PEDECIBA Informática
Instituto de Computación - Facultad de Ingeniería
Universidad de la República
Montevideo, Uruguay

Tesis de Maestría en Informática

Análisis de superficie basado en puntuación

Diego Garat Baridon

Febrero de 2006

Orientadora: Dra. Ing. Dina Wonsever

Resumen

A pesar que los símbolos de puntuación son fundamentales en la estructuración del texto, casi ninguna herramienta informática de análisis textual los aprovecha. Varios estudios confirman la importancia de su tratamiento en el idioma inglés; sin embargo, para el idioma español, es prácticamente inexistente la investigación sobre este tema dentro de la lingüística computacional.

El presente trabajo estudia el uso de la puntuación para el análisis de superficie de textos en español, y, como caso particular, se centra en la coma, por ser éste el signo que presenta la mayor variedad de usos en la estructuración de texto.

Con tal fin, se construye una categorización de las funciones de la coma que se adapte al procesamiento automático. Esto se realiza a partir del estudio de distintas clasificaciones existentes, pero corroborándola contra un corpus de textos periodísticos en español.

Finalmente, se construye un analizador sintáctico de superficie combinando métodos estadísticos y simbólicos. Por un lado, se obtiene un evaluador de la función de las comas a través de técnicas de aprendizaje automático. Por otro, se escriben reglas de análisis que aprovechan la clasificación realizada por el evaluador «aprendido», utilizando un formalismo de reglas de reescritura.

Se concluye que el tratamiento de la puntuación en el análisis sintáctico también es útil en el español. Además, se observa que la combinación de métodos simbólicos y estadísticos puede potenciar los resultados de ambos enfoques.

Palabras claves: puntuación, análisis sintáctico de superficie, aprendizaje automático, combinación de métodos simbólicos y estadísticos.

Agradecimientos

En primer lugar, mi agradecimiento es para Dina: sin sus ideas, su tiempo e infinita paciencia para conmigo, la tarea hubiera sido imposible.

Quiero agradecer a todos mis compañeros del grupo PLN por su aliento y apoyo en los momentos de desesperación. Mi agradecimiento también a los otros «valientes participantes» del seminario de Aprendizaje Automático del año 2002.

Durante todo este tiempo recibí el apoyo de mi familia «real» y de mi familia «adoptada»; gracias a todos ellos, en particular a Ana y a Natalia, que les tocó la peor parte.

Agradezco al PEDECIBA por haberme becado durante la duración de esta maestría.

Mi reconocimiento a todos los que, con su talento, permiten mejorar mi mundo con el suyo.

A Luisita, Lolo y Majo... los extraño.

I have spent most of the day putting in a comma and the rest of the day taking it out.

Oscar Wilde

Índice

Capítulo 1

Introducción

Los signos de puntuación cumplen un papel esencial en la correcta estructuración de la escritura: no sólo organizan el discurso, sino que permiten evitar ambigüedades en la interpretación del texto. A pesar de su importancia, el uso de los signos de puntuación es poco tratado o directamente ignorado en la mayoría de las herramientas informáticas.

A modo de ejemplo, los analizadores sintácticos los aprovechan mínimamente: su utilización se reduce, por ejemplo, a la segmentación en oraciones [36]. Briscoe y Carroll [8, 9], trabajando con la puntuación en el idioma inglés, demuestran que esto último es un error, en la medida que sus experimentos muestran una significativa mejora al considerarlos durante el análisis.

La coma es dentro de los signos de puntuación el que puede cumplir más funciones distintas en la escritura. Por lo general, se la asocia a las pausas en los enunciados, puestas únicamente a criterio del escritor. Sin embargo, hay casos en los cuales su uso es indispensable y otros en los cuales esta desaconsejado.

En el idioma español, según la normativa de la Real Academia Española (RAE) [49], es indispensable el uso de un par de comas para delimitar un inciso o separar elementos de una serie, y es un error la aparición de una entre el sujeto y el verbo de una oración o antes de una conjunción que cierra una serie de elementos (siempre que no se encuentre delimitando un inciso).

Se puede hablar, entonces, de distintos usos, funciones o *valores* que cumple o tiene este signo de puntuación en la estructuración del texto: las comas que separan elementos de una serie, las que introducen incisos, las que demarcan elementos transpuestos, etc.

Dado el rol fundamental que cumple este símbolo en la estructuración del texto, se decide estudiar en este trabajo la utilización de las comas como caso particular de análisis sintáctico guiado por la puntuación.

Con este objetivo, se busca crear un clasificador que valore las comas presentes en un texto, para luego emplearlas en un análisis sintáctico de superficie (*shallow parsing*). Este proceso consiste en estructurar parcialmente el texto, partiéndolo en segmentos (*chunks*), en vez de realizar un análisis completo [1]. A modo de

Texto:
Hasta se ve, gigante, el plano de una flor que crece y crece, rociada por la vacuna de la eterna juventud, y es la síntesis del héroe que renace.

Análisis:

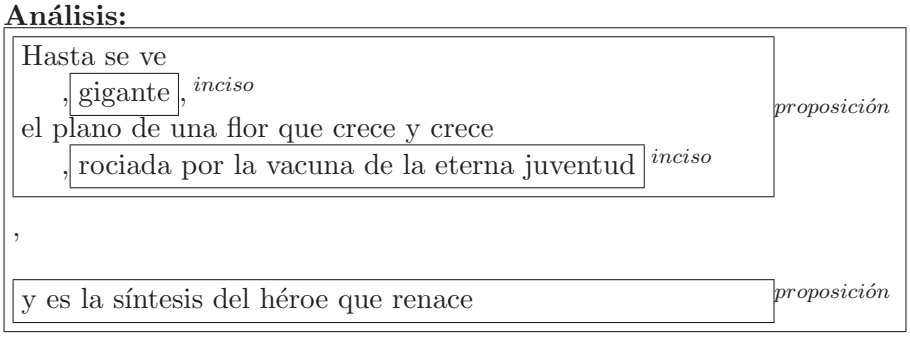


Figura 1.1: Ejemplo de un análisis

ejemplo, en la figura 1.1, se presenta un posible análisis de superficie sobre un texto dado.

1.1. Antecedentes

Existen varios estudios sobre el uso de la puntuación para el análisis de textos en inglés, a pesar que en este idioma es inexistente una normativa oficial. Esta carencia lleva a que se encuentren diferencias entre el uso de la puntuación en Estados Unidos y Gran Bretaña, e inclusive diferencias entre normativas dentro de un mismo país. A modo de ejemplo, cuando se realiza una cita, la coma puede ir antes o después de las comillas; otro ejemplo es la denominada *coma Oxford*, la cual se agrega —de forma opcional, según la normativa— antes de la conjunción que cierra una serie de elementos.

Sin embargo, y a pesar de que no hay una normativa única, el tratamiento de la puntuación en inglés ha sido ampliamente estudiado, no sólo exclusivamente desde un punto de vista lingüístico, sino también computacional. Varios trabajos se han enfocando en el uso de la puntuación como ayuda en el procesamiento automático de textos irrestrictos [54, 63].

Jones [29, 30], basándose en el marco teórico establecido por Nunberg [40, 41], experimenta con el uso de la puntuación en el análisis sintáctico. En su trabajo plantea una gramática que tiene en cuenta todos los símbolos de puntuación junto a las distintas reglas de interacción que entre ellos ocurren. Sus resultados preliminares demuestran que, si los símbolos de puntuación son quitados luego de separar en enunciados el texto, el análisis empeora significativamente, sobre todo en las oraciones más largas. Jones concluye que incluir los fenómenos de puntuación permite un mejor procesamiento del lenguaje natural.

Siguiendo la línea del trabajo anterior, Briscoe y Carroll [8, 9] desarrollan un analizador probabilístico que tiene en cuenta a la puntuación. Dado un texto, el resultado de este analizador es el conjunto de todos los posibles árboles de análisis según su gramática, junto a la probabilidad que tiene cada uno de éstos de ser el correcto. La puntuación, sostienen los autores, les permite disminuir la ambigüedad durante el proceso, mejorando sensiblemente los resultados:

> «(...) estos son los primeros experimentos que demuestran objetivamente la utilidad de la puntuación para resolver ambigüedades sintácticas y mejorar la cobertura del analizador»

lo que confirma las conclusiones preliminares de Jones.

Osborne [44] construye una gramática para el inglés de forma automática, partiendo de un modelo de análisis. Las reglas creadas toman a los signos de puntuación como otra categoría a ser tenida en cuenta. Los resultados de sus experimentos le permiten concluir que las reglas generadas con puntuación obtienen mejores análisis sintácticos que aquellas generadas sin este conocimiento.

Por otra parte, Doran agrega manualmente reglas que consideran la puntuación en el análisis de oraciones a la gramática XTAG para el inglés [21]. Sus resultados confirman la importancia de la puntuación:

> « (...) el uso de puntuación puede ser usado en el análisis de oraciones, reduciendo la ambigüedad de ciertas secuencias de palabras (...), y encontramos que las reglas de puntuación sí mejoran la cobertura de la gramática existente sin impactos negativos sobre el resto de la gramática»

En particular, Bayraktar, Say y Akman [4] realizan un estudio sobre la coma: su objetivo es analizar los distintos usos de este signo de puntuación en textos de la prensa escrita en inglés. Para alcanzar su meta, los autores adaptan una clasificación de los distintos usos de la coma a los fenómenos que encuentran en los textos. Luego, extraen los patrones sintácticos de las ocurrencias de las comas en textos anotados del *Wall Street Journal*[1]. Los patrones más frecuentes son manualmente etiquetados, asignándoles alguna de las categorías de su clasificación. Así, por ejemplo, un grupo nominal (GN) de la forma «GN, GN,» es un patrón que indica la presencia de un apositivo delimitado por comas (el 2do. grupo nominal del patrón).

Clasificando manualmente 211 de estos patrones —un 11 % de los 1978 patrones detectados— los autores logran cubrir el 80 % de las comas del corpus. Sin embargo, para llegar a cubrir un 90 %, les es necesario triplicar el número de patrones considerados. Luego, para caracterizar las clases de acuerdo a la variedad de sus patrones de uso, los autores establecen un criterio de «estabilidad» de

[1] Los textos forman parte del Penn Treebank, el cual es un corpus con anotaciones sintácticas realizado por la Universidad de Pensilvania.

las distintas clases de comas, según la cantidad de patrones que se precisan para cubrirlas.

Van Delden y Gómez [61, 62] aplican técnicas de estado finito para determinar el rol sintáctico de las comas en textos en inglés. Su propuesta es construir un conjunto de autómatas, de forma tal que cada uno de ellos captura un rol específico. Estos autómatas son construidos manualmente, a partir del análisis del corpus. En una primera fase, etiquetan las comas con los posibles valores asignados por los autómatas, para luego filtrar aquellos valores inválidos utilizando una matriz de coocurrencia de etiquetas.

A pesar de que todos los trabajos anteriores parecen indicar que el tratamiento de la puntuación es relevante, los trabajos sobre este tema para el español son prácticamente inexistentes dentro de la lingüística computacional.

1.2. Marco de trabajo

Este trabajo se inscribe dentro de las áreas del *aprendizaje automático* y la *lingüística computacional*. Cada una de estas áreas engloba una infinidad de subáreas, teniendo entre ellas muchos puntos de contacto.

Del área de aprendizaje automático se toman algunas herramientas que permiten construir el clasificador de comas. En la sección 1.2.1 se presenta una breve reseña del área, centrada en las aplicaciones de estas técnicas para la resolución de problemas relacionados con el procesamiento automático de texto.

En la sección 1.2.2 se introduce el «Método de Exploración Contextual». Este método sirve de marco teórico en la construcción del analizador sintáctico de superficie.

1.2.1. Aprendizaje automático

El área de *aprendizaje automático* estudia el uso de programas que «aprenden», entendiéndose por aprender el mejorar el desempeño en una tarea con la experiencia [34, 51].

Por ejemplo, los algoritmos de «*data mining*» intentan descubrir patrones en grandes bases de datos, en donde, debido a la gran cantidad de registros, un ser humano no podría encontrarlos fácilmente. De esta forma se puede establecer bajo qué condiciones se debe dar un crédito o qué tratamiento médico es el aconsejado para cierta enfermedad.

El aprendizaje automático también resulta útil en dominios donde el ser humano no puede determinar un algoritmo debido a la complejidad del problema, la falta de un modelo adecuado del dominio de trabajo o la gran cantidad de entradas a considerar. Por ejemplo, ¿cómo se puede reconocer una cara en una imagen?, ¿cómo puede un robot conducir un automóvil sin chocarlo?, etc. [34]

Otro dominio de aplicación incluye situaciones en las cuales las condiciones cambian dinámicamente y la información sobre el dominio de trabajo aumenta con el tiempo. Los sistemas de recomendación por filtrado colaborativo [6, 50]

sugieren al usuario posibles productos de su interés, basándose en los elementos considerados por otros usuarios con preferencias parecidas. Estas técnicas en general son utilizadas por sitios de compras en Internet (Amazon[2], Barnes & Nobles[3], etc.).

Dentro del área de la lingüística computacional, las técnicas estadísticas no son muy aplicadas en las décadas anteriores a la del noventa. Sin embargo, a partir de esa década, el uso de métodos estadísticos ha ido ganando terreno gracias a su capacidad de resolver problemas que los enfoques simbólicos, o bien no pueden resolver de forma adecuada, o bien requieren de un esfuerzo mucho mayor para su aplicación [15, 33]. El avance de estas técnicas puede ser observado en el gran número de publicaciones sobre su aplicación que se presentan en conferencias del área —por ejemplo, las «*Conference on Natural Language Learning*»—.

Prácticamente todos los métodos de aprendizaje automático han sido utilizados para resolver problemas de procesamiento de lenguaje natural. Por ejemplo, los analizadores morfosintácticos se encuentran implementados con redes neuronales [57], aprendizaje por revisión [39], programación lógica inductiva [19], árboles de decisión [38, 58] y combinaciones de varias técnicas en un único analizador [7].

Otros problemas dentro del área de la lingüística computacional en los que se han aplicado estas técnicas son : extracción y categorización de información [28], análisis sintáctico [25], reconocimiento y clasificación de entidades nombradas [52, 53], determinación del alcance de cuantificadores [26] y correspondencia del complemento verbal [2], por citar algunos ejemplos. Una buena reseña sobre este punto es la realizada por Márquez en [37].

1.2.2. Método de Exploración Contextual

El Método de Exploración Contextual (MEC), desarrollado por en el Laboratorio LaLICC[4], tiene como objetivo identificar información semántica contenida en textos [20].

Esta metodología plantea como hipótesis que los textos poseen unidades lingüísticas —léxicas, posicionales, signos de puntuación, etc.— que permiten establecer relaciones entre los distintos segmentos de un texto y dilucidar su sentido. El MEC se basa exclusivamente en información lingüística y no utiliza información específica del domino particular en el que se pueda inscribir el texto a tratar. Por ejemplo, se busca explotar en el análisis las expresiones que el propio autor incluye en el texto para estructurarlo, resaltarlo, etc.

El sistema cuenta con un conjunto de marcadores lingüísticos, los *indicadores*, que permiten detectar textos con cierto valor semántico. Así, «*en conclusión*», «*finalmente*», etc. son posibles indicadores discursivos de una conclusión por parte del autor.

[2] http://www.amazon.com
[3] http://www.barnesandnobles.com
[4] Universidad de París-IV Sorbonne, Francia

Sin embargo, un mismo indicador generalmente tiene asociado más de un valor semántico: identificarlo en el texto no asegura que su función sea la esperada. Luego, se plantea la necesidad de explorar el contexto de la ocurrencia detectada en busca de otros indicadores lingüísticos que permitan determinar el rol que juega esa ocurrencia particular del indicador en el texto.

El MEC se implementa con un conjunto de indicadores, cada uno de ellos con un grupo de reglas asociado. La presencia de un indicador en el texto «dispara» su conjunto de reglas, las cuales exploran el contexto del indicador en busca de los marcadores adicionales, los *índices*, que se establecen en las reglas para confirmar el valor semántico que el indicador sugiere. En caso que se cumplan todas las condiciones impuestas por la regla, se agrega una etiqueta semántica al texto. A modo de ejemplo, las etiquetas podrían ser «*conclusión*», «*definición*», etc.

Finalmente, las reglas se agrupan de acuerdo a la «tarea» que se quiere realizar. Esto permite establecer los indicadores, índices y reglas más adecuados al objetivo que se persiga, ya sea resumir, detectar definiciones, etc.

