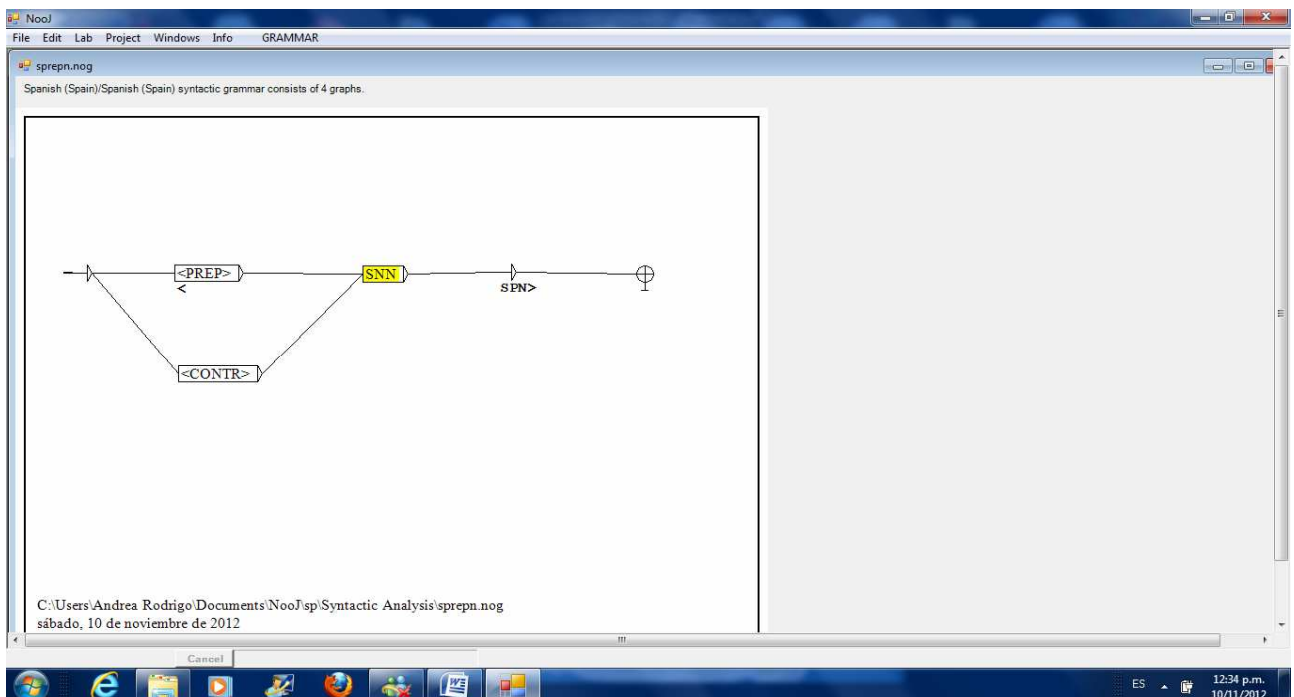


Y finalmente, las gramáticas correspondientes a cada uno de los sintagmas núcleos que lo integran:

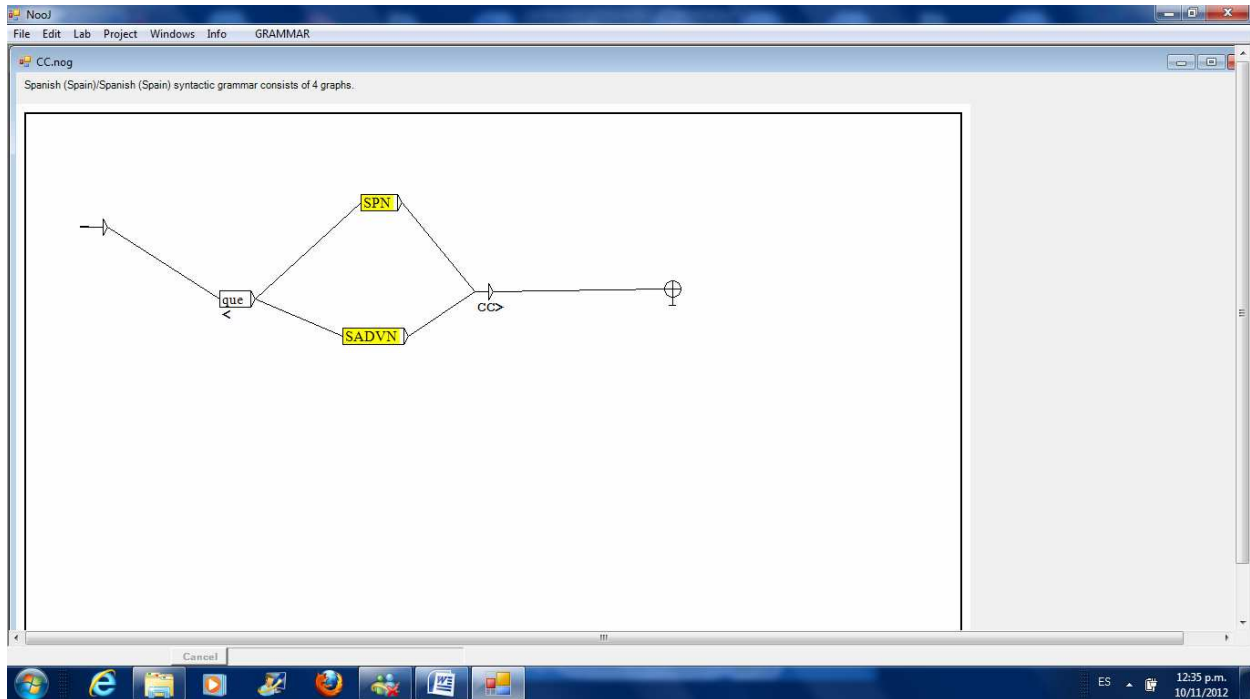
g. El SADVN



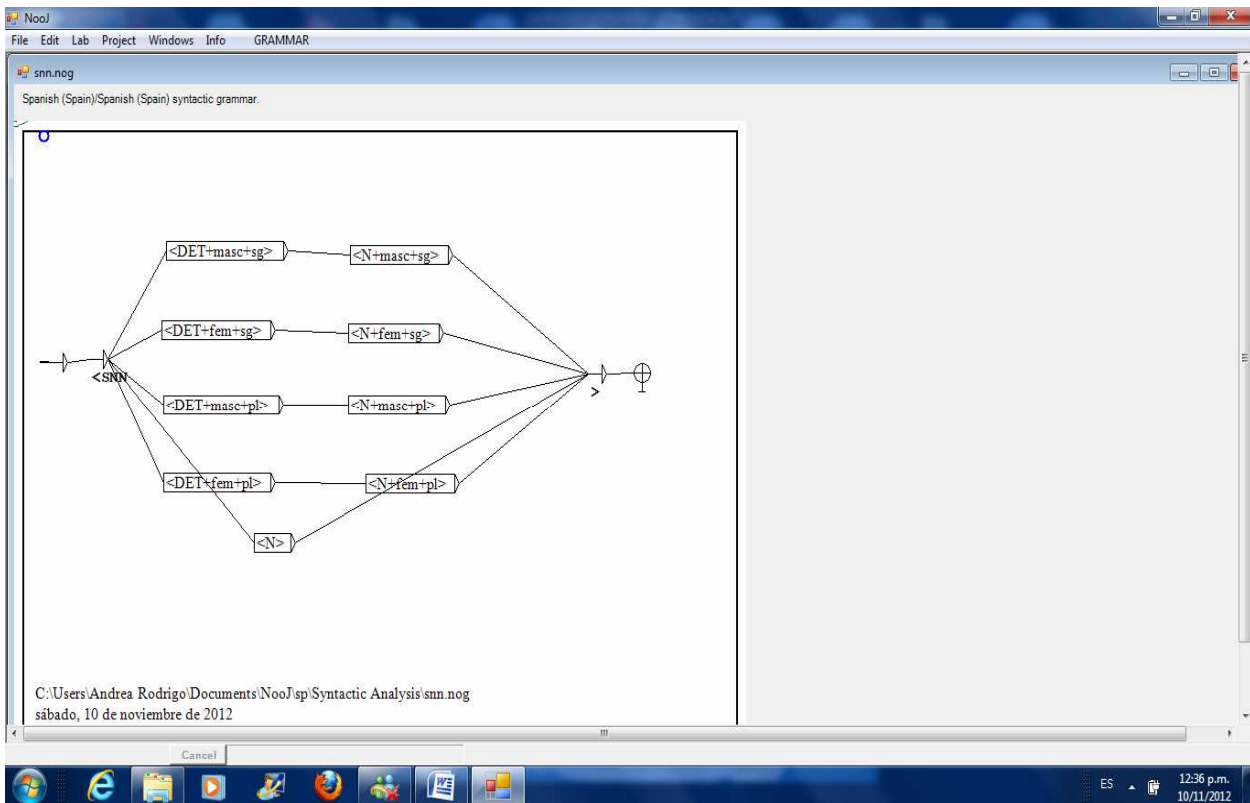
h. El SPN



## 6.8.La CC (construcción comparativa)



- i. El SNN que integra tanto al SPN como a la CC



## 5. APLICACIÓN DE NOOJ EN EL ANÁLISIS DE UN TEXTO

### 5.1. El texto

*Te espero bien enfrente de tu casa, poco después de la cena, pero esta vez no te escondas más abajo de la mesa.*