Dado que el MEC no se basa en conocimiento específico de un dominio, ni depende de análisis sintácticos profundos, se adapta a tareas de procesamiento de texto en las cuales no se busca un análisis completo de la entrada, sino que la tarea está dirigida hacia un objetivo particular. Este tipo de procesamiento, en donde se realiza únicamente el análisis requerido para completar una tarea, es común en los sistemas de extracción de información.

La metodología de exploración contextual se ha utilizado en varios trabajos, entre otros: resumen y filtrado de texto [17, 18]; y reconocimiento de proposiciones en español y francés [13, 67]. Este último se desarrolla en el proyecto Clatex, en donde se construye, entre otras cosas, un intérprete basado en el formalismo de «Reglas Contextuales» [67, 68]. La definición y un intérprete de estas reglas, utilizado en este trabajo, se presentan en el capítulo 4.

1.3. Objetivos

La puntuación cumple un rol relevante en el análisis de textos, siendo la coma el más polivalente de estos signos. Este trabajo tiene por objetivo el análisis sintáctico de superficie de textos irrestrictos en español, utilizando a las comas como orientadoras en este análisis. Se busca, entonces, que este analizador sea capaz de detectar estructuras —series, proposiciones, incisos, etc.— basándose principalmente en las comas presentes en el texto.

Para realizar el análisis, se precisa determinar qué función cumple cada coma. A su vez, esta tarea requiere: (a) establecer una clasificación de los posibles usos de la coma; y (b) valorar según esta clasificación a las comas presentes en un texto cualquiera.

En principio, la primera tarea se podría resumir a tomar la clasificación de la Real Academia Española. Sin embargo, y análogamente a lo que ocurre en

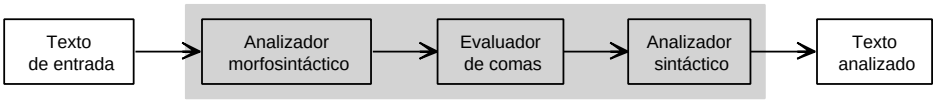


Figura 1.2: Etapas del análisis

los trabajos realizados en puntuación para el inglés, esta clasificación no resulta adecuada para el procesamiento automático de textos, debido a la gran cantidad de fenómenos que ésta contempla —alguno de los cuales no aparecen frecuentemente— y lo minucioso de su categorización. En definitiva, se debe desarrollar una clasificación específica para este trabajo que contemple los usos presentes en el texto, pero que también se adecue al procesamiento automático.

La segunda tarea consiste en construir un «clasificador» que logre valorar las comas según la categorización establecida. Aunque esto puede parecer análogo al trabajo que realizaron Bayraktar et al. pero para el español, su sentido es «inverso»: los mencionados autores determinan el valor de las comas (según su propia clasificación) a partir de la estructura sintáctica del texto, mientras que en este trabajo se pretende realizar el análisis sintáctico de la entrada partiendo de las comas previamente valoradas.

Pero entonces, dado que esta información se utiliza en el análisis sintáctico, la valoración de las comas debe lograrse —en la medida de lo posible— a partir de las características del propio texto o del resultado de su análisis morfosintáctico, y no de información que implique un análisis más elaborado.

Finalmente, se debe construir un analizador sintáctico de superficie que determine una posible estructura del texto a partir de las comas valoradas por el clasificador. La elección de este tipo de análisis se fundamente en que: (a) un análisis parcial es suficiente para algunas tareas de procesamiento, como, por ejemplo, en recuperación de información [24]; y (b) los analizadores de superficie suelen ser más eficientes y robustos que aquellos que realizan un análisis completo del texto [31].

En la figura 1.2 se presenta un diagrama del proceso de análisis.

1.4. Contribución

Este trabajo presenta contribuciones en tres niveles diferentes, marcados por las distintas etapas que conlleva el análisis estructural planteado.

En primer lugar, se realiza un estudio de un corpus para obtener como resultado una clasificación de los distintos usos de la coma en el español. Esta clasificación se adapta mejor al procesamiento automático de texto que la propuesta por la Real Academia.

En segundo lugar, utilizando distintas técnicas de aprendizaje automático, se obtiene un clasificador de comas. Este clasificador permite estimar qué función

cumple cada una de las comas a partir de la información morfológica de las palabras y «posicional» de las propias comas.

En tercer lugar, se construye un analizador sintáctico de superficie que estima una posible estructura del texto a partir de comas valuadas. Este analizador intenta sortear dos problemas: los errores cometidos por el analizador lexicográfico y los errores cometidos por el clasificador de comas.

Finalmente, se destaca la experimentación en combinar técnicas que tradicionalmente se presentan como enfrentadas dentro de la lingüística computacional: las que provienen de un enfoque estadístico *versus* las que lo hacen de uno simbólico.

1.5. Organización del documento

El primer paso de este trabajo es establecer una clasificación sobre el uso de las comas que se adapte al procesamiento automático de textos irrestrictos. En el capítulo 2 se define esta clasificación y se presentan otras dos que sirven de inspiración a la propuesta.

Una breve explicación de los algoritmos de aprendizaje utilizados, el proceso de construcción del clasificador de comas y los resultados obtenidos se encuentran en el capítulo 3.

En el capítulo 4, se describe el analizador sintáctico de superficie construido con el sistema de reglas contextuales: se presenta el formalismo de reglas contextuales, y se plantean las reglas empleadas para reconocer distintas estructuras presentes en el texto.

Finalmente, en el capítulo 5, se detallan las conclusiones y los posibles trabajos a futuro.

Capítulo 2

Usos de la coma

Por lo general, la coma se asocia con la realización de una breve pausa en la lectura de un enunciado. Sin embargo, su presencia no siempre marca una pausa, ni toda pausa es marcada por una coma. La coma cumple un rol esencial en la estructuración del texto, siendo fundamental en muchos casos para su correcta interpretación. Este es el caso en los siguientes ejemplos extraídos del Avance del Diccionario Panhispánico de Dudas de la Real Academia Española (RAE) [48]:

(2.1) a. Los soldados cansados volvieron al campamento con dos horas de retraso.

b. Los soldados, cansados, volvieron al campamento con dos horas de retraso.

En el primer enunciado, únicamente los soldados que se encontraban cansados volvieron al campamento con cierto retraso, habiendo presumiblemente otros soldados que no lo estaban y llegaron a tiempo (el adjetivo *cansados* tiene una función especificativa). En cambio, en la segunda frase, se deduce que todos los soldados estaban cansados, siendo ese el motivo de su retraso. Las comas cumplen en este último caso el cometido de delimitar un adjetivo explicativo.

El sentido del enunciado puede cambiar completamente por la presencia de una coma, como se muestra en los ejemplos 2.2 y 2.3, también extraídos del avance del diccionario de la RAE:

(2.2) a. Me he vestido como me indicaron.

b. Me he vestido, como me indicaron.

(2.3) a. Así no hubo quien lo convenciera.

b. Así, no hubo quien lo convenciera.

Así, en el ejemplo 2.2(a), le indicaron al sujeto la forma en la que debía vestirse; mientras que en 2.2(b) le indicaron únicamente que lo hiciera. En el ejemplo 2.3, la coma permite distinguir dos usos distintos de la palabra *así*: el de

modificador verbal, «de esa manera», en el caso (a), del de marcador discursivo, «por consiguiente», en el caso (b).

Este último fenómeno sucede con muchos de los marcadores discursivos. En el trabajo de Prada [46] se afirma que:

> «... la puntuación puede ayudar a delimitar no solamente el alcance de un marcador sino la presencia o no del mismo. Este aspecto es sumamente importante a la hora de hacer un análisis, en particular con términos que sean ambiguos y, por el hecho de estar comprendidos entre signos de puntuación, la ambigüedad desaparece.»

Aunque su uso puede estar supeditado al gusto del escritor, no siempre su uso es correcto y muchas veces, como se muestra anteriormente, es obligatorio. La Real Academia Española establece en su Manual de Ortografía [49] y en el Avance del Diccionario de Dudas [48] los usos correctos e incorrectos de la coma.

En particular, se destacan tres usos incorrectos: cuando una coma se escribe entre el sujeto y el verbo o entre el verbo y su complemento directo, sin separar un inciso; delante de la conjunción *que* precedida de *tan(to)* o *tal*; luego de la conjunción *pero* y antes de una oración interrogativa o exclamativa.

A pesar que la normativa sobre sus usos correctos e incorrectos es sencilla, existe un gran número de manuales de puntuación que explican su uso, y en varios trabajos se ha destacado la proliferación de texto mal puntuado. Penalvero [45], analizando el uso de la puntuación en la prensa de España, realiza la siguiente clasificación del mal uso de las comas:

- Ausencia de comas para señalar las estructuras y construcciones intercaladas.
- Ruptura de la relación entre el sujeto y el verbo o el verbo y el complemento directo por la presencia de una coma.

A modo de ejemplo, se extraen de ese trabajo las siguientes frases:

(2.4) *«El menor –15 años–, ha resultado ser analfabeto y el juez de Menores de Granada, al hilo de la nueva Ley de Responsabilidad Penal del Menor, le ha impuesto la obligación de que aprenda a escribir».

Diario ABC, 11 de noviembre de 2000.

(2.5) *«Cualquier idea que se le ocurriera a un progresista hacía temblar a los banqueros, pero lentamente el campo magnético de la seducción fue cambiando y agotada toda su carga aquella generación cayó en la tumba junto a sus guitarras».

El País, 8 de julio de 2001.

En la primer frase, la primera coma es incorrecta debido a que ocurre entre el sujeto y el verbo de la oración —rompiendo la relación verbo-sujeto—, sin ser su función la delimitación de un inciso. En la segunda, el problema consiste en la ausencia de comas que delimiten la construcción *agotada toda su carga*.

A continuación se presentan dos clasificaciones posibles de los usos de la coma, en las cuales se inspira la clasificación propuesta en este trabajo. La primera es la normativa de la RAE, un trabajo que busca cubrir todos los posibles usos de la coma en el idioma español; la segunda es la clasificación hecha por Bayraktar, Say y Akman [4] que, aunque es para el idioma inglés, está orientada hacia tareas de procesamiento automático de texto. Finalmente, se describe la clasificación propuesta y los resultados al aplicarla en un conjunto de artículos periodísticos.

2.1. Normativa de la RAE

En esta sección se presenta brevemente la clasificación para los usos lingüísticos de la coma, según el Avance del Diccionario de Dudas de la RAE [48].

En total, esta clasificación distingue más de veinte casos diferentes, agrupados en tres grandes clases: delimitadores de incisos, separadores de elementos y desambiguadores. A continuación, sólo se destacan aquellas que se consideran más importantes a los efectos del análisis estructural.

Algunos de los usos no detallados en este trabajo —que pueden encontrarse en el referido diccionario de dudas— corresponden a estructuras específicas, sin interés en el análisis sintáctico: datación de documentos, direcciones, referencias bibliográficas, etc.

2.1.1. Delimitadores de incisos

Las comas, junto con los paréntesis y las rayas, permiten introducir incisos (aposiciones, adjetivos explicativos, etc.) dentro de un enunciado. Bajo esta categoría, la Real Academia discrimina los siguientes seis usos:

1. Aposiciones explicativas.
2. Oraciones adjetivas explicativas.
3. Sustantivos que nombran o llaman al interlocutor.
4. Interjecciones.
5. Expresiones de carácter accesorio.
6. Comentarios o explicaciones de algo dicho.

Por ejemplo, en el enunciado 2.1 al comienzo de este capítulo, «*cansados*» entre comas cumple la función de adjetivo explicativo (punto 2).

En cambio, en el siguiente enunciado, se pueden observar un inciso apositivo (punto 1), «*Roberto Lavagna* », y al final un inciso que explica la «*preocupación empresaria*» (punto 6):

(2.6) El ministro de Economía, Roberto Lavagna, se mostró comprensivo con la preocupación empresaria, que reclama que no se los «castigue» con un doble costo cuando decidan echar a trabajadores sin causa.

En el ejemplo 2.7, se utilizan las comas para introducir en el texto a la interjección «*¡ay!*», interrumpiendo al enunciado sobre «*la solemnidad y el miedo*» (punto 4):

(2.7) (...) la solemnidad y el miedo al ridículo que, ¡ay!, tantos estragos suelen causar en los salones de lecturas poéticas.

Por lo general, como es el caso en todos los ejemplos presentados, se deben utilizar dos comas para delimitar el inciso. Sin embargo, se producen excepciones a esta regla. Así, cuando el inciso se encuentra al final del enunciado, el punto oficia al mismo tiempo como fin de inciso y oración; cuando figura al comienzo de la oración, el inicio de la oración marca a su vez el inicio del inciso. Este fenómeno, conocido como *absorción* [40] (el punto y el comienzo de oración *absorbieron* la coma de fin y de comienzo, respectivamente, del inciso). Esto no sólo ocurre entre la coma y el punto sino que también entre dos comas; para delimitar dos incisos consecutivos, se utilizan tres comas: la del medio marca el final del primer inciso y el comienzo del segundo.

2.1.2. Separadores de elementos dentro del enunciado

Bajo esta categoría, la Real Academia agrupa aquellos usos en que la coma cumple la función de separar elementos dentro del enunciado, ya sean simples (por ejemplo, enumeraciones de elementos simples o gramaticalmente equivalentes), ya sean elementos complejos (oraciones coordinadas adversativas o consecutivas, inversión en el orden regular de las partes del enunciado, etc.).

En total, se distinguen dieciocho casos diferentes, seis de los cuales no son considerados en este trabajo. Esto se motiva, como se menciona al comienzo de esta sección, en que se utilizan en estructuras que no se dan de forma libre en los textos, sino que están acotadas a secciones o construcciones muy particulares. A modo de ejemplo, se puede mencionar su uso en: las bibliografías, separando el nombre del autor, el nombre del artículo, etc.; los índices, al invertir el nombre y el apellido; los números, como separador de decimales o millares; la datación de documentos; etc.

Dentro de los casos que se consideran relevantes en este trabajo, y siempre según la RAE, se debe escribir una coma:

1. Para separar elementos de una enumeración.
2. Para separar miembros gramaticalmente equivalentes.
3. Para separar complementos de verbos elididos.

4. Delante de oraciones coordinadas encabezados por adverbios correlativos como, por ejemplo, *ya..., ya...*, y delante de la correlación disyuntiva *o bien..., o bien...*.
5. Delante de *excepto*, *incluso*, *salvo* y *menos* cuando funcionan como conjunciones.
6. Delante de conjunciones —*y*, *o*, *sino*, etc.— cuando unen oraciones compuestas.
7. Para separar los términos de construcciones de la forma *no solo..., sino (también)...*
8. Cuando se invierte el orden regular de las partes de un enunciado.
9. Detrás de ciertos enlaces, como ser: *sin embargo*, *por el contrario*, *además*, etc.
10. Detrás de los complementos encabezados por locuciones preposicionales, como ser: *en cuanto a*, *en ese caso*, etc.

y entre un par de comas:

11. Los sobrenombres o seudónimos luego del nombre verdadero.
12. El nombre del autor luego de mencionar una de sus obras.

A continuación se presentan algunas de las categorías mencionadas; otros casos, como el de verbo elidido o la transposición de elementos, se encuentran ejemplificados dentro de la clasificación propuesta en la sección 2.3.

En el enunciado del ejemplo 2.8 se observa dos usos diferentes de la coma:

(2.8) Además, Hugo nos advierte: «Un vértigo, un error, una derrota, una caída que dejó perpleja a toda la Historia (...)»

Por un lado, se utiliza una coma para marcar al conector «*Además*»; por otro, se introducen comas para separar a los elementos de la enumeración «*Un vértigo*», «*un error*» y «*una derrota*». Estos ejemplos se corresponden, respectivamente, con los puntos (9) y (1) de la enumeración previa.

En cambio, en el siguiente ejemplo se ejemplifica el uso de la coma luego de una conjunción para unir dos proposiciones, «*Corrientes está ...*» y «*la pelirroja corre...*»:

(2.9) Corrientes está cortada al paso de las limusinas, y la pelirroja corre y corre hacia la 9 de Julio.

Finalmente, la coma en el enunciado 2.10 se utiliza para crear una construcción de la forma «no tanto...sino», similar a la mencionada en el punto 7:

(2.10) Tal vez no tanto el arrepentimiento por no haber conseguido hacer la revolución, sino por haber alentado la creencia de que la revolución llegaría con sólo proponérselo seriamente.

2.1.3. Desambiguadores

Como se muestra en el comienzo de este capítulo en los ejemplos 2.2 y 2.3, la coma cumple una función sustantiva en la interpretación del texto, permitiendo desambiguar entre los posibles sentidos de un mismo enunciado. Según la Real Academia, su función no es en estos casos insertar o separar elementos dentro del enunciado sino definir cuál es el sentido de éste en un posible caso de ambigüedad.

Este uso «semántico» de la coma escapa, en principio, al alcance de este trabajo. Sin embargo, muchas veces la estructura que delimita la coma es la que termina dando sentido a la frase, como en el ejemplo 2.3 que, al escribir una coma luego de *así*, se establece que su función es de marcador conectivo y no de adverbio. Pero entonces, al tratar este uso de la coma, se contempla indirectamente alguno de estos casos.

2.2. Clasificación de Bayraktar, Say y Akman

Bayraktar et al.[4] proponen una clasificación de las distintas funciones que cumplen las comas en el idioma inglés. A diferencia de la normativa de la RAE, los autores tienen como objetivo el procesamiento automático de texto y no el análisis exhaustivo y minucioso de todas las posibles funciones de la coma.

A continuación se presentan estas categorías utilizando ejemplos extraídos del propio trabajo, junto a su traducción literal al español. En último lugar, se define la idea de *estabilidad sintáctica*, introducida por los autores para medir cuán sistemático es el patrón sintáctico en que se detecta cada uno de los usos definidos.

2.2.1. Categorías

Basándose en su corpus[1], los autores distinguen siete grupos: elementos de serie, elementos al comienzo de la oración, elementos al final de la oración, cláusulas no restrictivas, apositivos, interruptores y citas. Al igual que en este trabajo, no se tiene en cuenta la aparición de comas en números, fechas, direcciones, etc., por considerarlos usos no relevantes.

Elementos de serie

Las series de elementos, por lo general de un mismo tipo sintáctico, son separadas por comas. A modo de ejemplo, los autores dan los enunciados 2.11 y 2.12:

(2.11) Elsewhere, share prices closed higher in Amsterdam, Brussels, Milan and Paris.
En otras partes, los precios cerraron al alza en Amsterdam, Bruselas, Milán y París.

[1] Utilizaron para tal fin parte del Penn Treebank

(2.12) John went shopping, Mary cooked the meal and David washed the dishes.
John fue al centro comercial, Mary preparó la comida y David lavó los platos.

En el primer enunciado, la serie está compuesta por nombres propios («nominal simple», para los autores), mientras que la segunda es una serie un poco más compleja, al estar conformada por tres proposiciones.

El uso como separador de elementos de serie en esta clasificación es análogo al especificado por la RAE en los dos primeros puntos de la sección 2.1.2.

Elementos de comienzo de oración

En esta categoría se encuentran las comas que marcan los elementos iniciales de la oración. Éstos pueden ser, o bien modificadores sencillos —adjetivos, adverbios, etc.—, o bien enunciados o cláusulas más complejas.

Las comas que delimitan «*Running*», en el ejemplo 2.13, y «*Under two new features*» , en el enunciado 2.14, ejemplifican este uso.

(2.13) Running, he went up the stairs.
Corriendo, él subió las escaleras.

(2.14) Under two new features, participants will be able to transfer money from the new funds to other investment funds (. . .).
Bajo dos nuevas características, los participantes podrán transferir dinero desde los nuevos fondos hacia otros fondos de inversión (. . .).

Continuando con la comparación respecto a la clasificación de la Real Academia, estos casos comprenden a las comas por transposición (punto 8 de la sección 2.1.2), así como a los enlaces cuando ocurren al comienzo de la oración (punto 9 de la misma sección)

Elementos de fin de oración

Los elementos de fin de oración son análogos a los descriptos en el punto anterior, pero en este caso, se introducen al final y no al comienzo de la oración. Muchas veces, estos elementos se separan con una coma para evitar ambigüedades en la interpretación del enunciado.

En el siguiente ejemplo, la coma que separa «*though it will hardly eliminate it*» es considerada dentro de esta categoría:

(2.15) A face-to-face meeting with Mr. Gorbachev should damp such criticism, though it will hardly eliminate it.
Una entrevista cara a cara con el Sr. Gorbachev debería disminuir tal crítica, a pesar de que difícilmente la elimine.

Al igual que los elementos al comienzo de oración, este uso de la coma es contemplado por la RAE en la sección de separadores (punto 2.1.2).

Cláusulas no restrictivas

En esta categoría se consideran las comas que delimitan frases o cláusulas que modifican un sujeto, sin restringirlo.

Este es el caso de las comas que delimitan a la frase «*citing adverse developments in the market for high-yield "junk" bonds*» en el enunciado 2.16:

(2.16) A Western Union spokesman, citing adverse developments in the market for high-yield "junk" bonds, declined to say what alternatives are under consideration.
Un vocero de Wester Union, citando desarrollos adversos en el mercado para bonos «chatarra» de alto rendimiento, declinó decir qué alternativas están bajo consideración.

Esta función se corresponde con la de separación de oraciones explicativas, especificada en el punto 9 de la sección 2.1.1 en la clasificación de la RAE.

Apositivos

Al igual que en la clasificación de la RAE, se distingue el uso de las comas cuando rodean un apositivo.

(2.17) The new company, called Stardent Computer Inc., also said it named John William Poduska, former chairman and chief executive of Stellar, to the posts of president and chief executive.
La nueva compañía, llamada Sardent Computer Inc., también dijo que nombró a John William Poduska, anterior presidente y ejecutivo en jefe de Stellar, a los puestos de presidente y ejecutivo en jefe.

Por ejemplo, en el enunciado 2.17, se introduce el inciso apositivo «*former chairman and chief executive of Stellar*» para describir a «*John William Poduska*».

Interruptores

Los autores consideran interruptores a aquellas palabras o cláusulas que ocurren dentro de la oración y rompen con su sentido. Como ejemplo, presentan el siguiente enunciado:

(2.18) The new bacteria recipients of the genes began producing pertussis toxin which, because of the mutant virulence gene, was no longer toxic.
Las nuevas bacterias receptoras de los genes comenzaron a producir la toxina pertussis que, debido al virulento gen mutante, ya no era tóxica.

En este caso, la frase «*because of the mutant virulence gene*» es catalogada como una interrupción dentro del enunciado.

Las comas para delimitar interruptores están contempladas por la Real Academia como incisos de carácter accesorio o comentarios (puntos 4, 5 y 6 de la sección 2.1.1).

Citas

Bajo esta categoría se encuentran las comas que separan las palabras textuales, las cuales aparecen, por lo general, entrecomilladas.

(2.19) "The absurdity of the official rate should seem obvious to everyone," the afternoon newspaper Izvestia wrote in a brief commentary on the devaluation.
«*Lo absurdo de la cotización oficial debería parecer obvio para todos*», *escribió el diario vespertino Izvestia en un breve comentario sobre la devaluación.*

Observar que la coma aparece antes del cierre del entrecomillado en el inglés y no luego, como sucede en el español.

Este tipo de uso, a pesar de estar cubierto por la Real Academia en la sección de incisos (2.1.1), no tiene definido un ítem particular en esa clasificación.

2.2.2. Estabilidad sintáctica

En su trabajo, Bayraktar, Say y Akman construyen una base de patrones sintácticos a partir de un corpus anotado. Estos patrones se componen de su descripción, la cantidad de veces que se repite en el corpus y, como muestra, una de las oraciones en las que se encuentra presente. Por ejemplo:

Patrón: NP → NP, NP,

Clase: apositivo

Cantidad: 1880

Ejemplo:
Howard Mosher, president and chief executive officer, said he anticipates growth for the luxury auto maker in Britain and Europe, and in Far Eastern markets.
Howard Mosher, presidente y oficial ejecutivo en jefe, dijo que anticipa un crecimiento en la construcción de autos de lujo en Gran Bretaña y en Europa, y en los mercados del Lejano Oriente.

Los autores clasifican los patrones, comenzando con los más frecuentes, hasta llegar a cubrir el 80 % de las comas del corpus. Esto se logra luego de procesar

Clase	Comas	Patrones	Estabilidad
Series	2896	56	**52**
Elementos iniciales	2879	27	**107**
Elementos finales	722	32	**23**
Cláusulas no restric.	2476	29	**85**
Apositivos	3738	19	**197**
Interruptores	946	35	**27**
Citas	642	13	**49**

Cuadro 2.1: Estabilidad de las clases de comas según Bayraktar et al.

211 patrones, un 11 % del total. A partir de esto, definen la medida de estabilidad de una clase de coma:

$$estabilidad = \frac{número\ de\ comas}{número\ de\ patrones}$$

Esta medida permite estimar cuán regular es la estructura sintáctica definida por cada categoría. Los resultados les permiten afirmar que, en su corpus, la clase más estable es la de las comas apositivas, seguida de la que delimita elementos iniciales. Por el contrario, las comas más versátiles son aquellas que delimitan elementos finales, citas y series.

En el cuadro 2.1 se encuentra el valor de estabilidad para cada una de las clases presentadas en el punto 2.2.1.

2.3. Clasificación propuesta

En una primera instancia, se plantea una clasificación basada, principalmente, en los usos diferenciados por la Real Academia. Con esta clasificación se etiquetan cerca de mil comas para luego evaluar cuán apropiada resulta para capturar los fenómenos presentes en el corpus. Esto permite detectar varios problemas y desventajas, que se intentan solucionar en una nueva categorización. Ésta, a su vez, es probada sobre otro conjunto de textos, no utilizados en las instancias previas.

El corpus de trabajo al que se confrontan las dos clasificaciones está conformado por artículos de prensa extraídos del diario «Pagina 12» en su versión digital[2] y del corpus CORIN[3][14], consistente en textos periodísticos de publicaciones uruguayas. Los textos de CORIN son utilizados principalmente para evaluar la primera clasificación, mientras que los textos de «Página 12» se utilizan para la evaluación de la segunda.

El uso de artículos periodísticos tiene como principal ventaja la alta frecuencia de aparición de oraciones largas y, en consecuencia, un mayor uso de la coma para

[2] http://www.pagina12web.com.ar
[3] Corpus Informatizado (textos del español del Uruguay)

su estructuración.[4] Además, dado que en la mayoría de las publicaciones el texto es editado, sólo una pequeña proporción se encuentra mal puntuado.

Por otra parte, aunque la complejidad de estos textos permite estimar cuánto se adecúa la clasificación obtenida a ejemplos «reales», se debe tener en cuenta que esta misma complejidad juega en contra en el momento de procesar al texto automáticamente.

A continuación se presentan ambas clasificaciones, utilizando enunciados extraídos del corpus. En estos ejemplos, la categoría asignada se muestra en el propio texto como un superíndice de la coma.

2.3.1. Primera clasificación

La primera clasificación surge a partir de los distintos usos de las comas descriptos por la RAE, pero agrupando varios de éstos bajo una misma categoría, debido al gran número de casos planteados.

Se considera, entonces, el siguiente conjunto de etiquetas:

- comas de marcado de incisos iniciales y finales
- delimitadoras de series adjetivales, nominales, proposicionales y varias
- coma delimitadora de una transposición
- coma por verbo elidido

A continuación se describe el alcance y se presentan ejemplos de cada una de estas categorías. Luego se describen los resultados obtenidos al aplicar la clasificación, junto con los problemas que se intentan superar en una segunda —y definitiva— clasificación.

Incisos

Los incisos se limitan utilizando dos comas, asignándoles una etiqueta de comienzo («CMII») y de fin de inciso («CMIF») respectivamente. Por ejemplo:

(2.20) Sobre esa tumba que mira el mar,$^{\texttt{CMII}}$ reza la leyenda $^{\texttt{CMIF}}$, fue haciéndose Nápoles.

Como se señala en la sección 2.1.1, no todo inciso está delimitado por dos comas, dado que el comienzo y el final de una oración pueden actuar como límite. En este caso, se etiqueta únicamente a la coma que marca el comienzo o el fin con «CMII» o «CMIF» respectivamente. Las comas «absorbidas» por los límites de oración se denotan en los siguientes ejemplos entre paréntesis rectos.

[4] Por ejemplo, en pruebas preliminares realizadas sobre textos de manuales informáticos, se encuentra poca puntuación presente en las oraciones.

(2.21) [CMII]A un lado del mostrador CMIF,CMII custodiadas por una bandera de colores brillantes que recuerda el lugar donde todo comenzó CMIF, las «fresas» siguen recordando las costumbres del pueblo que creó las recetas.

(2.22) Dos grupos en dos turnos son los que se encargan de trabajar sobre mesadas larguísimas de madera masas generosas,CMII nacidas de harinas que se seleccionan de acuerdo con el origen y el uso.[CMIF]

Todos los casos mencionados en la sección «delimitador de incisos» de la clasificación de la RAE (sección 2.1.1) quedan comprendidos con estas dos marcas. También se encuentran contempladas algunas construcciones entre comas, catalogadas en la RAE como separadores: los enlaces, los sobrenombres, etc.

Respecto a la clasificación de Bayraktar et al., las categorías de inciso final e inicial delimitan los elementos de comienzo y fin de oración, las cláusulas no restrictivas, los apositivos, los interruptores y las citas.

Series

Al igual que en las otras clasificaciones, se consideran comas de series a aquellas que separan elementos que cumplen una función similar en el enunciado. Se distinguen dos grandes grupos: las series proposicionales («CMSP») y las series «simples». Estas últimas comprenden, a su vez, a las series adjetivales («CMSA»), las series nominales («CMSN») y a una tercera categoría «otras»(«CMSO») para todas las comas que delimiten series no consideradas por los casos anteriores (por ejemplo, las adverbiales).

(2.23) Es mediodía,CMSP y de un segundo al otro las baldosas gastadas quedan cubiertas por montones de pies.

(2.24) No se sabe, en realidad, si lo que atrae es la promesa de galletas tibias,CMSN el anisado de unas rosquitas,CMSN la hogaza de pan dorada en su punto justo o todo eso junto.

(2.25) En fin, en el espejo de la publicidad se ven mujeres flacas,CMSA bellas,CMSA exitosas,CMSA buenas madres,CMSA trabajadoras.

En las series simples, por lo general, los últimos dos elementos están separados por una conjunción —y, o, ni, etc.— aunque esto no es estrictamente necesario. En particular las series terminadas en *etc.* no deben llevar conjunción.

En el ejemplo 2.23, la serie está compuesta por las proposiciones: «*es mediodía*» y «*y de un segundo a otro las baldosas gastadas quedan cubiertas por montones de pies*». La serie nominal del ejemplo 2.24 tiene cuatro elementos: «*la promesa de galletas tibias*», «*el anisado de unas rosquitas*, «*la hogaza de pan dorada en su punto justo* y «*todo eso junto*». La conjunción «*o*» actúa como separador

de los últimos dos elementos; en cambio, la serie adjetival presente en el enunciado 2.25 ,«*flacas,. . . ,trabajadoras*», no tiene una conjunción como separador entre sus últimos elementos.

En Bayraktar et al., las comas de serie, tanto las «simples», como las proposicionales, están comprendidas en una única categoría (sección 2.2.1). En la clasificación de la Real Academia, estas comas abarcan prácticamente todas las categorías del grupo de separadores (sección 2.1.2).

Transposición

La coma como delimitador de una transposición («CMTR») se utiliza cuando se desea anteponer al verbo elementos que suelen ir luego de éste. Por ejemplo:

(2.26) Tras su muerte$^{\texttt{CMTR}}$, muchos de sus admiradores estaban en las filas judías.

en donde un posible orden «normal» sería: «Muchos de sus admiradores estaban en las filas judías tras su muerte.».

El uso de comas para marcar las transposiciones está implícitamente considerado por la RAE en su clasificación (punto 8 de la sección 2.1.2). En cambio, en la de Bayraktar et al., la transposición está comprendida dentro de los «elementos de comienzo de oración» (sección 2.2.1).

Elisión

Este uso de la coma («CMVE») corresponde, según la Real Academia, cuando se quiere «separar el sustantivo de los complementos verbales cuando el verbo está elidido por haber sido mencionado con anterioridad o estar sobrentendido». Por ejemplo[5]:

(2.27) Parafraseando a un dirigente soviético ya caído en desgracia: fuera del amor,$^{\texttt{CMVE}}$ nada; dentro del amor,$^{\texttt{CMVE}}$ todo.