*Cierto que arriba de la cama encontré las pruebas del delito y que fuiste más lejos que tu hermano...*

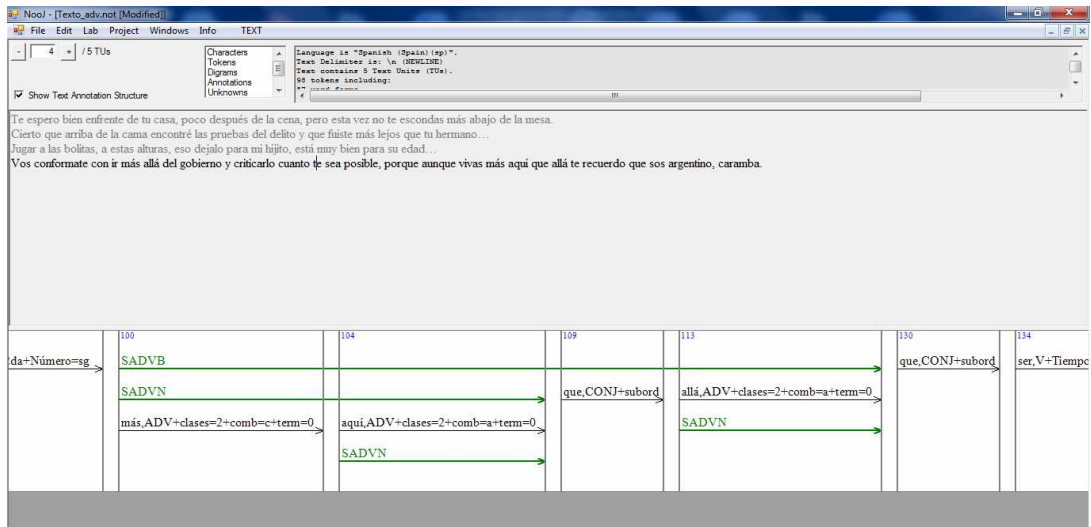
*Jugar a las bolitas, a estas alturas, eso dejalo para mi hijito, está muy bien para su edad... Vos conformate con ir más allá del gobierno y criticarlo cuanto te sea posible, porque aunque vivas más aquí que allá te recuerdo que sos argentino, caramba.*

### 5.2 .El resultado que arroja NooJ

Se observa cómo extrae el SADVA (*poco después de la cena*):

The screenshot shows the NooJ software interface. The main window displays the text: "Te espero bien enfrente de tu casa, poco después de la cena, pero esta vez no te escondas más abajo de la mesa. ... Certo que arriba de la cama encontré las pruebas del delito y que fuiste más lejos que tu hermano... Jugar a las bolitas, a estas alturas, eso dejalo para mi hijito, está muy bien para su edad... Vos conformate con ir más allá del gobierno y criticarlo cuanto te sea posible, porque aunque vivas más aquí que allá te recuerdo que sos argentino, caramba." The bottom panel shows the syntactic tree structure with various annotations. The annotations include: "SADVA", "SADVN", "poco, CUANT", "después, ADV+clases=2+comb=a+term=0", "de, PREP", "SNN", "la, DET+género=fem+número=sg", "cena, N+género=fem+número=sg", "pero, CONJ+coord", and "vez, N+género=fem". The interface also shows a menu bar (File, Edit, Lab, Project, Windows, Info, TEXT) and a status bar at the bottom.

Se observa cómo extrae el SADVB (*más aquí que allá*)<sup>8</sup>:

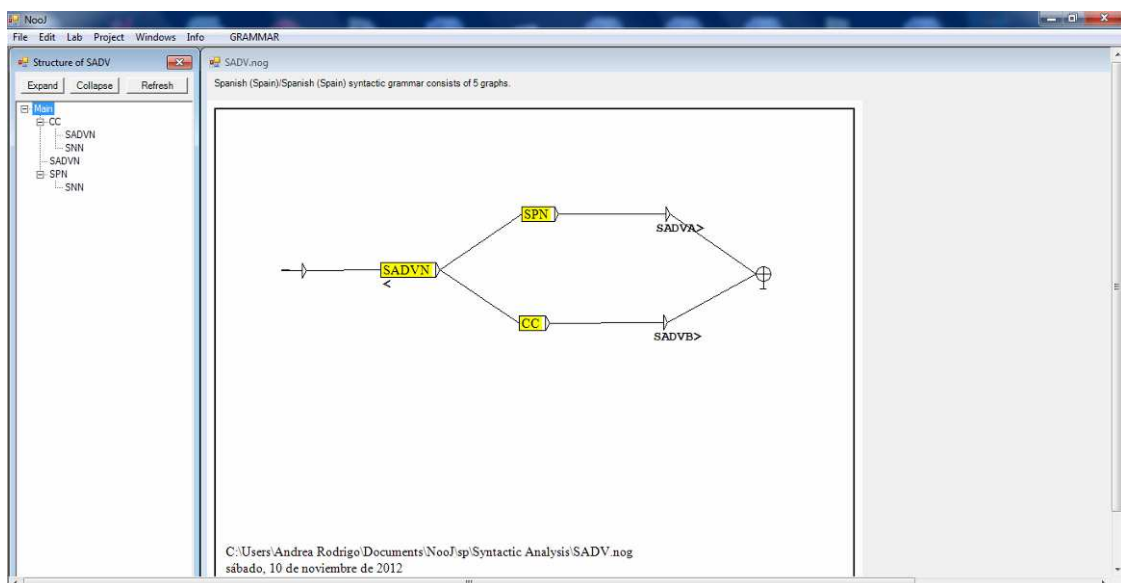


## 6. CONCLUSIONES

Según lo observado en 5.2., la herramienta logra extraer el SADV, aplicando la gramática, es decir, puede verse cómo el SADV aparece conformado por los sintagmas núcleos, según las estructuras planteadas en la Introducción:

- **SADVA** [SADVN + SPN (SNN)]: Ej: *más abajo de la mesa, arriba de la cama, bien enfrente de tu casa, poco después de las diez*<sup>9</sup>, *más allá del gobierno*.<sup>10</sup>
- **SADVB** [En una comparativa: SADVN + complemento comparativo (CONJ+ SNN) (CONJ+SNN)]: Ej: *más lejos que lo esperable, más aquí que allá*

Esto permitió corroborar las hipótesis lingüísticas planteadas en torno al SADV, a la vez que se puede visualizar cómo una estructura entra dentro de la otra:



<sup>8</sup> Se observa la extracción de uno SADV de cada clase, para no hacer tan extensa la exposición

<sup>9</sup> Ejemplo citado en Rev. Infosur N° 5, Extracción del Sintagma Nominal, (Solana, Rodrigo, Méndez), p. 18

<sup>10</sup> Pag. 12, 4/01/04

Se entiende así que en los SADV estudiados el SADVN puede ser seguido por un SPN, para los llamados clase SADVA o por una CC, para los llamados SADVB, según se ve en el rectángulo de la izquierda encabezado por **Structure of SADV**.