La RAE contempla este uso dentro de la categoría de «separadores dentro de un enunciado», descripta en la sección 2.1.2. Por otra parte, este tipo de fenómeno no está contemplado en la clasificación de Bayraktar et. al.

Otras

Dentro de esta categoría caen todas aquellas comas que no se encuentran comprendidas en ninguno de los usos arriba mencionados como ser: datación, separador de decimales, bibliografía, etc. También se incluye en esta categoría los casos del mal uso, como los señalados al comienzo de este capítulo.

[5] En realidad, en el ejemplo se dan concomitantemente los fenómenos de elisión y transposición.

Resultados

Con esta clasificación, se etiquetan veinte textos del corpus CORIN. El total de comas etiquetadas en estos artículos se eleva a 1174, asignándole una única etiqueta a cada coma. La distribución en las distintas categorías se puede observar en el cuadro 2.2.

Clasificación	**Cantidad**	**%Corpus**
Inciso Final (CMIF)	361	30,7
Inciso Inicial (CMII)	308	26,2
Serie proposicional(CMSP)	268	22,8
Serie nominal (CMSN)	85	7,2
Serie adjetival (CMSA)	21	1,8
Serie otra(CMSO)	27	2,3
Transposición (CMTR)	29	2,5
Verbo Elidido (CMCO)	7	0,6
Otras	68	5,8
Total	**1174**	**100,0**

Cuadro 2.2: Distribución de las comas según la primera clasificación

La ocurrencia de comas adjetivales es prácticamente inexistente, inclusive es superada por otros tipos de series simples —series adverbiales, por ejemplo—. Parece razonable, entonces, colapsar las etiquetas de series nominales, adjetivales y otras en una única etiqueta de serie simple.

Surge, además, otro problema: la clasificación anterior no tiene en cuenta el uso múltiple de una misma coma. Este es el caso del siguiente ejemplo:

(2.28) Hasta el momento, salvo en el caso de la ofensiva contra los desarmaderos, la estrategia que siempre primó fue la de mantener una especie de pacto de no-agresión con los pesados de la fuerza ...

Al analizarlo, surgen varias opciones a la hora de decidir cuál es la estructura del enunciado, y por ende, el valor de la coma a asignar:

(a). ¿Las dos comas marcan un único inciso, «*salvo en el caso de la ofensiva contra los desarmaderos*»?

(b). ¿O lo mejor sería considerar a «*Hasta el momento, salvo en el caso de la ofensiva contra los desarmaderos*» como una serie de incisos que afectan al resto del enunciado?

(c). ¿O considerar a «*salvo en el caso de la ofensiva contra los desarmaderos*» como un inciso anidado dentro del inciso «*Hasta el momento...* »?

Las distintas posibilidades se pueden ver gráficamente en el ejemplo 2.29.

(2.29) a. Hasta el momento
, salvo en el caso de la ofensiva contra los desarmaderos,
la estrategia que siempre primó fue...

b. Hasta el momento,
salvo en el caso de la ofensiva contra los desarmaderos,
la estrategia que siempre primó fue...

c. Hasta el momento
, salvo en el caso de la ofensiva contra los desarmaderos,
la estrategia que siempre primó fue...

En el enunciado 2.30, la coma marcada con un recuadro cumple una doble función: delimitar el final del inciso «*las buenas de las malas*» y continuar la serie compuesta de los elementos «las viejas pasiones», «los antiguos rencores», etc.

(2.30) ...en lugares no habilitados para bailar pero que se convertirán en pistas ardientes donde se derretirán todas las pasiones, las buenas y las malas, los antiguos rencores, los viejos amores, los que quedan a pesar de todo, los que se inician.

Para sortear este problema, se puede, por ejemplo, establecer un orden de preferencia sobre las categorías: si una coma cumple más de una función, se le asigna aquella etiqueta que tenga mayor preferencia.

Otra opción es crear nuevas categorías para diferenciar cada una de las posibles combinaciones de etiquetas —por ejemplo: coma de inciso final e inicial, inciso final y modificador, etc.—. Dada la gran cantidad de nuevas etiquetas que se generarían, esto último no parece ser una buena solución[6]. Se opta, en cambio, por permitir asignar más de una categoría a una misma coma.

Otra duda se presenta con los incisos especificativos y las series adjetivales. Muchas veces un inciso especificativo podría considerarse como una serie adjetival, en donde no está presente la conjunción final.

(2.31) a. ...masas generosas, nacidas de harinas ... *inciso*

b. ...masas generosas, nacidas de harinas ... *serie*

En el enunciado 2.22 se puede interpretar a «generosas» y «nacidas de harinas...» como una serie adjetival que califica a «las masas»; otra posibilidad es asociarle a las «masas generosas» la frase «nacidas de harinas...» como inciso especificativo.

[6]Probablemente esta opción aumente la cantidad de ejemplos bien etiquetados que hay que presentarle al algoritmo de aprendizaje para obtener un buen clasificador.

2.3.2. Clasificación final

Se modifica la clasificación anterior con el objetivo de superar alguno de los problemas encontrados. Estas modificaciones consisten en agregar nuevas categorías para contemplar mejor ciertos fenómenos, mientras que las categorías de baja aparición en el corpus son eliminadas.

Por ejemplo, en la nueva clasificación los incisos se «especializan», diferenciándose ahora los incisos discursivos, los modificadores, etc. Además, se agrega una categoría para distinguir las comas en los enunciados «bipolares» y las que delimitan los conectores discursivos.

Por otro lado, las series simples terminan colapsadas bajo una única etiqueta, y los elementos transpuestos desaparecen bajo una nueva categoría, un poco más general, los incisos «modificadores».

A continuación, se describen las modificaciones respecto al conjunto anterior de etiquetas. En el cuadro 2.3 se encuentra la clasificación final y su relación con las otras dos clasificaciones presentadas.

Clasificación propuesta	**Real Academia**	**Bayraktar et al.**
Inciso Inicial (CMII)	inciso	todos menos series
Inciso Final (CMIF)	inciso	todos menos series
Inciso Discurso (CMID)	inciso	citas
Serie Simple (CMSO)	separador (puntos 1 y 2)	serie
Serie Prop. (CMSP)	separador (puntos 5 y 6)	serie
Modificador (CMMO)	separador (puntos 8 y 10)	comienzo oración
Conector (CMCO)	separador (punto 9)	comienzo de oración
Bipolar (CMBI)	separador (punto 8)	—

Cuadro 2.3: La clasificación propuesta vs. la de la RAE y la de Bayraktar et al.

Incisos discursivos

Los incisos discursivos, muy comunes en artículos periodísticos, se pasan a distinguir con la etiqueta «CMID». Por ejemplo:

(2.32) «Usted tiene el poder con telecable»,[CMID] dice la publicidad de una empresa de televisión.[CMID]

Este tipo de comas precede en la mayoría de los casos a verbos de carácter discursivo, como ser: decir, opinar, comunicar, transmitir, pensar, etc. La coma en el ejemplo 2.32 es etiquetada como inciso inicial con la primera clasificación.

No existe una categoría particular para este tipo de incisos en la clasificación de la Real Academia. Por otra parte, Bayraktar et al. definen la clase «cita» (sección 2.2.1) análoga a estos incisos .

Series

Se juntan bajo una misma categoría a las comas de series nominales, adjetivales y otras, debido a la baja frecuencia de aparición de estos tipos de comas —sobre todo de las adjetivales—. Estas series se denominan bajo el nombre genérico de series simples, y se utiliza la marca «CMSO» para su etiquetado. Por otra parte, las series proposicionales no sufren ningún cambio.

Modificadores

Las comas etiquetadas como «modificadoras» («CMMO») aparecen luego de ciertos adverbios o frases adverbiales que afectan no sólo alguna parte sino a todo el enunciado. Por ejemplo: *hasta ahora*, *en París*, *luego de tres días*, etc.

(2.33) Hasta hoy,$^{\texttt{CMMO}}$ pasados diez meses,$^{\texttt{CMMO}}$ no se sabe por qué mataron a Piazza, pero todos los indicios van en dirección a un crimen policial.

Varios modificadores pueden «acumularse» al comienzo de la oración. Esto es lo que ocurre en el ejemplo 2.33 con «*hasta hoy*» y «*pasados diez meses*».

Esta etiqueta se basa tanto en la coma de «comienzo de oración» de Bayraktar et al. (sección 2.2.1), como en la anterior coma por transposición, incluida como parte de esta nueva categoría.

Conectivos

Como se menciona al comienzo de este capítulo, es frecuente que se utilice una coma para marcar la presencia de un conectivo porque, precisamente, es ésta la que le da tal carácter[7]. Esto motiva la creación de la categoría «conectivos» («CMCO»), que comprende a aquellas comas que son utilizadas luego de conectores discursivos, como ser *finalmente*, *en consecuencia*, etc.

Al igual que los modificadores, los «conectivos» ocurren, en la mayoría de los casos, al comienzo de la oración, posiblemente luego de otros conectivos o modificadores. Por ejemplo:

(2.34) Sin embargo,$^{\texttt{CMCO}}$ la audacia tenía una firma raigambre en el propio folklore.

Esta nueva categoría es, dentro de la clasificación de la RAE, uno de los casos de separador (punto 9 de la sección 2.1.2), y está comprendida dentro de los «elementos de comienzo de oración» en la clasificación de Bayraktar et al.

[7] Recordar el ejemplo 2.3 con el conector *así*.

Bipolares

Las comas «bipolares» («CMBI») se utilizan en ciertas construcciones que cuentan con dos miembros de igual importancia (frases hipotéticas, concesivas, etc.), a pesar que, en un enfoque tradicional, una es considerada como subordinada y otra como principal.

(2.35) Si se los deja avanzar en esa dirección,$^{\texttt{CMBI}}$ los problemas de salud pública van a volverse colosales.

En el ejemplo (2.35), *«si se los deja avanzar en esa dirección»* es la parte subordinada —en este caso, hipotética— dentro del enunciado.

Bayraktar et al. no distingue este uso de la coma. Sin embargo, esto sí ocurre en la clasificación de la RAE, y es inspirada en ésta que se agrega a la clasificación.

Resultados

El nuevo conjunto de clases se utiliza para etiquetar un nuevo corpus compuesto de 77 artículos periodísticos, esta vez extraídos del diario «Página 12». En total, se recolectan 5351 ejemplos; hay que tener en cuenta que una misma coma puede generar varios ejemplos, al ser clasificada simultáneamente en más de una clase. La distribución de las comas del corpus en las distintas categorías se encuentra en el cuadro 2.4.

Clasificación	**Cantidad**	**%Corpus**
Inciso Inicial (CMII)	1757	32,8
Inciso Final (CMIF)	948	17,7
Inciso Discurso (CMID)	141	2,6
Serie Simple (CMSO)	832	15,5
Serie Prop. (CMSP)	790	14,8
Modificador (CMMO)	542	10,1
Conector (CMCO)	198	3,7
Bipolar (CMBI)	91	1,7
Otras	52	1,0
Total	**5351**	100,0

Cuadro 2.4: Clasificación de las comas del corpus

Nuevamente, una categoría parece no ser muy significativa debido a su baja frecuencia de aparición: las comas bipolares prácticamente no se encuentran en el corpus. Por otra parte, la proporción de comas de inciso final disminuye significativamente respecto a la clasificación original, de un 26,2 % al 17,7 %. Presumiblemente, esto se debe a la «especialización» que sufre esta categoría al distinguirse los «modificadores», la mayoría de los cuales, por ocurrir al comienzo de los enunciados, son considerados como incisos finales en la clasificación anterior.

Resumen

A pesar que podría parecer sencillo determinar cómo utilizar una coma, en la práctica esto no resulta ser tal, en parte debido a las distintas funciones que éstas pueden cumplir en el texto. Las distintas clasificaciones lo reflejan de diferente forma según su finalidad. En particular, se presentan tres clasificaciones: (a) la normativa de la Real Academia Española, un estudio detallado sobre sus posibles usos; (b) la de Bayraktar et al., realizada para el inglés; y (c) la clasificación propuesta en este trabajo, inspirada en las dos clasificaciones anteriores y corroborada en un corpus conformado por artículos periodísticos.

Capítulo 3

Clasificador

Establecida una clasificación, se busca crear una herramienta que valore las comas presentes en un texto. Para la implementación de este clasificador, se pueden escoger entre diversas técnicas, ya sean analíticas, estadísticas o una combinación de ambas. Este trabajo explora, en particular, la utilización de dos algoritmos de *aprendizaje automático*.

El aprendizaje automático tiene como objetivo la adquisición automática de conocimiento a través de algoritmos que mejoran su desempeño con la experiencia [59]. En líneas generales, cualquier problema de aprendizaje puede verse como la aproximación de una función desconocida, o *función objetivo*, a través de algoritmos que utilizan un conjunto de ejemplos, el *conjunto de entrenamiento*. Esto implica que, dado un problema a resolver, se debe: (a) establecer una representación de la función objetivo que lo modele; (b) elegir el algoritmo que permita aproximar esta función; y (c) contar con un conjunto de entrenamiento a partir del cual aprender. [34]

La función objetivo es, en este caso, el clasificador de comas. Su salida se encuentra determinada por las nueve clases definidas en la sección 2.3. Ahora, el problema consiste en determinar qué espacio se va a utilizar para representar el dominio, esto es, ¿qué atributos de la coma son relevantes (¡y calculables!) para actuar como entrada de la función?

Dos algoritmos se utilizan para inducir el clasificador: *árboles de decisión* [34, 47] y *boostexter* [55, 56]. Ambos son algoritmos *supervisados*: requieren de un conjunto de entrenamiento correctamente etiquetado para inducir la función objetivo[1]; esto conlleva la clasificación manual de todas las comas presentes en el corpus.

En la sección 3.1 se presentan los algoritmos de aprendizaje utilizados. Los atributos de las instancias se describen en la 3.2, y los resultados obtenidos, en la sección 3.3. Finalmente, se comenta la herramienta creada para ayudar en el proceso de etiquetado y entrenamiento en la sección 3.4.

[1] En contraste a los algoritmos **no** supervisados, que deducen la función en base a ejemplos sin etiquetar, utilizando una función de distancia en el dominio.

3.1. Algoritmos

Se utilizan dos métodos para la obtención del clasificador: árboles de decisión y *boostexter*. Estos algoritmos son ampliamente utilizados dentro del área de la lingüística computacional para una gran variedad de tareas: análisis lexicográfico [43, 58], reconocimiento y clasificación de nombres propios [5, 11, 69], vinculación de grupos preposicionales [2], etc.

La principal ventaja de los árboles de decisión sobre otros algoritmos de aprendizaje es que se puede determinar bajo qué condiciones clasifican una instancia de una forma y no de otra. Esto permite, por ejemplo, «recodificar» el clasificador, incorporando el conocimiento adquirido en otros sistemas.

Por otro lado, los algoritmos basados en *boosting* dan por lo general mejores soluciones. Sin embargo, aunque más preciso, el resultado parece una caja negra: dada una instancia, se conoce la clase asignada pero no cómo se arriba a esta clasificación.

A continuación se presenta brevemente las versiones básicas de estos dos algoritmos. El lector interesado puede encontrar algunas extensiones en [47], [55] y [56].

3.1.1. Árboles de decisión

Este algoritmo toma instancias asignadas a un conjunto de clases y construye un clasificador en forma de árbol. Los árboles deben cumplir que:

a) los nodos internos «preguntan» por un atributo de las instancias;

b) sus ramas se etiquetan con una posible respuesta a ese atributo;

c) las hojas se etiquetan con una de las posibles clases.

En la figura 3.1 se muestra un posible árbol de decisión para el clasificador de comas.

Para clasificar una nueva instancia, se comienza aplicando el test que se encuentra en la raíz del árbol. Según la respuesta, se continúa por el subárbol que corresponda. Este procedimiento se aplica iterativamente hasta llegar a una hoja, la cual determina la clase de la instancia. Así, el árbol de la figura 3.1 clasifica a la primera coma del cuadro 3.3 como «CMSP» —«*verbo anterior*» es «OTRO» y «*distancia*» es «>3»— y como «CMII» a la segunda —«*verbo anterior*» es «OTRO» y «*distancia*» es «≤ 3»—.

Una de las principales ventajas de los árboles de decisión es que permiten, a diferencia de otros métodos, encontrar una representación «entendible» del clasificador inducido. En contraposición, entrenando una red neuronal se obtiene un clasificador en el cual es difícil interpretar qué es lo que efectivamente determina que una instancia sea etiquetada con una categoría y no con otra.

El algoritmo básico induce el árbol seleccionando un atributo para la raíz; luego, particiona el conjunto de entrenamiento de acuerdo a los posibles valores

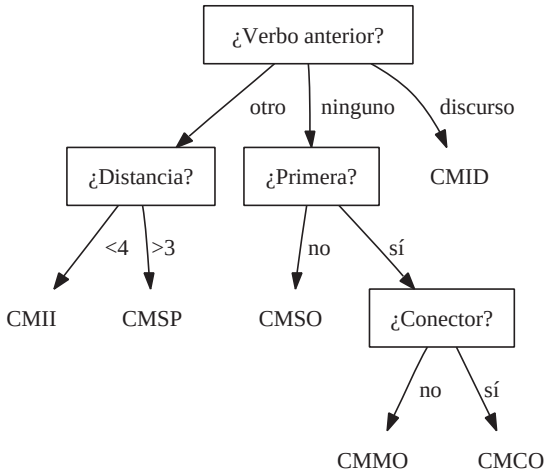


Figura 3.1: Ejemplo de un árbol de decisión

de ese atributo, y aplica recursivamente el procedimiento, armando para cada partición el subárbol de la rama correspondiente. El proceso continúa hasta que todos los elementos pertenecen a la misma clase o hasta que no hay más atributos para elegir.

La selección del atributo en cada paso varía de acuerdo a la implementación del algoritmo. En su versión original, se elige el atributo que da una mayor ganancia de información; en otras palabras, aquel atributo que, para todas las instancias del conjunto de entrenamiento que toman igual valor, la clasificación es lo más homogénea posible (a igual valor del atributo, igual valor de la clasificación).

Al intentar seleccionar en cada paso el atributo que mejor particiona al conjunto de entrenamiento, el algoritmo privilegia a los árboles de menor sobre los de mayor profundidad. Esta es una técnica *ávida*: luego de seleccionado el atributo por el que preguntar en un nodo, jamás se reconsidera si esa opción permite obtener el árbol que mejor clasifica a todos los ejemplos, no sólo en el conjunto de entrenamiento, sino también en todo el dominio de trabajo. En consecuencia, el algoritmo es susceptible a caer en óptimos locales.

El cuadro 3.1 muestra un algoritmo básico, el algoritmo ID3, para obtener un árbol de decisión (extraído de [34]), el cual induce funciones *booleanas* sobre atributos también *booleanos*.

El árbol obtenido puede *sobreajustarse* al conjunto de entrenamiento. Por sobreajuste se entiende a la existencia de otras hipótesis distintas a la obtenida que se comportan peor sobre el conjunto de entrenamiento, pero tienen mejor resultado sobre la totalidad del espacio de instancias. Esto se puede deber tanto al ruido[2] en el conjunto de entrenamiento, como al apartamiento de la distribución del conjunto de entrenamiento respecto a la distribución del conjunto de todas

[2]Instancias mal clasificadas.

Algoritmo **ID3**(*Ejemplos*, *Atributos*)

- Si todos los ejemplos son de una misma clase, etiquetar la raíz con esa clase.
- Si el conjunto *Atributos* es vacío, etiquetar con el valor más común
- En caso contrario,
  - Preguntar por A, el atributo que maximiza $Ganancia(Ejemplos, A)$
  - Para cada posible valor v_i de A:
    - Genero una rama.
    - Sea $Ejemplos_{v_i}$ el subconjunto de *Ejemplos* donde $A = v_i$
      - Si $Ejemplos_{v_i}$ es vacío etiquetar con el valor más probable
      - En caso contrario **ID3**($Ejemplos_{v_i}$, *Atributos* -$\{A\}$)

En donde:

$$Ganancia(Ejs, A) = Entrop(Ejs) - \sum_{v_i \in Valores(A)} \frac{|Ejs_{v_i}|}{|Ejs|} Entrop(Ejs_{v_i})$$

$$Entrop(Ejs) = \sum_{c_i \in Clases} -p_i \log_2 p_i$$

siendo p_i la proporción de ejemplos pertenecientes a la clase i-ésima.

Cuadro 3.1: Algoritmo ID3

- Dados: $(x_1, Y_1), \ldots, (x_m, Y_m)$ donde $x_i \in \mathcal{X}$, $Y_i \subseteq \mathcal{Y}$
- Inicializar: $D_1(i, l) = (m\,k)^{-1}$
- Para $t = 1, \ldots, T$:
  - Utilizar el aprendiz débil con distribución D_t
  - Obtener la hipótesis débil $h_t : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$
  - Elegir $\alpha_t \in \mathbb{R}$
  - Actualizar:

$$D_{t+1}(i, l) = \frac{D_t(i, l)\, \exp(-\alpha_t\, Y_i[l]\, h_t(x_i, l))}{Z_t}$$

- Dar como salida:

$$f(x, l) = \sum_{t=1}^{T} \alpha_t\, h_t(x, l)$$

Cuadro 3.2: Algoritmo AdaBoost.MH

las posibles instancias.

Una técnica para evitar el sobreajuste es la «poda» del árbol. Esta técnica consiste en quitar subárboles enteros, sustituyéndolos por una hoja que evalúa en el valor más frecuente dentro de las instancias utilizadas para crear ese subárbol. Por lo general, se utiliza un conjunto de validación o un test estadístico para estimar si efectivamente la poda mejora el resultado.

Por otra parte, cada rama del árbol puede verse como una regla de la forma:

$$\textbf{Si } nodo_1 \text{ es } valor_{1j_1} \text{ y } \ldots \text{y } nodo_i \text{ es } valor_{ij_i} \textbf{ entonces } categoría_p$$

en donde $nodo_1 \ldots nodo_i$ son todos los nodos internos de la rama, $valor_{kj_k}$ es la respuesta según la rama al test del $nodo_k$ y $categoría_p$ es su hoja. En definitiva, a partir de un árbol de decisión se puede obtener un conjunto de reglas, transformando cada una de sus ramas a la regla asociada. Por ejemplo, el árbol de la figura 3.1 se puede traducir como:

- Si «*verbo*» es «OTRO» y «*distancia*» no supera a tres, la etiqueta es «CMII»
- Si «*verbo*» es «OTRO» y «*distancia*» es mayor a tres, la etiqueta es «CMSP»
- ...
- Si «*verbo*» es «DISCURSO», la etiqueta es «CMID»

Las reglas también son susceptibles a una «poda» a los efectos de evitar sobreajustes: cada regla puede ser generalizada quitando alguno de los atributos por los que se pregunta, siempre que esto mejore su precisión estimada.

Varias extensiones se le han hecho al algoritmo básico, entre las que se encuentran el manejo de rangos para atributos numéricos y continuos (como el atributo «*distancia*» del ejemplo), la utilización de atributos con distintos costos o prioridades, heurísticas para la reducción del sobreajuste, etc. La versión del algoritmo que se utiliza en este trabajo, *C4.5*, incluye estas mejoras. (el código puede encontrarse en [47])

3.1.2. Boosting

Informalmente, la técnica de *boosting* consiste en combinar muchos clasificadores simples, llamados *hipótesis débiles*, para obtener un clasificador más complejo [55, 56]. Las hipótesis débiles pueden ser árboles de decisión, redes neuronales o cualquier otro algoritmo de aprendizaje.

Las hipótesis débiles se obtienen en forma secuencial utilizando un *aprendiz débil* sobre el conjunto de entrenamiento. El aprendiz débil entrena poniendo énfasis en los ejemplos sobre los cuales las hipótesis obtenidas hasta el momento cometen más errores. En otras palabras, la nueva hipótesis se concentra en clasificar correctamente aquellas «zonas» del conjunto de entrenamiento donde las hipótesis predecesoras fallan. No todas las hipótesis débiles tienen igual peso en el resultado final: el clasificador se obtiene de una combinación lineal de todas ellas.

En el cuadro 3.2 se encuentra el pseudocódigo del algoritmo *AdaBoost:MH*, un algoritmo de *boosting* para la clasificación con múltiples etiquetas (extraído de [56]). En este algoritmo, $\mathcal{X}$ representa al dominio de las instancias, mientras $\mathcal{Y}$ es el conjunto de clases. Debido a que se admite que una instancia pueda estar en más de una posible clase, el conjunto de entrenamiento está formado por elementos (x_i, Y_i) en donde $x_i \in \mathcal{X}$ y $Y_i \subseteq \mathcal{Y}$. Cada Y_i puede verse como una función del conjunto de etiquetas hacia el rango $[0 \dots 1]$: $Y_i[l]$ es la confianza de que la instancia x_i pertenezca a la clase l.

D_t es una distribución sobre el conjunto de entrenamiento. A medida que se ejecuta el ciclo del algoritmo, la distribución se modifica de forma que tengan una mayor probabilidad de ocurrencia aquellas instancias y clases mal clasificadas. De esta forma, los ejemplos más «complicados» tienen cada vez mayor peso durante el entrenamiento, haciendo que el algoritmo se vea «forzado» a generar hipótesis débiles que resuelvan correctamente su clasificación. El factor Z_t garantiza que, al actualizar D_{t+1}, lo que se obtenga sea efectivamente una distribución.

La salida es la combinación lineal de las hipótesis débiles obtenidas, ponderadas por el factor α_t correspondiente. Este factor se elige en cada iteración buscando disminuir el error de clasificación.

En este trabajo se utilizan como hipótesis débiles reglas que preguntan por un único atributo para etiquetar una instancia (*decision stumps*). A modo de ejemplo, para el clasificador de comas serían de la siguiente forma:

- Si «*primera*» es «VERDADERO», la categoría es «CMMO»
- Si «*última*» es «VERDADERO», la categoría es «CMIF»
- Si «*verbo*» es «DISCURSO», la categoría es «CMID»
- Si «*verbo*» es «OTRO», la categoría es «CMIF»
- ...

en donde, para los atributos discretos (etiquetas morfosintácticas, patrones, tipo de verbos presentes, etc.) se pregunta si éste toma un valor determinado, mientras que para el atributo continuo (la distancia a la siguiente coma) se pregunta si supera o no cierto valor establecido.

A diferencia de los árboles de decisión, la salida de esta familia de algoritmos no es necesariamente «legible». Un algoritmo como *AdaBoost.MH* puede necesitar miles de hipótesis débiles para lograr un buen resultado final. Inclusive tomando aprendices simples como los «*decision stumps*», la gran cantidad de reglas y el peso que efectivamente tiene cada una de éstas en la clasificación de una nueva instancia hacen que, en definitiva, el algoritmo actúe en la práctica como una caja negra.

Los detalles de *AdaBoost.MH* y otros algoritmos de *Boosting* pueden encontrarse en el trabajo de Schapire et al. [55, 56]. *Boostexter*, una implementación de estos algoritmos especializada en la categorización de texto, se puede obtener

Dado el siguiente enunciado

(3.1) Horario central es una obra para cinco personajes, un poco loca, pero a mí me gusta deformar lo que aparenta ser normal.

estos son los valores de los atributos calculados para la primera coma:

Conector: NO	$\text{Cat}_{anterior}$: NINGUNA	$\text{Cat}_{posterior}$: CONJ
$\text{Verbo}_{anterior}$: OTRO	$\text{Verbo}_{posterior}$: OTRO	Distancia: 4
Patrón: NINGUNO	Primera: SÍ	Última: NO
Cat_{-5}: DET	Cat_{-4}: NOMC	Cat_{-3}: PREP
Cat_{-2}: DET	Cat_{-1}: NOMC	Cat_{+1}: DET
Cat_{+2}: ADV	Cat_{+3}: ADJ	Cat_{+4}: CM
Cat_{+5}: CONJ		

y estos para la segunda:

Conector: NO	$\text{Cat}_{anterior}$: NOMC	$\text{Cat}_{posterior}$: NINGUNA
$\text{Verbo}_{anterior}$: OTRO	$\text{Verbo}_{posterior}$: OTRO	Distancia: 1
Patrón: NINGUNO	Primera: NO	Última: SÍ
Cat_{-5}: NOMC	Cat_{-4}: CM	Cat_{-3}: DET
Cat_{-2}: ADV	Cat_{-1}: ADJ	Cat_{+1}: CONJ
Cat_{+2}: PREP	Cat_{+3}: PRON	Cat_{+4}: PRON
Cat_{+5}: VERBFIN		

Cuadro 3.3: Ejemplos de cálculos de atributos

en la página de Schapire en el sitio *web* de la Universidad de Princeton[3].

3.2. Conjunto de entrenamiento

El clasificador se construye para actuar entre el analizador morfosintáctico y el analizador sintáctico de superficie. Por tanto, los atributos en base a los cuales se describe cada una de las comas presentes en el texto deben ser determinados a partir de la salida del primero de éstos o directamente a partir del texto de entrada.

En primer lugar, se establece que todos los atributos considerados sean calculados utilizando como contexto únicamente las palabras que ocurren en la oración donde la coma está presente. Esto resulta natural si se tiene en cuenta que este signo de puntuación tiene un alcance «local» a la oración.

Luego, se realizan sucesivas pruebas para determinar las características a ser tomadas como entrada por los clasificadores. Durante este proceso, se experimenta agregando y modificando los atributos, observando cómo estas variaciones

[3]http://www.cs.princeton.edu/∼schapire (último acceso en enero de 2005)

influyen en el proceso de aprendizaje.

Como punto de partida se toman en cuenta únicamente las categorías morfosintácticas determinadas por el analizador. Los resultados obtenidos no son buenos: el clasificador inducido por el algoritmo C4.5 presenta una tasa de error cercana al 50 %.

En consecuencia, se intenta agregar nuevos atributos que, sin requerir de cálculos costosos, mejoren el proceso de aprendizaje. Por ejemplo, se agregan atributos posicionales: ¿es la coma la primera de la oración?, ¿es la última?, etc. Los primeros resultados son positivos: estos atributos simples mejoran el clasificador resultante, provocando una caída de aproximadamente cinco puntos en la tasa de error.

Se decide, entonces, ampliar aún más el conjunto de características. Algunos de estos nuevos atributos se inspiran en los elementos que son considerados al momento de establecer manualmente la función que cumple cada una de las comas en los textos del corpus.

Por ejemplo, se constata que existen algunos fenómenos simples que permiten, sin un análisis profundo por parte del lector, determinar qué función cumple cada coma en un texto: la presencia de un conector discursivo al comienzo de la oración, la repetición de un patrón simple —nombres comunes, adjetivos, etc.— como indicador de serie, etc.

Sin embargo, se debe tener en cuenta que, como no se realiza ningún análisis sintáctico profundo —no tiene ningún sentido hacerlo, puesto que se perdería la finalidad de contar con la evaluación de las comas para realizar el análisis de superficie— y aún menos una interpretación semántica de la oración, hay factores considerados por una persona al momento de la clasificación que no son trasladables de forma automática a un atributo.

Los resultados de los experimentos parecen indicar que, si bien el uso de estos atributos logra una mejora sustantiva en el clasificador, existe un tope en el proceso de aprendizaje que ronda el 40 % de error[4].

Lamentablemente, no es posible realizar una comparación efectiva entre las tasas obtenidas en cada prueba, debido a que entre ellas varía la categorización o el analizador morfosintáctico empleados. En el apéndice B se detalla la secuencia de atributos probados y los resultados obtenidos en cada uno de estos experimentos.

A continuación se describe el conjunto de atributos utilizado para la construcción del clasificador final, explicando el por qué de su elección. En el cuadro 3.3 se ejemplifica su cálculo en las comas de una oración del corpus.

3.2.1. Etiquetas morfosintácticas

Se establecen diez atributos que cubren una ventana de etiquetas morfosintácticas: las cinco categorías anteriores («cat_{-5}», ..., «cat_{-1}») y las cinco posteriores («cat_{+1}», ..., «cat_{+5}»). Estos atributos se toman directamente de la salida del analizador morfosintáctico, por ende, poseen por valor cualesquiera

[4] Esto se mide únicamente para el algoritmo C4.5.

de las categorías en las que éste clasifica. Aunque puede tomarse una ventana más grande —por ejemplo, que cubra toda la oración—, las categorías más alejadas no demuestran tener gran influencia en los resultados prácticos.

3.2.2. Conector

El atributo «*conector*» indica la ocurrencia de un marcador discursivo antes o después de la coma considerada. Determinar el valor de este atributo corresponde, simplemente, a comprobar la existencia de uno de estos marcadores antes o después de la coma. En el apéndice A se encuentra el listado de los marcadores discursivos considerados.