## Referencias

- [1] Rodrigo, A. Tratamiento automático de textos, el sintagma adverbial núcleo. Tesis doctoral, Escuela de Posgrado, Facultad de Humanidades y Artes, UNR, Ediciones Juglaría, 2011.
- [2] Silberztein Max, 2003-. NooJ Manual. Available for download at: [www.nooj4nlp.net](http://www.nooj4nlp.net)
- [3] Rodrigo, A. Análisis automático de textos, el sintagma nominal núcleo. Tesis de Maestría, Escuela de Posgrado, Facultad de Humanidades y Artes, UNR, 2006.
- [4] Bès G.G. La phrase verbale noyau en français. En Recherches sur le français parlé, Volume 15, 1999.
- [5] Hagège C. Analyse syntaxique automatique du portugais. Tesis de Doctorado GRIL, Univ. Blaise Pascal, 2000.
- [6] Solana Z., Rodrigo A. El sintagma nominal núcleo. En Desarrollo, implementación y utilización de modelos para el procesamiento automático de textos. Compilado por Víctor Castel, Trabajos de las Segundas Jornadas de Lingüística Informática: Modelización e Ingeniería, Facultad de Filosofía y Letras, Universidad Nacional de Cuyo, 2005.
- [7] Abney, S. Parsing by Chunks. En Berwick et al. Principle-Base Parsing. Kluwer Academic Publishers. Dordrecht, 1991.
- [8] Aït-Mokhtar, S. L'analyse presyntaxique en une seule étape. Tesis de Doctorado, GRIL, Univ. Blaise Pascal Clermont-Ferrand, 1998.
- [9] Bès G.G., Solana Z. Análisis morfológico y gramáticas locales: introducción y una aplicación concreta. En Jornadas Argentinas de Lingüística Informática Modelización e Ingeniería, JALIMI, Rosario, 2004.
- [10] Bonino, R. "Una propuesta para la implantación de la morfología verbal del español en NooJ" en *Revista Infosur*. Nro. 5, [en línea] <<http://www.infosurrevista.com.ar/>>



# **Análisis automático de la gramática temprana<sup>1</sup>**

## **Automatic analysis of early grammar**

**Zulema G. Solana**  
GRUPO INFOSUR  
UNR  
zsolana@arnet.com.ar

### **Abstract**

This work aims to provide a general view about the syntactic and morphological aspects of early grammar and to determine some parameters to implant it into the machine. First, the order in which the categories N (noun), A (adjective), V (verb) and P (preposition) appear in the child and then, the system of morphological and semantic inflectional marks of these categories are determined. The study of prefixes and suffixes which may appear in word formation and of a more specific determination of the evolution stages is left for a later stage in the analysis. Only one sample from three children between the specified ages has been collected. They belong to the CHILDES Data Bank and the resulting generalizations are of a provisional nature.

**Keywords:** early grammar, morphological inflectional marks, semantic marks

### **Resumen**

Este trabajo se propone presentar una descripción general de la gramática temprana en sus aspectos morfológicos y sintácticos y determinar algunos parámetros para implantarla en máquina. Se comienza por establecer el orden en que aparecen en el niño las categorías N(nombre), A(adjetivo), V (verbo) y P (preposición) y el sistema de marcas morfológico-semánticas de flexión de estas categorías. Se deja para más adelante el estudio de la aparición de sufijos y prefijos en la formación de palabras y una determinación más específica de las etapas de evolución. Se ha tomado una muestra de sólo tres niños, en las edades consideradas, pertenecientes al Banco de datos de CHILDES, razón por la cual las generalizaciones tienen carácter provisorio.

**Palabras clave:** gramática temprana, marcas inflexionales morfológicas, marcas semánticas

## **1. INTRODUCCIÓN**

Me propongo presentar una descripción general de la gramática temprana<sup>2</sup> en sus aspectos morfológicos y sintácticos y determinar algunos parámetros para implantarla en máquina. Comenzaré por establecer el orden en que aparecen en el niño las categorías N(nombre), A(adjetivo), V (verbo) y P (preposición) y el sistema de marcas morfológico-semánticas de flexión de estas categorías. Se deja para más adelante el estudio de la aparición de sufijos y prefijos en la

---

<sup>1</sup> Este trabajo pertenece al PID de Ciencia y Técnica 2012 del mismo nombre

<sup>2</sup> Con “gramática temprana” se hace referencia al conocimiento lingüístico de las primeras etapas del desarrollo

formación de palabras y una determinación más específica de las etapas de evolución. Se ha tomado una muestra de sólo tres niños, en las edades consideradas, pertenecientes al Banco de datos de CHILDES, razón por la cual las generalizaciones tienen carácter provisorio.

## **2. ESTADO ACTUAL DE LOS CONOCIMIENTOS SOBRE EL TEMA**

En lo que respecta a las investigaciones sobre la informatización del análisis del lenguaje infantil hay que mencionar en primer lugar a todo lo producido en el marco de CHILDES. La base de datos CHILDES[1] (Child Language Data Exchange System) contiene 44 millones de palabras de 28 lenguas diferentes. La información contenida en el sistema corresponde a 4500 investigadores que trabajan sobre éste y realizan sus contribuciones.

Los datos corresponden a transcripciones de conversaciones de niños con sus madres, padres, investigadores, etc. Todos estos archivos de transcripciones, pertenecientes al sistema CHILDES, se encuentran en el formato CHAT (Codes fr de Human Analysis of Transcripts). El sistema CHAT provee un formato estandarizado para producir transcripciones informatizadas de conversaciones. Estas conversaciones involucran niños y padres, niños, profesionales/investigadores y padres o maestros.

Los objetivos de un sistema informático para datos de este tipo pueden resumirse en:

- a) automatizar el proceso de análisis de los datos
- b) obtener mejores datos en un sistema de transcripciones consistente y completamente documentado
- c) obtener más datos de una mayor cantidad de niños de variadas edades y lenguas

Estos archivos con formato CHAT pueden ser analizados mediante el programa CLAN (Computerized Language ANalysis). El programa fue diseñado para facilitar las investigaciones y análisis sobre las transcripciones en formato CHAT.

En nuestra investigación acudiremos al banco de datos de CHILDES, pero analizaremos los datos con nuestras propias herramientas informáticas.

## **3. METODOLOGÍA Y ETAPAS**

Se trabajará con lenguaje espontáneo dejando para una etapa posterior la elaboración de test o pruebas, es decir, se recurrirá por el momento a la observación dejando para después el uso de métodos experimentales.

La posibilidad de guardar en base de datos informatizadas grandes corpus de lenguaje infantil hace posible el trabajo con el lenguaje espontáneo, en general producto de conversaciones de adultos con el niño o de niños entre sí, metodología (la de la observación del lenguaje espontáneo) que en décadas pasadas había sido desplazada por la experimentación ya que, no mediada por grandes corpus informatizados, la observación era un proceso costoso en tiempo.

Para la implantación en máquina se recurrirá a las herramientas que tiene el grupo INFOSUR [2].

### **3.1. Herramientas informáticas con que se cuenta**

- El software Smorph, analizador y generador, realiza la tokenización y el análisis morfológico, en una sola etapa y da como resultado las formas correspondientes a un lema con sus valores. Lo presentó en su tesis doctoral Salah Aït- Mokthar [3]. La tesis fue desarrollada en la Universidad Blaise-Pascal de Clermont-Fd bajo la dirección de Gabriel G. Bès.

Consideramos que el proyecto INFOSUR, desarrollado en el ámbito de nuestra universidad, es no sólo un antecedente sino la base que nos proporciona los elementos de partida en el aspecto informático y el trabajo de modelización del español. [4]

- MPS trabaja SMORPH en una organización modular. La salida de smorph, en lenguaje Prolog, es la entrada de MPS, también procede de una tesis(cf. Abacci) dirigida por el Dr. Bès en Clermont-Fd.

- xfst (Xerox Finite State Tools): Esta herramienta informática ha sido usada por Xerox Research Centre Europe (XRCE) y Palo Alto Research Centre (PARC.) Es un autómata de estados finitos en el que se ingresan las propiedades lingüísticas en forma de reglas, que pueden ser testeadas en el proceso de generación.

### 3.2. Hipótesis de partida

Respecto del sistema de marcas morfológico-semánticas de las categorías N, V, A, ADV, P, PRON, se plantea la siguiente hipótesis:

Respecto del conocimiento y uso de marcas morfológicas de flexión, los niños pasan por tres momentos:

- categorías sin marcas de flexión: por ejemplo, *nene* aplicado a masculino, femenino, singular y plural, o *come* aplicado a primera o tercera persona singular.
- primeras marcas en algunas categorías: por ejemplo diferenciar persona con *como* y *come*.
- un desarrollo morfológico complejo

### 3.3. Corpus

Para esta etapa de la investigación, el corpus consiste en una muestra [5] procedente de expresiones espontáneas de niños hablantes de español menores de cuatro años del Banco de Datos CHILDES.

La investigación tiene propósitos metodológicos y no va a evaluar precisión y cobertura de la implementación de las herramientas informáticas sino sólo la posibilidad de ser empleadas en adquisición del lenguaje.

### 3.4. Etapas

Consideraremos dos etapas:

#### 3.4.1. Antes de los dos años

##### 3.4.1.1. Categorías morfológicas

Predominan los nombres y las primeras producciones son palabras bisílabas, sin sufijos ni flexivos ni derivativos Ejs: *nene*, *pupa*, *tete*, *pie*, *guaguau*, *agua*, *peño*(por *pelito*), *bota*, *calle*, *silla*, *ponja* (por *esponja*).

Posteriormente aparecen anteceditos en varios casos por determinantes[6]: Ejs: *e nene*, *a calle*, *a silla*, *\*este silla*, *a botas*. Aquí estoy considerando como determinantes propios de esta etapa a (e/o/a/a) que explicitan los rasgos que no están gramaticalizados dentro de N.

Finalmente antes de los dos años aparecen nombres trisílabos. Ejs: *cabeza*, *mañana*, *chaqueta* y se

enriquecen los determinantes, con formas de los indefinidos un/una/unos/unas, propios de la gramática adulta, pero, en algunos casos con variaciones fonéticas: *u' botó'*, *un caballo*

En cuanto a los verbos, se reafirma lo que se sabe por la literatura especializada [7] se encuentran verbos aparentemente en tercera persona singular, pero en realidad sin persona asignada desde el momento en que no constituyen ninguna oposición. Ejs: *está*, *no'sta*, *pincha*, *chupa*, *gusta*, *come*, *quita* luego *hay*, *abe*(por *abre*), *ueve*( por *llueve*), *vio*, *caió* (*cayó*), *pasó*. Quedan para remarcar dos casos: un verbo irregular diptongado en primera y tercera singular: *siento*, *sienta* y *cae* acompañado de *se*, que es el primer clítico registrado *se cae* El hecho se reafirma un mes más tarde con *se acabó*

Al final de esta etapa aparece el clítico de primera persona *me*, se refuerza el *se* y ambos aparecen con verbo en pretérito perfecto. *m'a chupa'o*, *s' ha perdido*, *s' ha loto* .

Adjetivos solamente: *malo*, *bonito* y *caente* (por *caliente*)

### 3.4.1.2. La sintaxis antes de los 2 años

Vamos a caracterizar la sintaxis de la gramática temprana a partir de los siguientes ej:

a.*Se cae e nene*

b.*Papá a calle*

c.*E'nene a botas*

d.*Quita guauguau*

e.*Sienta mamá*

f.*Siento aquí*

g.*Echo agua*

h. *Ota(otra) vez cayó*

i.No apabó

j. *Mía(mira) mama mi pie.*

k. *No hay papú (champú)*

l.*Mamá, m'a chupa'o e' guauguau .*

ll.*S' ha perdido e' pendiente .*

Si se toma en consideración sintagmas, sintagmas núcleos y oración pueden hacerse algunas observaciones de conjunto y otras particulares:

-La mayor parte de las oraciones cuentan entre dos y cuatro palabras.

-Algunos sintagmas están formados por N solo, otros por det+N

-Entre los determinantes, se encuentran posesivos(*mi* en j), indefinidos (*ota* en h), artículos (*e*, *a* en a,b y c)

-Hay SN sujetos y objetos directos.

-Aparición de verbo en primera persona(e), lo que habilita a decir que (f) está en tercera, dado que hay una oposición de persona con el mismo verbo. La primera (a) con un verbo de clítico obligatorio, otro clítico en l (me (m´) es el objeto de “*a chupa'o*”)y en (ll) aparece la primera

expresión de impersonalidad.

### 3.4.2. A los 2 años

#### 3.4.2.1 .Categorías morfológicas

A los 2;02 los Adjetivos. *secos, mojado, mojados, buena, buenísima, guapa, frío (por frío), malo, bonito, pequeño, novo (por nuevo)*

Sufijos derivativos, sólo el diminutivo : *sapatito, bonito a los 2;02.*

#### Sistema verbal María 2;00- 2;03

dej -o v/pres/ind/1a/sg

-as v/pres/ind/2a/sg

-a v/pres/ind/3a/sg

dej -e v/pres/subj/1a/sg

-es v/pres/subj/2a/sg

-e v/pres/subj/3a/sg

#### Clíticos María 2 a 2;03

me	se me
te	te los
le	se le
lo	me lo
se	

#### 3.4.2.2. La sintaxis a los 2 años

##### 2;02

a.No, no te gusta las galletas .

b.Papá, tú que me, tú me que(d)as conmigo, a qué sí ?

c.Te vo a mojar, a Papá .

d.Venió un bolo feroz .

e.Estaba una niña guapa .

f.Estaba una niña gua:pa que se llama Maniña y viene un bolo feroz

Hay falta de concordancia en (a), en la persona del clítico (*me* en lugar de *te*) en (b), “*veníó*” por “*vino*” en (d). Aparece una relativa y una coordinación en (f)

##### 2;04

a.Espera que voy a coger un muñeco .

b.Hacemos una película al Aito .

c.Si no jubamos la tiro la tapa .

d. *Yo sabo hasé muchas casas*

e. *Ahora voy a hasé **ot'a casa más bonita***

Puede observarse que se va aumentando el número de palabras por oración, que hay subordinadas, entre ellas una condicional en (c). Crece la variedad de determinantes (*muchas casas, ota casa*) y continúa la conjugación regular donde en la lengua estándar es irregular (*sabo*). Llama particularmente la atención un SN como ***ot'a casa más bonita*** (indef + N + SAdj (adv + adj)). En esta etapa se observa la concordancia correspondiente a la lengua adulta.

### 3.4.3. Síntesis de la evolución

	Antes 2 años	2 a 2;6
morfología N	ausencia flexión y derivación	flexión plural concordancia género y número en el SN
Determinantes	artículos ( <i>e/a/o/a</i> ) indefinidos ( <i>ot(r)o/a, u(n)</i> ) posesivos ( <i>mi</i> )	artículos ( <i>la, las</i> )
Adjetivos	<i>malo, bonito, ca(li)ente</i>	<i>mojado, mojados, buena, buenísima, guapa, fío (por frío), pequeño, novo(por nuevo)</i>
morfología V	3a-pers.sg.presMI : <i>está, no 'sta, pincha, chupa, gusta, come, quita, hay, abe(por abre), ueve( por llueve),sienta.</i> 3a-prts.sg.presMI <i>vio, caió (cayó), pasó.</i> 1a-pers.sg: <i>siento</i> 3a-pers.sg.pret.perf.MI. <i>m'a chupa'o, s' ha perdido, s' ha loto</i>	1a-2a-3a.pers-sg-MS 1a.pers.fut.MI : <i>voy a + infinitivo</i>
Clíticos	<i>me/se</i>	<i>te/le/lo/se me/ se le/te lo/me lo</i>
compl verbo	OD	
Prep	-	<i>a, con, para</i>

## 4. IMPLANTACIÓN EN MÁQUINA

### Implantación de la morfología del español estándar

El software Smorph, mencionado en 3.1., 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 un lema (o a un subconjunto de lemas) con los valores correspondientes. Es una herramienta declarativa, la información utilizada por Smorph está separada de la maquinaria algorítmica, esto hace que se la pueda adaptar al uso que quiera darse, de modo tal que con el mismo software se puede tratar cualquier lengua siempre y cuando se modifique la información lingüística declarada en sus archivos.

Se ha trabajado con información pertinente para el francés, para el portugués y para el español [6]. En todos estos casos se trata de la lengua estándar, en este trabajo daremos cuenta de la investigación que estamos realizando respecto de las producciones de la gramática temprana. A continuación explicaremos como se implementa en general y luego las modificaciones realizadas para que se puedan analizar las producciones infantiles.

Esta herramienta compila, minimiza y compacta la información lingüística de modo que quede disponible en un archivo binario. Los códigos fuente se dividen en cinco archivos:

- Códigos Ascii
- Rasgos
- Terminaciones
- Modelos
- Entradas

En el archivo **entradas**, se ingresan los ítemes léxicos acompañados por un indicador del modelo correspondiente. Este indicador de modelo oficia de enlace con el archivo **modelos**, donde se especifica la información morfológica, género y número y las terminaciones que se requieren en cada ítem.

En el archivo **modelos**, se introduce la información correspondiente a los modelos de flexiones morfológicas. Un modelo de flexión agrupa todas las flexiones de una misma clase de palabras. Esto se describe asociando a un conjunto de terminaciones 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 se declara la cantidad de caracteres que hay que sustraer a la forma lematizada. Este valor debe ser una cifra entre 0 y 9 y estar precedida del signo "-". En tercer lugar se declara la terminación, que debe estar declarada previamente en el archivo de **terminaciones**. La declaración morfológica corresponde a una cadena de caracteres sin espacios en blanco.

En el archivo de **terminaciones** es necesario declarar todas las terminaciones que son necesarias

para definir los modelos de flexión. Si en la definición de un modelo se especifica una terminación no declarada en este archivo, el programa emite un mensaje de error. Las terminaciones se declaran una a continuación de otra, separadas por un punto. Es posible declarar una terminación vacía mediante el carácter "@" y una terminación distinguida asociando a una terminación la definición morfológica correspondiente.