3.2.3. Posiciones

Tres atributos establecen la posición de la coma en la oración y respecto a otras comas o conjunciones:

- *Primera*: este atributo toma el valor «SÍ», cuando es la primera ocurrencia de una coma en la oración; toma el valor «NO» en caso contrario.
- *Última*: ídem a «*primera*», pero en caso de ser la última coma de la oración.
- *Distancia*: mínima distancia, medida en unidades léxicas, hasta la próxima coma o conjunción, ya sea la anterior o posterior; en caso de ser la única coma de la oración, toma el valor arbitrario de cien.

Estos tres atributos pueden ser útiles para detectar, por ejemplo, series simples, dado que sus elementos están relativamente próximos entre sí en la mayoría de los casos, o para detectar modificadores, que usualmente se encuentran al comienzo de la oración, delimitados por la primera coma.

3.2.4. Verbos

Utilizando dos atributos, «*verbo anterior*» y «*verbo posterior*», se establece si hay verbos conjugados antes o después de la coma, con la única condición que ocurran en la misma oración. En caso de existir más de un verbo antes o luego de la coma, sólo se toman aquellos que se encuentren más cercanos a ésta.

Además, en caso de existir, se determina si los verbos son de discurso, dado que los incisos discursivos siempre cuentan con un verbo de esta clase. Luego, ambos atributos toman alguno de los siguientes valores: «NINGUNO», «DISCURSO», «OTRO».

(3.2) El descalabro se instala cuando la inscripción implica el cuerpo; cuando, **dice** Mathis , «la caligrafía **toma** al cuerpo como material en el lugar de la materia mineral».

En el ejemplo 3.2, hay un verbo antes y otro luego de la segunda coma; el primer verbo es discursivo, mientras que el segundo no lo es. En consecuencia, el atributo «*verbo anterior*» toma el valor «DISCURSO», y el atributo «*verbo posterior*», el valor «OTRO».

La lista de los verbos discursivos considerados se encuentra en el apéndice A.

3.2.5. Patrón

Se busca la aparición de algún patrón en la zona que rodea a la coma, entendiéndose por patrón a la repetición de alguna de las categorías dadas por el analizador morfosintáctico. Esto se fundamenta en que las series simples presentan patrones, ya sean unitarios —sustantivos, nombres propios, adjetivos, etc.—, ya sean más complejos.

Por ejemplo, al final del siguiente texto se puede encontrar una serie que presenta un patrón dado por una secuencia de verbos:

(3.3) Si se observa la práctica nacional se verá que, en realidad, sólo hay una desigual distribución de libertades para **detectar**, **denunciar** y **ejercer** la violencia.

Luego, se decide que el atributo tome alguno de los siguientes valores: «NOM» (nombre común), «PROP» (nombre propio), «ADJ» (adjetivo), «VERB» (verbo), «OTRO» o «NINGUNO».

En principio, la búsqueda de patrones se implementa de una forma sencilla:

1. Se determina la categoría posterior a la coma.
2. Se cuenta las veces que aparece esa categoría separada por otras comas y de forma continua a partir de la coma en cuestión, tanto sea hacia adelante o hacia atrás.
3. Se tiene en cuenta, además, que el patrón aparezca luego de una conjunción —en caso de existir— no pudiéndose buscar más allá de ésta.
4. Un procedimiento análogo se realiza con las categorías que preceden a la coma.

El problema con el método anterior surge cuando los posibles elementos de las series cuentan con más de una categoría, siendo, inclusive, heterogéneos:

(3.4) Al terminar el bachillerato ya había escrito **seis novelas** , **un libro de poesías** y **tres ensayos de literatura clásica**.

En este caso, se elige una categoría «representativa» para describir al posible patrón. Ésta se determina buscando la aparición, en orden de preferencia, de una de las siguientes categorías: verbo, nombre común, nombre propio y adjetivo. En caso que no se encuentre ninguna, se etiqueta al posible patrón como «otros».

A modo de ejemplo, si se cuenta con la siguiente secuencia de categorías (se recuadra la coma considerada):

SENT NOMCS [CM] ADJSG NOMCS CONJ NOMCS VERBFIN... [5]

Los posibles elementos de la serie son:

SENT [NOMCS] CM [ADJSG NOMCS] CONJ [NOMCS VERBFIN...]

Las categorías consideradas hacia «adelante» son «adjetivo nombre-común» («ADJSG NOMCS»). En consecuencia, según la prioridad establecida, se busca al patrón «nombre común» («NOM»), el cual ocurre con frecuencia tres (una vez antes y dos veces luego de la coma).

3.2.6. Categoría anterior y posterior

Los atributos «*categoría anterior*» y «*categoría posterior*» toman el valor de las etiquetas morfosintácticas del elemento anterior a la coma precedente y posterior a la siguiente coma respectivamente (en caso de existir).

(3.5) El éxtasis de la muerte se conjuga con la belleza en El pabellón de **oro**, donde el protagonista[,] atrapado y obsesionado por la idea de la perfección, **incendia** y destruye el pabellón admirado.

Para la segunda coma del ejemplo 3.5, el atributo «*categoría anterior*» toma como valor la etiqueta de la palabra *oro*, «NOUN», mientras que «*categoría posterior*» toma como valor la etiqueta de *incendia*, «VERBFIN».

El objetivo de estos atributos es ayudar a determinar si la coma se encuentra delimitando un inciso, ya sea como marcador inicial (en cuyo caso, «*categoría posterior*» posiblemente sea de ayuda) o como marcador final (en este caso, lo sería «*categoría anterior*»).

3.3. Resultados

Para el análisis numérico, se utilizan las medidas de *precisión* y *recuperación*, definidas como:

$$precisión = \frac{nro.\ de\ instancias\ correctamente\ etiquetadas}{nro.\ total\ de\ instancias\ etiquetadas}$$

$$recuperación = \frac{nro.\ instancias\ correctamente\ etiquetadas}{nro.\ total\ de\ instancias}$$

[5] Las categorías deben interpretarse de la siguiente forma: SENT = fin de oración, NOMCS = nombre común singular, CM = coma, ADJSG = adjetivo singular, CONJ = conjunción, VERBFIN = verbo finito.

	Árbol	Reglas	Boostexter
CMII	68,4	66,2	**72,6**
CMIF	51,0	57,7	**64,2**
CMID	**77,8**	**77,8**	72,7
CMSO	55,1	56,6	**61,3**
CMSP	54,8	48,3	**55,1**
CMCO	47,1	69,4	**76,9**
CMMO	52,0	48,3	**64,4**

Cuadro 3.4: Medida-f de los clasificadores obtenidos

Informalmente, la precisión mide cuán confiable es el sistema, dado que es el porcentaje de elementos correctamente etiquetados[6]; en cambio, la recuperación da una idea del cubrimiento o alcance que presenta la solución: estima cuál es el porcentaje de elementos correctamente etiquetados sobre el total que existe en el universo.

Estas medidas se encuentran por lo general en una relación inversa: a medida que una aumenta, la otra disminuye. Por esto, se utiliza una combinación de ambas, la *medida-f*, definida de la siguiente forma:

$$medida\text{-}f = \frac{2 \times recuperación \times precisión}{recuperación + precisión}$$

Las comas del corpus se divide en dos partes: el conjunto de entrenamiento (4724 ejemplos) y el conjunto de verificación (484 ejemplos). En estos conjuntos no se tuvieron en cuenta ni las comas mal utilizadas, ni las comas clasificadas como bipolares, dada la baja frecuencia de aparición de estas últimas en el corpus (1,7 %).

La línea base a tomar como punto de comparación inicial, tal como es usual en el área, es el clasificador más sencillo que se puede construir: consiste en seleccionar sistemáticamente la categoría más frecuente del corpus, la coma de inciso inicial. Éste tiene una precisión sobre todo el conjunto de ejemplos que ronda el 33 %.

Los resultados globales obtenidos —sin discriminar por categoría— revelan una tasa de acierto cercana al 60 % para los árboles de decisión y sus reglas asociadas y del 66 % para el clasificador obtenido con *Boostexter*, prácticamente duplicando la línea base.

Cabe señalar que en ninguno de los métodos los resultados son parejos si se los observa por categoría. Sin embargo, como se detalla en el cuadro 3.4, en todos ellos las comas de inciso de discurso y las comas de inciso inicial se encuentran dentro de las categorías con mejores resultados. Por otra parte, en ninguno de los clasificadores obtenidos la medida-f de las comas de series proposicionales supera

[6] Es la medida inversa al *ruido*, porcentaje de falsos positivos.

el 55,1 %.

Estos resultados reflejan la complejidad de las estructuras involucradas en cada caso. Por ejemplo, los incisos discursivos son fácilmente reconocibles por la presencia de un verbo de discurso, mientras que no existe un patrón tan claro de uso de las comas que son parte de series proposicionales.

El trabajo más cercano con el que comparar los resultados es el etiquetador de van Delden y Gómez, mencionado en la sección 1.1. Este etiquetador consiste en un conjunto de reglas representadas con autómatas finitos, más un algoritmo de filtrado de etiquetas erróneas mediante una matriz de coocurrencia calculada a partir del corpus.

La precisión reportada por los autores sobre textos extraídos del *Penn Treebank 3* es del 95 %, muy por encima de los resultados obtenidos con cualquiera de los métodos aplicados. Sin embargo, hay que señalar que: (a) van Delden y Gómez no inducen sus reglas automáticamente del corpus sino que las generan manualmente; y (b) su sistema realiza un preprocesamiento de la entrada —también con autómatas finitos— en donde se reconocen cláusulas relativas o infinitivas [7].

En su evaluación, los autores toman como línea de base el etiquetador de Brill, al cual entrenan con sus comas etiquetadas. La precisión en este caso es del 56 %, cuatro puntos por debajo de los árboles y las reglas y diez puntos por debajo del clasificador obtenido con *Boostexter*.

Por otro lado, no es posible realizar una comparación directa con el trabajo de Bayraktar et al., dado que su enfoque es totalmente distinto: parten de los patrones encontrados, para luego asignarles un posible valor de la clasificación.

A pesar de esto, y salvando las diferencias entre las clasificaciones propuestas, los resultados en ambos casos son en cierta forma similares: las clases con menor estabilidad en el idioma inglés (cuadro 2.1) tienden a corresponderse con clases que resultan «difíciles» de aprender. Esto parece razonable si se observa que, a menor estabilidad, mayor cantidad de patrones son necesarios para «cubrir» la clase y, por ende, mayor dificultad se encuentra en su aprendizaje. Existe, sin embargo, una diferencia notoria en las comas de incisos discursivos («citas» en su clasificación), sin encontrarse una razón de peso que la explique: se podría pensar que el atributo que indica el tipo de verbo influye en la evaluación, pero como se muestra en las siguientes secciones, este atributo no es determinante en ninguno de los clasificadores obtenidos.

A continuación, se presenta en mayor detalle los resultados obtenidos en la aplicación de cada uno de los algoritmos.

3.3.1. C4.5

Al utilizar C4.5 con el conjunto de entrenamiento como entrada, se obtiene un árbol de 25279 nodos (interiores y hojas), el cual erra con una tasa del 46,3 % sobre el conjunto de verificación. Luego de la poda para evitar el sobreajuste al

[7] Este tipo de análisis se descarta en este trabajo, puesto que la finalidad de clasificar las comas es facilitar el análisis posterior del texto y no utilizarlo como parte de su entrada.

	CMII	CMIF	CMID	CMSO	CMSP	CMCO	CMMO
CMII	**135**	9		11	4		4
CMIF	27	**37**		12	3		
CMID	1	1	**7**	1			
CMSO	28	9		**46**	6		3
CMSP	23	7		2	**34**	1	7
CMCO	7				2	**8**	5
CMMO	11	3	1	3	1	3	**22**

Cuadro 3.5: Matriz de confusión para el árbol de decisión podado

	CMII	CMIF	CMID	CMSO	CMSP	CMCO	CMMO
CMII	**129**	6		9	8	6	5
CMIF	25	**41**		8	2		3
CMID	1	1	**7**	1			
CMSO	33	4		**47**	3		5
CMSP	24	7		7	**28**	1	7
CMCO	3					**17**	2
CMMO	12	4	1	2	1	3	**21**

Cuadro 3.6: Matriz de confusión para las reglas

conjunto de entrenamiento, la cantidad de nodos se reduce a 4404, y la tasa de error es de un 40,3 % (289 comas correctamente etiquetadas). A pesar del gran número de nodos, el árbol tiene profundidad cuatro; esto se explica porque los nodos que preguntan por categorías morfosintácticas abren una rama por cada posible marca[8].

En el cuadro 3.5 se detalla la matriz de confusión para el árbol obtenido. En cada fila se indica la clase correcta de las instancias y qué etiqueta es la asignada. En la diagonal se encuentra, entonces, la cantidad de instancias correctamente clasificadas.

Por otro lado, se realiza el entrenamiento utilizando validación cruzada. Este método consiste en dividir el corpus en varios subconjuntos de aproximadamente el mismo tamaño y por turnos aplicar el algoritmo utilizando un subconjunto para verificar y los restantes para entrenar[9]. En este caso, partiendo el conjunto de entrenamiento en 10 partes, el error promedio de los árboles es de 39,1 %.

Dados los resultados de ambos experimentos, se puede afirmar que la tasa de error del árbol de decisión ronda el 40 %, con una caída de 27 puntos sobre la línea base.

[8] Son 92 etiquetas, luego de simplificado el conjunto de marcas del analizador morfosintáctico.
[9] Cuando se parte en n subconjuntos, se habla de validación cruzada de tamaño n.

	Árbol			Reglas		
	Precisión	Recup.	Medida-F	Precisión	Recup.	Medida-F
CMII	58,2	82,8	**68,4**	56,8	79,1	66,2
CMIF	56,1	46,8	51,0	65,1	51,9	**57,7**
CMID	87,5	70,0	**77,8**	87,5	70,0	**77,8**
CMSO	61,3	50,0	55,1	63,5	51,1	**56,6**
CMSP	68,0	45,9	**54,8**	66,7	37,8	48,3
CMCO	66,7	36,4	47,1	69,4	77,3	**69,4**
CMMO	53,7	50,0	**52,0**	48,8	47,7	48,3

Cuadro 3.7: Precisión, recuperación y medida-f para el árbol y las reglas

Luego de obtenidos los árboles, se transforman en reglas «si... entonces», seleccionándose las 115 reglas que tienen mejor resultado. En este caso, el error obtenido asciende al 40,1 % (290 comas correctamente etiquetadas), prácticamente el mismo valor que se obtiene al podar directamente el árbol. La matriz de confusión para las reglas se presenta en el cuadro 3.6. En el caso de validación cruzada de tamaño diez, la tasa de error promedio es del 38,6 %.

Observar que, a pesar que las tasas de error son similares, el árbol y las reglas se comportan diferente: las reglas clasifican notablemente mejor las comas de conectores y las de inciso final, mientras que el árbol se comporta mejor con las comas de series proposicionales, las de inciso inicial y las de incisos modificadores. En el cuadro 3.7, se detallan las medidas de precisión, recuperación y medida-f para ambos resultados.

Para obtener el árbol y el conjunto de reglas finales se entrena con las 5208 comas del corpus (el conjunto de entrenamiento más el de verificación). Luego de realizada la poda, se obtiene un árbol con 4688 nodos y un conjunto de 243 reglas.

El atributo más discriminante, la raíz del árbol, es «*primera*»; en sus dos subárboles se pregunta por el valor de «cat_{+1}». Estos dos atributos combinados son, entonces, los que más discriminan entre las distintas instancias. Recién en un tercer nivel aparece el resto de los atributos, menos «*patrón*», por el cual no se pregunta en todo el árbol. En la figura 3.2 se puede visualizar los primeros niveles del árbol resultante.

El mejor resultado de las reglas se da con las comas de incisos discusivos. Las reglas parecen basarse fuertemente en el hecho que la coma «discursiva» se utiliza inmediatamente luego de la cita entrecomillada y el verbo de discurso. Contra lo esperado, ninguna de las reglas pregunta por los atributos «$verbo_{anterior}$» y «$verbo_{posterior}$». En el cuadro 3.8 se muestran las tres mejores reglas para este tipo de comas.