Para construir los modelos se recurre a rasgos morfológico- sintácticos (categoría, género, número, etc). En el archivo de **rasgos**, se organizan jerárquicamente las etiquetas, por ejemplo, nombre, adjetivo, etc. Asimismo, se puede incorporar la etiqueta que indica, por ejemplo, el tipo de nombre y se adicionan los rasgos de concordancia, género y número:

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 archivos descriptos anteriormente.

A continuación se presentará la implementación de los archivos nombrados, para el análisis de ejemplos pertenecientes al español estándar y luego **los cambios necesarios para su adaptación de modo de poder realizar el análisis de expresiones de la gramática temprana.**

Hemos encontrado en la gramática de los 2 años el siguiente paradigma para los artículos:

e/a/o/a

En el español estándar tenemos:

#### Entradas

el	el	/det/art .
el	los	/det/art .
el	la	/det/art .
el	las	/det/art .
el	lo	/det/art .

En cada fila, primero el lema (elegimos convencionalmente "el"), luego la ocurrencia lingüística y después los rasgos ("det": determinante, "art": artículo).

Para poder analizar las ocurrencias lingüísticas infantiles, introducimos:

el	e	/dett/artt .
el	o	/dett/artt .
el	a	/dett/artt .
el	a	/dett/artt .

Equivale "dett":determinante gramática temprana y "artt":artículo gramática temprana.Además de modificar el rasgo en "entradas" tenemos que agregarlo en el archivo "rasgos".

Así, si analizamos "e nene"(el nene), obtenemos:

'e'.

[ 'el', 'EMS','dett', 'TDETT','artt'].

'nene'.

[ 'nene', 'EMS','nom', 'GEN','masc', 'NUM','sg'].



-cuestiones fonológico- fonéticas que llevan a duplicar las “entradas”

Ejs: ota/otra, frío/frío, apabó/acabó, jubamos/jugamos

-cuestiones morfológicas que llevan a modificar los “modelos”

Ejs: sabo/sé, venió/vino

Aquí se trata de la conjugación de verbos irregulares como si fueran regulares.

“venir” como regular sigue el modelo 3 ej.”partir” (vení/veniste/veníó)

@v3 -2

+í v/prets/ind/1a/sg/c3/r

+iste v/prets/ind/2a/sg/c3/r

+ió v/prets/ind/3a/sg/c3/r

(rasgos: V(verbo), prets(pretérito simple), ind(indicativo), 3ª(tercera), sg(singular), c3(3ª conjugación), r(regular))

Para que este modelo, propio de “partir/partió”, se adapte a venir/veníó, debe convertirse en un nuevo modelo, al que asignamos otro número:

@v4 -2

+í vt/prets/ind/1a/sg/c3/r

+iste vt/prets/ind/2a/sg/c3/r

+ió vt/prets/ind/3a/sg/c3/r

Cambiamos el rasgo “v” por “vt” (verbo gramática temprana) y, en consecuencia en el archivo “rasgos” hay que agregar el nuevo rasgo.

## A MODO DE CONCLUSIÓN

Respecto de la evolución sintetizada en 3.4.3 puede decirse que en el SN se parte de una forma invariable que pronto toma variaciones de género y número, lo que permitirá el proceso de concordancia. Cuando todavía N no flexiona ya presenta un determinante (e/a/o/a) que anuncia su género y número en una especie de morfología flexional pre-posicionada. Los adjetivos avanzan en un sentido cuantitativo y cualitativo, en la segunda etapa se triplican y aparecen los participios adjetivales.

El verbo también parte de una forma invariable hasta lograr primero oposición de persona y en segundo término de tiempo. Se establece el sistema de clíticos que incide en las posibilidades sintácticas de la lengua.

En la propuesta de implementación en máquina hemos duplicado entradas por razones fonológico-fonéticas y hemos modificado modelos para dar cuenta del tratamiento de verbos irregulares como si fueran regulares. Esta reestructuración del sistema lleva a agregar nuevos rasgos. El hecho de contar con una investigación e implantación en máquina de la morfología del español estándar nos permite la implementación presentada para la gramática temprana. Sólo cuando ampliemos la muestra estaremos en condiciones de evaluar la precisión y cobertura de lo realizado.

## Referencias

- [1] Carrasco González, M. y Celis Sánchez, C. (2004): CHILDES Project: Child Language Data Exchange System. Sistema de Transcripción CHAT. Publicación electrónica disponible en <http://childes.psy.cmu.edu/intro/spanish.pdf>
- [2] Nuestro equipo de investigación, INFOSUR, en una trabajo que ha desarrollado entre 2005 y 2008, ha realizado la implantación en máquina mediante el software SMORPH de 4000 verbos del español, 6000 sustantivos, 3000 adjetivos, preposiciones, pronombres, etc.  
Hasta el momento ha logrado los siguientes resultados:
  - a) Delimitación de límites de oraciones (parte de la tesis de Doctorado 2008 de Celina Beltrán bajo la dirección de Gabriel G. Bès)
  - b) Descripción, formalización y análisis automático del sintagma nominal núcleo (tesis de Maestría 2006, de Andrea Rodrigo, dirigida por Zulema Solana)
  - c) Descripción, formalización y análisis automático del sintagma verbal núcleo (trabajos de Gabriel Bès y Zulema Solana y tesis doctoral 2010 de Rodolfo Bonino.
  - d) Descripción, formalización y análisis automático del sintagma adverbial núcleo (tesis doctoral de Andrea Rodrigo)
  - e) Análisis automático de la puntuación en español (tesis doctoral de Walter Koza)
- [3] Aït Mokhtar, Salah 1998 L'analyse présyntaxique en une seule etape tesis doctoral dirigida por Gabriel G. Bès en el GRIL Université Blaise-Pascal, Francia. Aït-Mokhtar, Salah y Rodrigo Mateos, José Lázaro. 1995 Segmentación y análisis morfológico de textos en español utilizando el sistema SMORPH. SEPLN, 17, 29-41.
- [4] Solana, Zulema, Bonino Rodolfo y Valenti, Viviana 2005 Modelización de las fuentes declarativas en una herramienta de análisis y conjugación automáticos de verbos del español en Desarrollo, implementación y uso de modelos para el procesamiento automático de textos (ed. Víctor Castel) Facultad de Filosofía y Letras, UNCUIYO
- [5] La muestra recoge expresiones de María, Irene y Jakshon (banco de datos CHILDES)
- [6] Mariscal, S. 2008 "Early acquisition of gender agreement in the Spanish noun phrase: starting small" en J. Child Lang. 35, 1-29. Cambridge University Press
- [7] López Ornat, S.1995 Adquisición de la sintaxis española, Siglo XXI,

## **Análisis automático de la interlengua de aprendientes de español: la presencia de determinantes indefinidos en el sintagma nominal núcleo.**

### **Automatic analysis of the interlanguage of Spanish learners: the presence of indefinite determiners in the head noun syntagm.**

**Carolina Paola Tramallino**

CONICET, UNR

Rosario, Argentina

carotramallino@hotmail.com

### **Abstract**

The presence or absence of determiners in Spanish grammar is a matter of considerable interest not only at the syntactic level but also at the semantic one as these can increase or decrease reference (Leonetti: 1999). This becomes even more evident in the interlanguage constructions produced by Spanish learners.

The aim of this work is to provide a description and achieve the automatization of head noun syntagms initiated by determiners which are found in the interlanguage of Spanish learners whose source languages are English, French, German and Portuguese. To these ends, our object of analysis and discussion are source language constructions matching those of the target language as well as the constructions derived from them.

First, a linguistic description is made about head noun syntagms in the standard language; following are described and analyzed the interlanguage own structures and, lastly, both constructions are recognized automatically. For the automatic analysis, we use the computer-based tool MPS implanted by Faiza Abbaci, who has been working in a modular way with Smorph. In this case, clustering and recomposition rules will have to be created for head noun syntagms.

**Keywords:** automatic analysis, indefinite determiners, head noun syntagm, interlanguage, Spanish learners.

### **Resumen**

En la gramática del español, la presencia o ausencia de determinantes interesa no sólo a nivel sintáctico sino también semántico, ya que reduce o amplía la referencia (Leonetti: 1999). Esto se hace aún más evidente en las construcciones de interlengua producidas por aprendientes de español.

El objetivo del presente trabajo es realizar una descripción y lograr la automatización de los sintagmas nominales núcleos iniciados por determinantes que se hallan en la interlengua de

aprendientes de español cuyas lenguas de origen son el inglés, francés, alemán y portugués entre otras, analizando tanto las construcciones coincidentes con la lengua meta, así como también las que se desvían de la misma.

Para esto, en primer lugar se realiza una descripción lingüística de los snn en la lengua estándar, luego se describen y analizan las estructuras propias de interlengua y por último se reconocen automáticamente ambas construcciones. Para el análisis automático se utiliza la herramienta informática MPS implantada por Faiza Abbaci, que trabaja de manera modular con el software Smorph. En este caso, habrá que crear reglas de recomposición y agrupamiento para los snn.

**Palabras claves:** análisis automático, determinantes indefinidos, sintagma nominal núcleo, interlengua, aprendientes de español.

## 1. INTRODUCCION

El objetivo del presente trabajo será realizar la descripción, la modelización y la implantación en máquina de ciertos sintagmas nominales núcleos divergentes de los de la lengua estándar que se hallan en la interlengua de aprendientes de español como segunda lengua.

El corpus está conformado por cincuenta textos escritos, pertenecientes a estudiantes de un nivel inicial que poseen variadas lenguas de origen como el holandés, francés, italiano, portugués, inglés y alemán. La investigación, enmarcada en la lingüística computacional, tiene como objetivo lograr el análisis automático a partir de herramientas informáticas, no sólo de las construcciones estándares sino también de las estructuras propias que presenta el corpus de interlengua.

Para ello, en primer lugar, se explicará a qué llamamos determinantes, qué valores poseen, en qué tipo de construcciones pueden aparecer posesivos, demostrativos e indefinidos y si permiten la presencia o no de otro determinante.

A continuación, se mostrará brevemente el funcionamiento de Smorph [1] y de MPS [2], que son las herramientas que se utilizarán para el análisis automático y por último, se expondrá la implantación en máquina y las modificaciones pertinentes que deberán ejecutarse en los archivos que componen el software Smorph, para el reconocimiento de las construcciones desviadas. Por último, una vez obtenidos los resultados del análisis morfológico y sintáctico realizado por los softwares mencionados, se presentarán algunas consideraciones finales a modo de cierre.

## 2. EL SINTAGMA NOMINAL NÚCLEO

El sintagma núcleo, tal como define Abney (1991) [3] es una secuencia de categorías que se encuentran en un orden, por lo tanto las combinaciones entre ellas son limitadas.

Este segmento se inicia con la primera categoría del sintagma y finaliza en su núcleo, por lo tanto, el sintagma nominal núcleo (snn) es una construcción que va desde el comienzo hasta el núcleo del sintagma nominal. (Solana-Rodrigo: 2005) [4]

En este trabajo se considerarán los snn presentes en la interlengua deteniéndose en la clase de determinantes que modifican al núcleo.

### 3. LOS DETERMINANTES EN EL ESPAÑOL ESTÁNDAR

Los determinantes especifican la referencia del núcleo del SN. Dentro de los determinantes se identifican por un lado a los artículos, que permiten delimitar la denotación del grupo nominal que componen e informar de su referencia y por otro, a los posesivos y demostrativos. Además se agrupan como tales a los numerales cardinales y a los indefinidos, para los cuales se propondrá una clasificación.

Tabla 1: Clasificación propuesta para determinantes

<b>DETERMINANTES</b>
Artículos (la/las/lo/los/el)
Posesivos (mi/tu/su...)
Demostrativos (este/aquel...)
Numerales cardinales (dos/tres/noventa...)
Indefinidos (algún, un, otros, pocos, todos, ...)

Tanto el artículo, como los demostrativos y posesivos pueden preceder a los numerales (*Mis dos hijos/ Estos tres alumnos*). Con respecto a los indefinidos, debemos distinguir dentro de estos, a pequeños subgrupos para poder establecer las relaciones de combinación. En el siguiente apartado se mencionan algunas clasificaciones que han sido consignadas para el tratamiento de los mismos y a continuación se expondrá la propia, atendiendo a las posibilidades de combinación y exclusión de los indefinidos dentro del snn, respecto de los otros determinantes.

#### 3.1. Clasificaciones realizadas

Los determinantes, también llamados cuantificadores existenciales, poseen características propias que han llevado a numerosos estudios. Bosque – Gutiérrez Rexach (2008) [5] retoman las observaciones de Milsark (1974, 1977) sobre el tema, quien distingue entre determinantes fuertes (distributivos, demostrativos y posesivos) y determinantes débiles, y advierte que, mientras los primeros no pueden aparecer en las construcciones existenciales, los últimos sí pueden hacerlo: (*\*Hay ese estudiante en el jardín. Hay muchos libros sobre la mesa*)

Según Milkar, los determinantes débiles (*algún (os), numerales cardinales, muchos, pocos, varios, etc.*) no son cuantificadores, sino marcadores de cardinalidad. Por lo tanto, al estar dichas oraciones cuantificadas existencialmente de forma inherente, no pueden aparecer cuantificadores pero sí marcadores de cardinalidad. Si hubiera cuantificadores en la posición posverbal de esas oraciones, reflexiona Milkar, tendríamos dos cuantificadores para el SN, lo cual resultaría anómalo, por ejemplo: *\*Algunos todos los niños*.

La Real Academia Española (2010) [6] distingue entre cuantificadores fuertes y débiles. Dentro de estos últimos (también designados como indefinidos) realiza una subclasificación: existenciales, numerales cardinales, evaluativos, comparativos y de indistinción o de elección libre.

Con respecto a la posibilidad de combinación, aclara que los cuantificadores universales, los existenciales y los de indistinción no aparecen luego de los artículos, demostrativos y posesivos

(\**estos todos*, \**los algunos*, \**este nadie*), a excepción del indefinido existencial alguno que puede combinarse con sustantivos no contables (algún líquido) y de la fórmula: el poco o ningún + sustantivo.

### 3.2. Los determinantes indefinidos

El determinante indefinido se utiliza para señalar que lo designado por el grupo nominal no es identificable por el oyente: Un hombre entró por la puerta, y si la construcción está en singular, recibe interpretación genérica: *Una madre siempre entiende a su hijo*.

No obstante, diversos trabajos complejizan la alternancia de los artículos definido/indefinido en función de la (in)especificidad semántico-pragmática del referente. Cf. Leonetti, M. (1999) [7], Alcina Caudet (1999). [8]

El término *otro* cuenta con propiedades que lo asemejan tanto a los adjetivos, como a los determinantes y a los cuantificadores. Acepta complementos partitivos y se acerca a los adjetivos por el hecho de que puede ir precedido de determinante, por ejemplo: [*las otras cuestiones*], sin embargo no puede ir precedido del indefinido: la combinación \*un otro por ejemplo, se rechaza. Con respecto al significado, podemos atribuirle dos valores semánticos a otro: a) alteridad: Juan se mudó a otra casa y b) aditiva: Juan editó otro disco. (Egurén-Sánchez: 2003 [9])

### 3.3. Los posesivos y demostrativos pronominales

Los posesivos y demostrativos se encuentran en distribución complementaria respecto de los otros determinantes en el español contemporáneo. Por lo tanto, puede decirse: el, este, mi, algún, ningún/problema, pero no \*el mi problema o \* algún su problema. Excepto si se combinan con los numerales (*eran tres mis/los/aquellos niños*) y con algunos indefinidos (*sus pocas pertenencias/estos otros jurados*).

Es decir, la presencia de un posesivo o demostrativo en el snn excluye a los artículos y a ciertos indefinidos a la vez que permite la presencia de numerales y de algunos indefinidos. Por lo tanto, para analizar automáticamente tanto a los snn presentes en la interlengua que coinciden con las restricciones expuestas, como a los que no las evidencian, es necesario formular una clasificación para los indefinidos que contemple las particularidades observadas respecto de sus posibilidades de combinación y exclusión con los otros determinantes.

## 4. CLASIFICACIÓN PROPUESTA PARA DETERMINANTES INDEFINIDOS

Para las restricciones del español señaladas en el apartado anterior y con el objetivo de poder analizar automáticamente los snn presentes en la interlengua, propongo una clasificación que agrupe a los indefinidos según las relaciones de combinación y exclusión con respecto a los otros determinantes dentro del snn. Para eso los divido en tres grupos: indefinidos 1, 2 y 3 y a los dos primeros los subclasifico a su vez en a y b.

Indefinidos 1a: ningún – ninguna/ algún – alguna/s- alguno/s.

Excluyen a artículos, demostrativos y posesivos (\*Las algunas cuestiones/ \*Mi algún trabajo/ \*Estos algunos problemas)

Preceden a los indefinidos 2 (Algunos pocos profesores/ Ningún otro juicio)

Indefinidos 1b: Un- una- unas- unos

Excluyen a artículos, demostrativos y posesivos (\*Un mi computadora/ \*Unos aquellos objetos/ \*Una la posibilidad)

Excluyen a indefinidos 2b (\*Unas otras personas)

Indefinidos 2a: Varios/as, poco/s- poca/s, mucho/s -muchas/s, demasiado/s- demasiada/s- bastante/s.