En primera instancia, las comas de conectivos se caracterizan por: (a) ocurrir

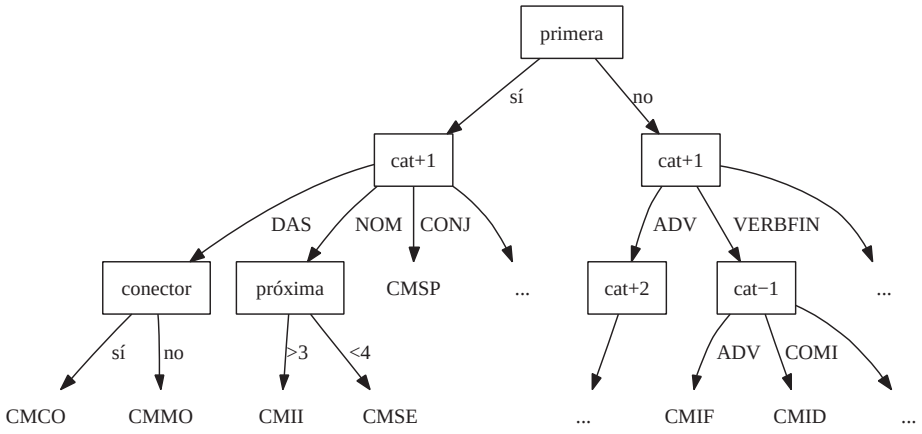


Figura 3.2: Resultado (parcial) de aplicar C4.5

- Si cat_{-1} es «COMILLAS» y cat_{+1} es «VERBFIN» $\Rightarrow$ «CMID»
- Si cat_{-1} es «COMILLAS» y cat_{+1} es «PRON» $\Rightarrow$ «CMID»
- Si primera es «NO», cat_{-1} es «COMILLAS» y cat_{+1} es «ADV» $\Rightarrow$ «CMID»

Cuadro 3.8: Tres reglas para detectar comas de incisos discursivos

preferentemente al comienzo del enunciado y (b) delimitar un marcador discursivo. Esto sí se ve reflejado en las reglas deducidas para etiquetar «CMCO», como se muestra en el cuadro 3.9: en la primera regla se exige la cercanía de un conectivo; en la segunda, se exige además que la coma a etiquetar sea la primera; en la tercera, que la antepenúltima categoría anterior a la coma sea la marca de fin de oración[10], lo cual es, a los efectos prácticos, exigir que sea la primera coma de la oración.

- Si conector es «SÍ» y cat_{-2} es «PREP» $\Rightarrow$ «CMCO»
- Si conector es «SÍ» y primera es «SÍ» $\Rightarrow$ «CMCO»
- Si cat_{-3} es «SENT», cat_{-1} es «NOM» y cat_{+1} es «DET» $\Rightarrow$ «CMCO»

Cuadro 3.9: Tres reglas para detectar comas de conectivos

Las reglas para detectar comas que delimitan series simples confirman que, sin duda, lo que priman son los patrones. Sin embargo, el atributo «*patrón*» no es utilizado por las reglas deducidas. Como se ve en el cuadro 3.10, las cuatro primeras reglas capturan las comas que separan series de nombres (del estilo «Juan, María. . . »); la quinta y la séptima, series con determinantes («el perro, el

[10]Recordar que en el comienzo de una oración siempre hay un marcador «SENT».

gato,...»); y la sexta series de adjetivos («era bella, sencilla, ...»). Todas ellas se basan fundamentales en las categorías léxicas que rodean a la coma a clasificar.

Esto último pareciera indicar que la falta de uso del atributo «patrón» se debe a que el mismo fenómeno que captura —la repetición de ciertas categorías morfosintácticas— se puede inducir directamente de la ventana de etiquetas sin necesidad de realizar un cálculo previo, con el agregado de que estas últimas son, además, determinantes al momento de decidir por otras categorías.

- Si cat_{ant} es «NOM», cat_{+1} es «DETS» y cat_{+4} es «DETS» ⇒ «CMSO»
- Si cat_{pos} es «NOM», cat_{+1} es «NOM» y cat_{+2} es «CM» ⇒ «CMSO»
- Si próxima es «≤ 3», cat_{+1} es «NOM» y cat_{+3} es «NOM» ⇒ «CMSO»
- Si cat_{pos} es «NOM», primera es «NO» y cat_{+1} es «NOM» ⇒ «CMSO»
- Si cat_{-1} es «ADJS», cat_{+1} es «ADJS» y cat_{+2} es «CM» ⇒ «CMSO»
- Si cat_{pos} es «DET», primera es «NO», cat_{-2} es «DET», cat_{+4} es «DET»⇒ «CMSO»
- Si cat_{-3} es «ADJS» y cat_{+1} es «ADJS» ⇒ «CMSO»

Cuadro 3.10: Siete reglas para detectar comas de series simples

Las reglas para detectar comas de inciso final e inicial —cuadro 3.11 y 3.12 respectivamente— utilizan no solo las «ventanas» de categorías léxicas sino el atributo «*primera*» y las categorías de los atributos «cat_{ant}» y «cat_{pos}». Ambos grupos de reglas parecen capturar las mismas reglas de escritura. En particular, hay dos reglas, las primeras de cada etiqueta, que preguntan la categoría que está antes del posible inciso: si esta categoría es un pronombre relativo, entonces lo marcado por las dos comas es un inciso. Estas reglas pueden deberse a los incisos de la forma «...quien, $\boxed{\textit{inciso}}$, estuvo...».

- Si cat_{-1} es «PRONREL» ⇒ «CMII»
- Si primera es «SÍ», cat_{-1} es «VERBFIN» y cat_{+2} es «CM» ⇒ «CMII»
- Si cat_{pos} es «VERBFIN» y cat_{+2} es «CM» ⇒ «CMII»
- Si primera es «SÍ» y cat_{+1} es «PRONREL» ⇒ «CMII»
- Si primera es «NO» y cat_{-1} es «CONJ» ⇒ «CMII»
- Si cat_{pos} es «PRON» y cat_{+2} es «CM» ⇒ «CMII»
- Si cat_{-2} es «VERBFIN» y cat_{+1} es «DETINDEF» ⇒ «CMII»

Cuadro 3.11: Siete reglas para detectar comas de inicio de inciso

Las reglas de inciso inicial reflejan que son buenos como indicadores de incisos: (a) las comas antes de los verbos, dado que esto no es válido a menos que haya

una serie o un inciso (tercera regla); (b) los pronombres relativos luego de la coma (cuarta regla), que parecen indicar la ocurrencia de incisos especificativos; (c) que haya una conjunción inmediatamente antes de una coma (quinta regla) —en el caso de las proposiciones, es la coma la que se coloca antes de la conjunción—.

Las reglas para el etiquetado de final de inciso son, de alguna forma, análogas a las de inciso inicial. Además de la mencionada primera regla, nuevamente se marca que (a) no puede haber una coma entre el grupo nominal y el verbo a menos que sea de inciso (segunda regla) y (b) una conjunción inmediatamente antes de la coma anterior es un indicador de inciso (cuarta regla). La tercera regla indicaría que un complemento del verbo (preposición más un grupo nominal) sólo puede estar separado del verbo por un inciso. La última regla considera que un adverbio entre comas es un inciso.

- Si cat_{ant} es «PRONREL» $\Rightarrow$ «CMIF»
- Si primera es «NO», cat_{-1} es «NOMP» y cat_{+1} es «VERBFIN» $\Rightarrow$ «CMIF»
- Si cat_{-2} es «CM», cat_{+1} es «PREP» y cat_{+5} es «NOMC» $\Rightarrow$ «CMIF»
- Si cat_{ant} es «CONJ», primera es «NO» y cat_{-1} es «NOMC» $\Rightarrow$ «CMIF»
- Si cat_{-2} es «CM» y cat_{-1} es «ADV» $\Rightarrow$ «CMIF»

Cuadro 3.12: Cinco reglas para detectar comas de final de inciso

Las comas de separación de series proposicionales son de las más difíciles de aprender. A pesar que las reglas obtenidas no tienen una precisión marcadamente menor que las reglas de otras categorías —las de inicio de inciso, en este sentido, se comportan aún peor—, su capacidad de «recuperación» es marcadamente menor.

- Si cat_{-2} es «VERBINF» y cat_{-1} es «PRONINDEF» $\Rightarrow$ «CMSP»
- Si cat_{-3} es «VERBFIN» y cat_{+1} es «PRON» $\Rightarrow$ «CMSP»
- Si verbo_{ant} es «OTRO», primera es «SÍ» y cat_{+1} es «VERBFIN» $\Rightarrow$ «CMSP»
- Si cat_{+1} es «CONJ» $\Rightarrow$ «CMSP»

Cuadro 3.13: Cuatro reglas para detectar comas de series proposicionales

En el cuadro 3.13 se pueden observar algunas de las reglas obtenidas para esta etiqueta. La regla más clara es la última de las cuatro; en particular, esta regla es análoga a la de los incisos pero cuando la coma se encuentra inmediatamente luego de la conjunción y no antes.

3.3.2. Boostexter

Se entrena con *Boostexter* utilizando las 4724 instancias del conjunto de entrenamiento. El conjunto de verificación se utiliza para evitar el sobreajuste: el

	CMII	CMIF	CMID	CMSO	CMSP	CMCO	CMMO
CMII	**131**	7	1	15	6	2	1
CMIF	13	**51**		9	4		2
CMID	1		**8**				1
CMSO	23	8	3	**53**	4		1
CMSP	19	10		3	**35**		7
CMCO	4					**15**	3
CMMO	7	4		1	4		**28**

Cuadro 3.14: Matriz de confusión para el clasificador

	Precisión	Recup.	Medida-F
CMII	66,2	80,4	72,6
CMIF	63,8	64,6	64,2
CMID	66,7	80,0	72,7
CMSO	65,4	57,6	61,3
CMSP	66,0	47,3	55,1
CMCO	88,2	68,2	76,9
CMMO	65,1	63,6	64,4

Cuadro 3.15: Precisión, recuperación y medida-f del clasificador basado en *boosting*

entrenamiento se detiene cuando la tasa de error sobre este conjunto deja de disminuir y empieza a aumentar.

El clasificador construido por este método consta de aproximadamente 6800 reglas simples y comete un error del 33,7 % sobre el conjunto de verificación (27 % sobre el conjunto de entrenamiento). La matriz de confusión para el clasificador obtenido se muestra en el cuadro 3.14. Para medir el error se toma en cuenta únicamente una etiqueta, aquella a la cual el clasificador asigne el mayor índice de confianza.

En el cuadro 3.15 se observa la precisión, recuperación y medida-f del clasificador y la medida-f del árbol y las reglas obtenidas con C4.5. El nuevo clasificador supera ampliamente a sus predecesores: según la medida-f, en todas las categorías se comporta mejor, salvo en el marcado de comas de inciso de discurso debido a la menor precisión obtenida en esta clase. Por otro lado, es notable la mejora en el etiquetado de comas de modificadores.

Dentro de las 6800 reglas, existen varias que preguntan por el mismo atributo y valor. El algoritmo guarda un clasificador «condensado», en donde cada pregunta aparece una única vez, combinando los pesos de todas las reglas. En total se generan 388 reglas distintas.

En el cuadro 3.16 se muestran las diez reglas más influyentes, ordenadas de

	CMII	CMIF	CMID	CMSO	CMSP	CMCO	CMMO
cat_{+1} es PREP	0,337	0,173	0	0,142	0,231	0,110	0,186
	0,661	0,046	0,035	0,066	0,079	0,078	0,020
cat_{-1} es COMILLAS	0,364	0,573	0	0,268	0,368	1	0,821
	0,224	0,406	1	0,374	0,371	0	0
cat_{+1} es PRONREL	0	0,642	0,873	0,408	0,665	0,522	0,945
	1	0,317	0	0,426	0,152	0,182	0
cat_{+1} es CONJ	0,818	0,278	0,911	0,199	0,151	0,903	0,786
	0,128	0,376	0	0,613	0,883	0	0
cat_{-4} es SENT	0,521	0,558	0,069	0,421	0,605	0,078	0,234
	0,520	0,442	0,915	0,275	0,414	0,892	0,120
cat_{+1} es VERBFIN	0,958	0,165	0	1	0,360	0,361	0,275
	0,052	0,638	1	0	0,621	0,257	0,599
cat_{-1} es PRONREL	0	0,854	0,510	0,967	1	0,631	0,873
	1	0	0	0	0	0	0
cat_{-2} es CM	1	0,132	0,890	0,155	0,465	0,270	0,695
	0	0,827	0	0,543	0,082	0,209	0
cat_{+1} es PREGUNTA	0,967	0,523	0,381	0,886	0	0,189	0
	0	0	0	0	1	0,532	1
cat_{ant} es PRONREL	0,980	0	0,592	0,790	1	0,506	0,991
	0	1	0	0	0	0	0

Cuadro 3.16: Las diez reglas más influyentes

forma descendente. La votación es un número entre cero y uno para cada clase; cuanto más cerca de uno se encuentre el valor para una clase, mayor confianza tiene el clasificador simple de que esa sea la clase correcta. Para cada pregunta hay dos renglones: el primero, la votación por la respuesta negativa; el segundo, la votación por la respuesta afirmativa.

Aunque es difícil determinar cómo influye cada atributo en la clasificación de una instancia, es posible observar que se repiten muchos de los indicadores que surgen en las reglas del algoritmo C4.5. Por ejemplo, en la segunda regla, la presencia de comillas inmediatamente antes de un inciso es una buena señal que la coma marca un inciso discursivo y su ausencia lo contrario (se observa un fenómeno similar, pero de signo contrario, con las comillas y las comas que marcan conectores discursivos). La presencia de una coma seguida de un pronombre relativo nuevamente indicaría el comienzo de un inciso (tercera regla), y su presencia antes de la coma el comienzo de un inciso (octava regla); en este último caso, se inhiben todas las otras categorías. Reafirmando el resultado anterior, la última regla establece que, si la categoría que precede a la coma anterior es un pronombre relativo, se está frente al final de un inciso.

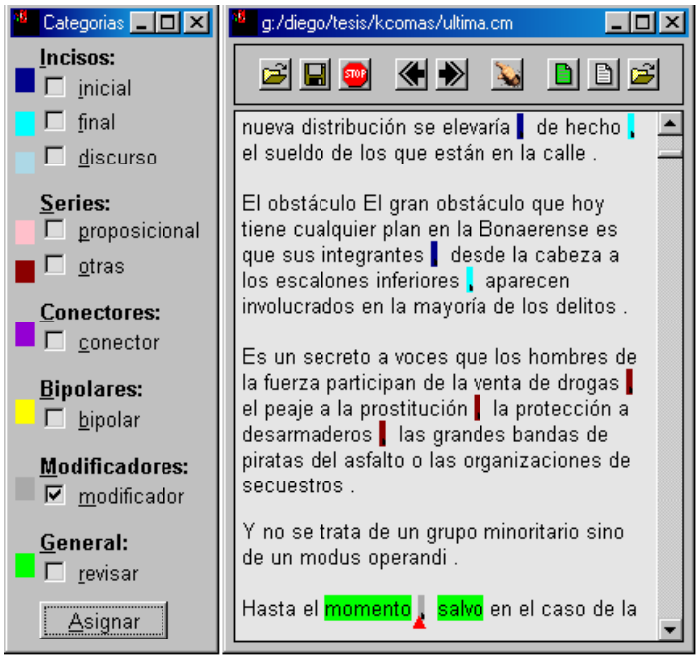


Figura 3.3: Interfaz para el etiquetado y generación del conjunto de entrenamiento

3.4. Herramienta para el entrenamiento

Como se menciona en las secciones anteriores, ambos algoritmos, debido a su calidad de supervisados, requieren de un conjunto de entrenamiento etiquetado. Se desarrolla una herramienta con el objetivo de facilitar las tareas de etiquetado, entrenamiento y evaluación de resultados. Esta herramienta permite:

- Visualizar los textos con las comas resaltadas, permitiendo etiquetarlas con múltiples categorías.
- «Navegar» por el texto en función de sus comas: ir de una coma a la siguiente o a la anterior, ir hasta una coma marcada como de clasificación dudosa, etc.
- Calcular los distintos atributos —posición, patrones, etc.— y generar el conjunto de entrenamiento según el formato requerido por cada uno de los algoritmos.
- Aplicar el clasificador aprendido sobre un texto cuyas comas se encuentran sin clasificar y ver el resultado de la evaluación en el propio texto.