Admiten a artículos, demostrativos y posesivos y a indef1a y 1b siempre en posición posterior (Los pocos turistas/ los bastantes conflictos/ Estos muchos televidentes/ Sus varios pedidos).

Es el único tipo de indefinido que puede combinarse entre sí en ciertas ocasiones (bastantes pocos indicios/ demasiados pocos consumidores).

Indefinidos 2b: Otro/a- otros/as

Se combinan con indef1a e indef2a (Ningún otro llamado/ alguna otra persona/ otras tantas palabras)

Preceden a indef2a (Otros muchos jugadores/ otras tantas personas) pero no pueden preceder a indef1a (\*Otras algunas tareas/ \*otro ningún estudio).

Excluyen a indef 1b (\*unas otras docentes/ \*otro un deseo).

Indefinidos 3: todo/a- todos/as

Preceden a artículos, posesivos, demostrativos, nombres (Todos los días /Todos mis parientes/ Todas estas monedas/Todas canciones hermosas / toda canción)

Solo en singular se permite la combinación: el todo pero con un sentido de generalidad que ubica a todo como núcleo del sintagma y excluye a los nombres: (\*el todo dilema)

Excluyen a indef1a (\*todo algún empleo/\* toda ninguna situación)

Excluyen a indef2a y b (\*varios todos/ \*todos muchos/ \*todos otros anuncios / \*otras todas maestras) a excepción de intercalar otro determinante como artículo, demostrativo y pronombre entre los dos indefinidos (todos los otros avisos).

Si siguen a los indefinidos 1b, estos últimos no actúan como determinantes sino como pronombres (una toda roja).

No se incluyen a indefinidos del tipo: sendos/as, ciertos/as, tantos/as, ambos/as, cada por las escasas posibilidades de combinación con artículos, posesivos, demostrativos, numerales e indefinidos.

En el siguiente cuadro se visualizan las relaciones de exclusiones y combinaciones entre artículos, demostrativos, posesivos y nombres y los distintos tipos de indefinidos propuestos: 1a, 1b, 2a, 2b y 3. Además se indica si estos pueden o no combinarse entre sí. El signo – indica que la presencia de uno excluye al otro, mientras que el signo + indica que pueden combinarse, es decir, hallarse juntos en un snn. No se especifican las relaciones de orden, es decir si inician la construcción o se encuentran pospuestos. Los espacios en blanco significan que esa relación ya fue efectuada en el cuadro.

Tabla 2: Cuadro de exclusiones y combinaciones

	Art, pos, dem, nom	Indef1a	Indef1b	Indef2a	Indef2b	Indef3
Indef1a (algunos)	–	–	–	+	+	–
Indef1b (unos)	–		–	+	–	+
Indef2a (pocos)	+			+		–
Indef2b (otros)	+			+	–	–
Indef3 (todos)	+					–

(\*1) un todo: en esta combinación todo funciona como pronombre y no como determinante.

Existen ciertas expresiones estilísticas como: “Carlos es todo un hombre” o “...estudió toda una noche”. En el primer caso todo equivale a completo, en el segundo caso toda equivale a entera y marca el fin del evento denotado por el verbo.

(\*2) Puede decirse: bastantes pocas personas asistieron al evento.

## 5. ANÁLISIS DE LA INTERLENGUA

Encontramos construcciones encabezadas por el cuantificador todos, por ejemplo: [Todos los días], [todos mis amigos], el empleo de cada con un valor partitivo [cada noche...] numerales [un pasajero, tres remeras, dos pantalones, cuatro hermanos]; determinantes posesivos: [este edificio], [mis vacaciones], [sus muchos negocios], [su casa], [ese correo]; artículos indefinidos: [Unos cuarenta y cinco minutos] (valor aproximativo), [una de las chicas] (valor partitivo), [pocos lugares], [un poco morena] (evaluativo), [tantos kilómetros], [algunas horas], [...cualquier manera], [todos los colores], [todos mis amigos], [otro hotel].

snn coincidentes con la lengua estándar:

- “Todos los días voy a mi escuela”
- “¿Te gusta este país?”
- “Tenía cuatro hermanos...”
- “Sobre cada uno de los discos de masa...”
- “... tres remeras están deformadas...”
- “... los dos amantes planearon huir...”
- “... durante unos cuarenta y cinco minutos...”
- “... de cualquier manera.”



- i. "... en el otro lado hay empresas..."
- j. "Cada noche me gustaba más..."
- k. "...estaba escondido en su casa..."
- l. "¿Te gusta este país?"
- m. "Una cierta emoción..."
- n. "... su matrimonio no funcionó..."
- ñ. "... dejalo varios días..."
- o. "... he aprendido español en un curso..."
- o. "... en el otro lado hay empresas..."
- p. "...se reúnen algunas condiciones climáticas ideales..."
- q. "... sus muchos negocios me gustaron"
- r. "Queremos otro hotel..."
- s. "...hicimos tantos kilómetros"

### **5.1. Snn propios de interlengua**

En el corpus de interlengua hallamos estructuras que presentan diferencias con respecto a los sintagmas nominales núcleo del español estándar. Los agrupamos en cuatro casos:

#### **1. Ausencia de artículos en construcciones encabezadas por todos:**

- a. "Recibiré todas fotos..." (inglés)
- b. "Me gusta mucho todas días en Rosario." (francés)
- c. "...son diferentes en todos aspectos" (portugués)

#### **2. Artículo seguido de pronombre posesivo más sustantivo común:**

- a. "Es (...) mui linda tanto cuanto la suya ciudad Rosario."
- b. "...pero la mía ciudad natal..." (portugués)

#### **3. Artículo seguido de posesivo más sustantivo común:**

- a. "Los compañeros de la mi nueva clase son muy simpáticos..." (italiano)
- b. "... nos estaba esperando Siao, el nuestro perro negro." (italiano)
- c. "Yo soy mucho grato al pueblo de Rosario por la su hospitalidad" (portugués)

#### **4. Indefinido seguido de la palabra otro más sustantivo común:**

- a. "...lo deja para una otra persona..." (francés)

Por una cuestión de productividad solo me ocuparé en esta ocasión de los dos últimos casos.

#### 5.1.1 Construcciones encabezadas por artículo seguidas de determinantes posesivos

Este tipo de construcción, encontrada solo en dos lenguas se debe al hecho de que tanto en italiano como en portugués, el posesivo puede aparecer precedido por artículos y por indefinidos.

#### 5.1.2 Construcciones encabezadas por indefinidos y seguidas por el término otro

Se observan en el corpus construcciones encabezadas por un determinante indefinido y seguidas por la palabra otro que en la mayoría de los casos concuerda en género y número con el sustantivo que le sigue. A continuación, se dividen los casos encontrados según el valor que toma otro, mencionado en el apartado 3.2.

Tabla 3: indefinido + otro

leng	Valor de alteridad	Valor aditivo
ALEMÁN	(6) a. “El hospital quería encontrar una otra familia para María...”	
INGLÉS	(6) b. “Una otra característica es el nivel de educación...”	(7) a. “Espero volver una otra día...”
ITALIANO	(6) c. “Un otro lugar que no puedes perderte es la capital: Roma...”	(7) b. “...me gustaría volver allá una otra vez!”
FRANCÉS	(6) d. “...lo deja para una otra persona...”	(7) c. “Deseo ir un otra vez en los próximos años.”
HOLANDÉS	(6) e. “Una otra solución es transformar rutas...”	(7) d. “Su padre tenía contactos con Mersenne, un otro matemático...”

## 6. ANÁLISIS AUTOMÁTICO

Para el análisis automático emplearé el software Smorp, que es un analizador y generador morfosintáctico, desarrollado en el GRIL por Salah Aït-Mokhtar bajo la dirección de Gabriel Bès y el Módulo Post Smorph (MPS), ya que ambos constituyen herramientas declarativas que se encuentran separadas de la maquinaria algorítmica y por lo tanto permiten al lingüista trabajar con cualquier lengua y, además, realizar adaptaciones de acuerdo a su propósito de investigación.

Smorph realiza la tokenización y el análisis morfológico en una sola etapa, produciendo las formas correspondientes a un lema con los valores adecuados, es decir, brindando información morfológica y categorial.

MPS trabaja sobre la salida de los datos de Smorph. Este programa es un tokenizador de estado finito que segmenta textos sobre la base de la información semántica recibida y realiza un análisis morfosintáctico, reconociendo sintagmas y etiquetándolos.

Su funcionamiento permite adecuar la herramienta para el tratamiento automático de snn propios de la interlengua descripta.

MPS ejecuta dos funciones principales que son la recomposición y la correspondencia. Dentro de la recomposición trabajaré con el tipo de regla de reagrupamiento a la que me referiré más adelante.

## 6.1. Implementación en Smorph

Se emplean los siguientes procedimientos:

### a) Modificaciones en el archivo Rasgos

Debe crearse la etiqueta morfosintáctica que corresponda a determinantes indefinidos y dentro de ésta ubicar las distintas clases.

TDET	tipdet
art	artic
dem	demos
pos	poses
num	numer
indf	indef .

TINDF	tipindf
indf1	indef1a
	indef1b
indf2	indef2 a
	indef2b
indf3	indef3 .

### b) Modificaciones en el archivo Entradas

Se declaran en las entradas los indefinidos con el rasgo que les corresponde, de la siguiente manera:

algún	/indef1a .
un	/indef1b .
pocos	/ indef2a .
otros	/indef2b .
todos	/indef3 .

## 6.2. Análisis morfológico a partir de Smorph de textos de aprendientes de español

*“Una otra solución es transformar rutas...”*

*“Deseo ir un otra vez en los próximos años.”*

*“Los compañeros de la mi nueva clase...”*

'una'.

[ 'una', 'EMS','det', 'TDET','indef1b'].

'otra'.

[ 'otra', 'EMS','det', 'TDET','indef3'].

'solución'.

[ 'solución', 'EMS','nom', 'GEN','fem', 'NUM','sg'].

'es'.

[ 'ser', 'EMS','v', 'MODOV','ind', 'PERS','3a', 'NUM','sg', 'TPO','pres'].

'transformar'.

[ 'transformar', 'EMS','v', 'MODOV','infin', 'TR','r', 'TC','c1'].

'rutas'.

[ 'ruta', 'EMS','nom', 'GEN','fem', 'NUM','pl'].

'la'.

[ 'el', 'EMS','det', 'TDET','art'].

'mi'.

[ 'mi', 'EMS','det', 'TDET','pos'].

'nueva'.

[ 'nuevo', 'EMS','adj', 'GEN','fem', 'NUM','sg'].

'clase'.

[ 'clase', 'EMS','nom', 'GEN','fem', 'NUM','sg'].

### **6.3. Reglas de reagrupamiento en MPS para las construcciones coincidentes con la lengua estándar**

Para analizar los sintagmas nominales núcleos que coinciden con los de la lengua meta se crean reglas de reagrupamiento en el software MPS que funciona de manera modular con Smorph, es decir, toma el resultado del análisis automático realizado por éste como punto de partida.

Luego de analizar el sintagma [sus muchos negocios] el software arroja los siguientes datos:

'sus'.

[ 'pos', 'EMS','pos'].

'muchos'.

[ 'mucho', 'EMS','indf2a', 'GEN','masc', 'NUM','pl'].

'negocios'.

[ 'negocio', 'EMS','nom', 'GEN','masc', 'NUM','pl'].

A partir de esta información morfológica se crea una regla en el archivo rcm.txt de la siguiente manera:

REGLA N° 1: % pos+ indef2a+nom da snomn%

S1 [L1,'EMS', 'det', 'TDET','pos'] S2 [L2, 'EMS','indef2a"TDET"] S3 [L3, 'EMS','nom'] --> S1+S2+S3 [L1+L2+L3, 'EMS','snomn' ].

A continuación debe ordenársele a la herramienta que agrupe los lemas designados según las etiquetas morfosintácticas correspondientes y el orden indicado y que la suma de ellos dé como resultado un sintagma nominal núcleo. La operación se hará visible en el archivo smorphg de la siguiente manera: 'sus muchos negocios'. ['pos indf2a nom', 'EMS', 'snn' ].

REGLA N° 2: % indef1a+ nom da snn%

S1 [L1,'EMS','indef1a"TDET"] S2 [L2, 'EMS','nom'] --> S1+S2 [L1+L2, 'EMS','snn' ].

Algunas condiciones

REGLA N° 3: % indef1b+ nom da snn%

S1 [L1,'EMS','indef1b', 'TDET'] S2 [L2, 'EMS','nom'] --> S1+S2 [L1+L2,'EMS','snn' ].

Un curso

REGLA N° 4: % pos+ nom da snn%

S1 [L1,'EMS','det', 'TDET','pos'] S2 [L2, 'EMS','nom'] --> S1+S2[L1+L2,'EMS','snn' ].

Su matrimonio

### 6.3.1. Resultados obtenidos para los sintagmas coincidentes con el español estándar

Los resultados obtenidos en el archivo smorph se muestran a continuación:

'algunas condiciones'. [ 'algunas condiciones', 'EMS', 'snn' ].

'un curso'. [ 'un curso', 'EMS', 'snn' ].

'su matrimonio'. [ 'su matrimonio', 'EMS', 'snn' ].

## 6.4. Reglas para las construcciones propias de interlengua

Para analizar las construcciones halladas en la interlengua tendré que crear nuevas reglas y lo más importante, indicar que el agrupamiento de estas palabras dé como resultado un sintagma nominal desviado, por lo tanto emplearé la expresión desv para distinguirlos de los otros sintagmas nominales.

REGLA N° 1: % indf1b + indf2b+nom da snndesv %

S1 [L1,'EMS','indef1b', 'TDET'] S2 [L2,'EMS','indef2b', 'TDET',] S3 [L3, 'EMS','nom'] --> S1+S2+S3 [L1+L2+L3, 'EMS','snndesv' ].

%Una otra solución%

REGLA N° 2: % art + det pos + adj + nom da snndesv%

S1 [L1,'EMS','det', 'TDET','art'] [L2,'EMS','det', 'TDET','pos'] S3 [L3, 'EMS','adj'] S4 [L4, 'EMS','nom'] --> S1+S2+S3+S4 [L1+L2+L3+L4, 'EMS','snndesv' ].

% la mi nueva clase%

#### 6.4.1. Resultados obtenidos para los sintagmas propios de la interlengua

Los resultados alcanzados en el archivo Smorph son los siguientes:

'una otra solución'. ['una otra solución', 'EMS', 'snndesv']

'una otra familia'. ['una otra familia', 'EMS', 'snndesv' ].

'la mi nueva clase'. ['el mi nuevo clase', 'EMS', 'snndesv' ].

'el nuestro perro'. ['el nuestro perro', 'EMS', 'snndesv'].

## 7. CONCLUSIONES

Se hallaron sintagmas nominales núcleos de interlengua que presentaban divergencias respecto a los de la lengua estándar. Luego de establecer que estas anomalías residían en la elección y sobre todo en la combinación de los determinantes, se especificó a éstos como artículos, posesivos, demostrativos, numerales e indefinidos y se propuso una clasificación para los últimos, de acuerdo a la posibilidad de combinación entre ellos y con los otros determinantes, para lo cual hubo que efectuar modificaciones tanto en el archivo Entradas como en el de Rasgos de Smorph.

Se agruparon las estructuras propias de interlengua en cuatro casos, de los cuales se eligieron los dos últimos para analizar automáticamente: específicamente los sintagmas que presentaban artículo seguido de posesivo más sustantivo común y los de indefinido seguido de la palabra otro más sustantivo común.

Con el objetivo de distinguir a estas estructuras se crearon reglas de reagrupamiento en MPS que las identifican con el rasgo snndesv.

## Referencias

- [1] Aït- Mokhtar Salah y Rodrigo Mateos J. (1995).“Segmentación y análisis morfológico de textos en español utilizando el sistema SMORPH”. *SEPLN*, 17, 29-41.
- [2] MPS ha sido especificado en el GRIL por Caroline Hagège, José Rodrigo, Gabriel G. Bès y Faiza Abacci, e implantado en C++ en un contexto Windows por Faiza Abacci.

- [3] Abney Steven (1991) “Parsing By Chunks” en Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht.
- [4] Solana, Zulema y Rodrigo Andrea (2005). “El sintagma nominal núcleo”, en *Desarrollo, implementación y uso de modelos para el procesamiento automático de textos* (ed. Victor Castel). Mendoza: Facultad de Filosofía y letras, UNCUIYO.
- [5] Bosque Ignacio -Gutiérrez-Rexach J. (2009) *Fundamentos de sintaxis formal*. Madrid: Akal.
- [6] Real Academia Española (2010) *Nueva gramática de la lengua española*. Buenos Aires: Espasa.
- [7] Leonetti, M. (1999), *Los determinantes*, Madrid: Arco/Libros.
- [8] Alcina Caudet M. Amparo (1999) *Las expresiones referenciales. Estudio semántico del sintagma nominal*, tesis doctoral, Col.lecció Tesis doctorals en microfitxa. Universitat de València, Valencia.
- [9] Eguren, L y Sánchez López, C. (2003), “La gramática de otro”, *Revista española de Lingüística*, 33,pp. 69-122.