El lenguaje utilizando para su implementación es *Prolog* [16, 32, 42]. La principal ventaja de este lenguaje es que permite un rápido prototipado, además de un acoplamiento con el intérprete de las reglas contextuales, que también está realizado en este lenguaje (ver el capítulo 4). Específicamente, se utilizó SWI-Prolog

como intérprete [60, 64, 65] y la biblioteca XPCE [66] para la implementación de la interfaz gráfica.

Resumen

Para obtener un clasificador de comas, se aplican dos técnicas de aprendizaje automático: árboles de decisión y *boosting* con reglas simples. La primera cuenta con la ventaja que se pueden obtener una solución «legible», en forma de un conjunto de reglas «si... entonces...»; mientras que la segunda, un mejor resultado.

Ninguno de los clasificadores obtenidos presentan muy buenas tasas de precisión y recuperación. Sin embargo, los métodos estadísticos parecen confirmar algunos fenómenos lingüísticos que caracterizan a los incisos, las series, etc.

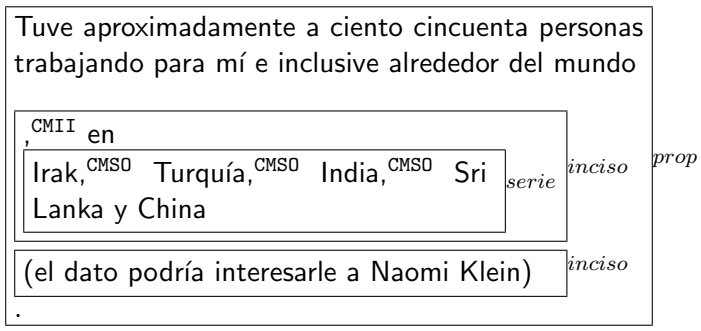
Ambas técnicas utilizadas requieren de un conjunto de instancias correctamente etiquetadas. Para tal fin, se desarrolla una herramienta en *Prolog*, que permite etiquetar las comas, calcular el valor de los atributos, generar el conjunto de entrenamiento y ver el resultado del clasificador sobre un nuevo texto.

Capítulo 4

Analizador sintáctico

¿Es la información brindada por el clasificador de alguna utilidad a pesar de su tasa de error? Para intentar contestar esta pregunta, se plantea como objetivo implementar un analizador sintáctico de superficie que logre determinar algunas estructuras presentes en el texto —series proposicionales, series simples, incisos, etc.— indicadas, principalmente, por sus comas, las cuales son etiquetadas con el clasificador obtenido con *Boostexter*. Algunas de las estructuras buscadas se muestran en el ejemplo 4.1, sobre un texto con sus comas ya clasificadas.

(4.1) a. Tuve aproximadamente a ciento cincuenta personas trabajando para mí e inclusive alrededor del mundo,$^{\texttt{CMII}}$ en Irak,$^{\texttt{CMSO}}$ Turquía,$^{\texttt{CMSO}}$ India,$^{\texttt{CMSO}}$ Sri Lanka y China (el dato podría interesarle a Naomi Klein).

b. 

En principio se desea aprovechar al máximo la información brindada por el clasificador de comas. Sin embargo, dado que el clasificador no es totalmente confiable, el análisis también se apoya en otros fenómenos del texto que permiten «sortear» los errores de clasificación.

El analizador construido se basa en el formalismo de *reglas contextuales* y se implementa utilizando un intérprete para estas reglas realizado en el marco del proyecto Clatex [67, 68].

Este capítulo se divide en tres secciones: en la sección 4.1 se describen el formalismo y su intérprete; las reglas para el reconocimiento de las distintas

estructuras, en la sección 4.2; finalmente, la evaluación del analizador se encuentra en la sección 4.3.

4.1. Reglas contextuales

El formalismo de las *reglas contextuales* tiene como objetivo el análisis de porciones de texto, con su consecuente etiquetado. El análisis se realiza a partir del propio texto a ser etiquetado, parte del texto que sucede antes, el *contexto izquierdo*, y parte del que sucede después, el *contexto derecho*.

Cada regla determina qué secuencia de elementos —texto, signos de puntuación y etiquetas— debe estar presente para ser etiquetada por ésta. Estos elementos no tienen por qué ser contiguos: el formalismo permite «subespecificar» la secuencia estableciendo *zonas de exclusión*, esto es, porciones de texto de largo acotado[1] en las cuales no deben ocurrir ciertos elementos.

Este formalismo se encuentra fuertemente motivado por las necesidades que plantea el reconocimiento de proposiciones: (a) realizar una especificación parcial del texto a reescribir; (b) condicionar esta reescritura al contexto y (c) expresar el anidamiento de proposiciones a través de la recursión. Estas tres características hacen que este sistema de reescritura sea equivalente, en principio, a las gramáticas sensibles al contexto [27].

La aplicación de reglas contextuales se demuestra exitosa en el reconocimiento de marcadores discursivos [46] y, en el proyecto Clatex, en la detección de proposiciones en textos en francés y español [12, 13, 67].

4.1.1. Definición

Sea V un conjunto finito de símbolos. Una regla sobre V es una expresión de la forma:

$$A \rightarrow ContextoIzq \backslash Cuerpo \, / ContextoDer; \\ EspecifConjuntos$$

donde:

- $A \in V$.
- Una *zona de exclusión* es una expresión de la forma $*(ConjExc, n)$, donde $ConjExc$ es el nombre de un conjunto de elementos de V y n un natural que refiere al tamaño máximo, medido en unidades léxicas, de una porción de texto.
- $ContextoIzq$, $Cuerpo$ y $ContextoDer$ son tiras de símbolos de V o zonas de exclusión.
- $Cuerpo$ no puede ser vacío; en cambio, $ContextoIzq$ y $ContextoDer$ sí pueden serlo.

[1] Medido en unidades léxicas.

- La *condición* de la regla es la tira $ContextoIzq.ContextoDer$. En la condición, toda zona de exclusión debe estar precedida y seguida de un símbolo de V.
- $EspecifConjuntos$ es la definición, por enumeración, de los conjuntos mencionados en las zonas de exclusión de la regla.

Por ejemplo, para encontrar un «inciso modificador» al comienzo de la oración a partir de una coma marcada como «CMMO», se puede plantear una regla de la forma:

$$\begin{aligned} (4.2) \qquad & inciso_{modif} \rightarrow \text{SENT} \setminus *(NoCm, 10) \ / \ \text{CMMO} \ ; \\ & NoCm = \{\text{SENT, CMII, CMIF, CMID, CMSO, CMSP, CMMO, CMBI, CMCO}\} \end{aligned}$$

Esta regla indica que debe marcarse una zona del texto como «inciso modificador» si no tiene más de diez unidades léxicas, hay un coma del tipo «modificador» como contexto derecho, hay un límite de oración como contexto izquierdo y en la zona a marcar no ocurre ni una coma, ni otro límite de oración.

Aplicada al ejemplo 4.3, la regla marcaría como modificador «*Hasta hoy*» [2]:

(4.3) Hasta hoy ,CMMO pasados diez meses,CMMO no se sabe por qué mataron a Piazza, pero todos los indicios van en dirección a un crimen policial.

Por otra parte, la regla no marca al texto «*Hasta hoy, pasados diez meses*» porque, a pesar que se encuentra enmarcado en el contexto adecuado, la ocurrencia de la primera coma está excluida por la zona $*(NoCm, 10)$.

En principio, el formalismo sólo permite elementos atómicos en el conjunto de etiquetas V. Sin embargo, existe una extensión en la cual se admiten términos estructurados, expresados como una lista de elementos atributo–valor:

$$[atributo_1 = valor_1, \ldots, atributo_n = valor_n]$$

Por ejemplo, las comas de tipo «inciso discursivo» se representan con la etiqueta «[categoría=CM, tipo=inciso, subtipo=discursivo]». Ahora, una etiqueta satisface la condición impuesta en una regla si, de sus atributos, los que aparecen especificados en la regla satisfacen las restricciones impuestas por ésta.

Esto permite reescribir de forma más compacta al conjunto $NoCm$ del ejemplo 4.2, aplicando que todas las comas —y nada más que éstas— satisfacen la restricción «[categoría=CM]»:

$$NoCm = \{\ [\text{categoría=SENT}],\ [\text{categoría=CM}]\ \}$$

[2] Al comienzo de la oración hay siempre un límite de oración; éste se encuentra dado por el final de la oración anterior, salvo en la primera oración del texto, para la cual siempre se agrega esta marca.

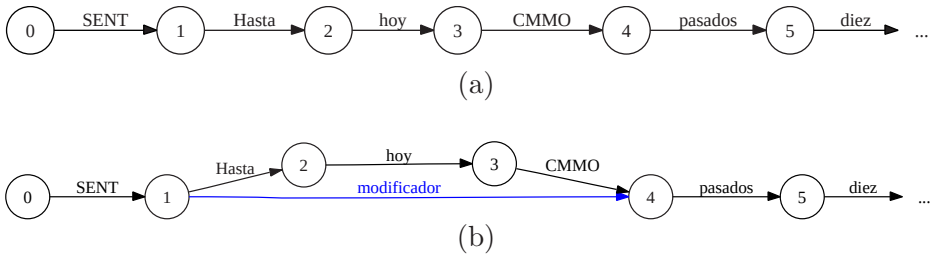


Figura 4.1: Grafo de texto para el ejemplo 4.3 (a) antes y (b) luego de aplicada la regla 4.2.

En las reglas se pueden establecer elementos opcionales, con el operador *op()*, los cuales no condicionan la aplicación de una regla y permiten escribir un módulo utilizando menos reglas. En el siguiente ejemplo, las dos reglas en (a) se pueden condensar en la regla de (b):

(4.4)
a. $gnom \rightarrow \backslash$ DET NOM /;
$gnom \rightarrow \backslash$ DET ADJ NOM /;
b. $gnom \rightarrow \backslash$ DET op(ADJ) NOM /;

Finalmente, se destaca que, a pesar que el formalismo recuerda a las reglas de las gramáticas generativas (libres de contexto, irrestrictas, etc. [27]), las reglas contextuales tienen como único objetivo el etiquetado de texto, resultante de un análisis posiblemente parcial de éste.

4.1.2. Intérprete

El intérprete de las reglas contextuales trabaja con conjuntos finitos de reglas, llamados *módulos*. Dentro de un módulo pueden coexistir más de una regla con la misma etiqueta del lado izquierdo; esto se entiende como una disyunción de las reglas: expresan todas las distintas maneras en que una porción de texto puede ser marcada con la misma etiqueta.

El analizador aplica los distintos módulos en pasadas secuenciales sobre el texto. De esta manera, las marcas realizadas por un módulo forman parte de la entrada en el análisis con los módulos subsiguientes.

Para representar el texto y las etiquetas que se le agregan, el intérprete utiliza un grafo dirigido. Cada posición del texto tiene unívocamente asociada un nodo, y cada marca entre las posiciones n y m del texto se corresponde con la etiqueta de una arista que une al nodo n con el m. En la figura 4.1 se muestra parte del grafo correspondiente al texto del ejemplo 4.3 y el resultado al aplicar la regla 4.2.

El análisis se realiza utilizando la técnica de *right-corner chart-parsing*[3]. El intérprete procesa el texto de izquierda a derecha, moviéndose una posición por vez. Para cada etiqueta que finalice en una posición dada, se busca una regla cuyo lado derecho la tenga por última categoría (*right corner*). Luego, se

verifica la existencia (o inexistencia, en caso de zonas de exclusión) del resto de las categorías de su lado derecho. En caso de encontrarlos, se inserta la arista etiquetada con el correspondiente lado izquierdo entre los nodos que marcan el fin del contexto izquierdo y el comienzo del contexto derecho reconocidos. Finalmente, para habilitar que la nueva etiqueta actúe a su vez como disparador de reglas, se retoma el procesamiento de la entrada en el punto donde se terminó el marcado.

Visto en el grafo, el análisis consiste en ir procesando nodo a nodo de forma ascendente. En cada nodo n se revisan las etiquetas de las aristas entrantes, buscando las posibles reglas que tengan sus etiquetas por *right corner*. Dada una posible regla a aplicar, se recorre el grafo «hacia atrás», controlando la existencia de las otras categorías de esta regla. En caso de determinar que efectivamente el lado derecho se encuentra en el texto, el analizador inserta una nueva arista entre los nodos j y k, siendo j y k el fin del contexto izquierdo y el comienzo del contexto derecho respectivamente ($j < k \leq n$). El analizador retoma el procesamiento en el nodo k.

Existen varias extensiones disponibles en el intérprete, dos de las cuales son utilizadas en este trabajo: (a) la priorización de reglas dentro de un módulo; y (b) la posibilidad de especificar elementos opcionales.

Según el algoritmo visto, en cada posición, todas las reglas cuyo *right corner* coincide con la etiqueta son sistemáticamente probadas. Cada una de las etiquetas exitosas genera un arco en el grafo. Al establecer prioridades, el intérprete intenta primero etiquetar con las reglas de mayor prioridad. La primera regla exitosa inhibe la aplicación de todas las otras —esto es, las de menor prioridad que ella—. Establecer prioridades permite una ejecución más eficiente cuando de antemano se conoce que las reglas son autoexcluyentes. Además, permite escribir de un modo más compacto, al evitar explicitar en una regla las «diferencias» que determinarían su no aplicación en caso de ser exitosa alguna otra de mayor prioridad.

El intérprete anterior se encuentra implementado en *Prolog* [67], pero también hay otras versiones: una implementada como un sistema deductivo, utilizando *Constraint Handling Rules* [3] [22]; otra con técnicas de estado finito, para módulos de reglas no recursivas [35].

4.2. Reglas propuestas

En esta sección se presenta el analizador basado en los valores de las comas, implementado con reglas contextuales. Como se menciona al comienzo de este capítulo, el objetivo planteado es realizar análisis de superficie de textos utilizando, en la medida de lo posible, la información brindada por el clasificador de comas obtenido con *Boostexter*.

[3] Sistema de restricciones lógicas.

El proceso que se aplica al texto es el siguiente: primero es etiquetado con el analizador lexicográfico; luego, sus comas son evaluadas con el clasificador; finalmente, se procesan por el intérprete de reglas contextuales.

Dado que el clasificador implementado asigna un valor de confianza a cada una de las posibles etiquetas, se elige de todas la que tiene mayor valor. Además, en la construcción del analizador se tiene en cuenta que, como consecuencia de la tasa de error que presenta el clasificador, la valoración de las comas no es totalmente confiable. Luego, las reglas requieren además de la presencia o ausencia de otros fenómenos en el texto para confirmar o desestimar la ocurrencia de las estructuras buscadas. Sin embargo, en todo momento se busca minimizar el uso de esta información «extra–coma».

El analizador se implementa con un conjunto de reglas contextuales, utilizando un método heurístico: las reglas se crean a partir de la observación y análisis manual del corpus de trabajo, intentando capturar fenómenos generales a partir de casos particulares.

Una alternativa a este método es utilizar nuevamente técnicas de aprendizaje automático, como por ejemplo realizan Li y Roth para el inglés [31]. Sin embargo, se descarta esta opción debido a que no se cuenta con un corpus en español correctamente analizado y con sus comas ya valuadas (ni con los recursos para hacerlo).

Se decide detectar y etiquetar los siguientes tipos de estructuras:

- *Series simples*: nombres comunes y propios, adjetivos, etc. los cuales, si el clasificador no cometiera errores, deberían encontrase separados por comas del tipo «CMSO».

- *Series proposicionales*: segmentos de texto con un verbo conjugado y delimitados por puntuación[4], posiblemente separados por comas etiquetadas como «CSMP».

- *Incisos*: segmentos de texto correspondientes a incisos en cualquier posición dentro de la oración, delimitados por comas de inciso (inicial, final, modificador, etc.) o comas de conectivos. Se opta por incluir dentro de esta categoría a los conectivos discursivos, dado que tienen un comportamiento similar al de los incisos modificadores —ocurren al comienzo de la oración, y se «acumulan» con los incisos, teniendo, de alguna forma, un «alcance global» a toda la oración—. Esto permite tratarlos de igual forma y a un mismo tiempo que al resto de los incisos.

Notar que no es usual hablar de modificadores y conectivos como incisos. Las estructuras a reconocer se corresponden a dos tipos básicos de comas: las aditivas,

[4]Se utiliza como noción de proposición a un segmento de texto con un verbo conjugado y sus argumentos y modificadores. Además, se incluyen aquellos segmentos cuyos verbos se encuentran elididos.

que separan elementos de series, y las sustractivas, que delimitan segmentos que interrumpen el desarrollo del enunciado.

El proceso de análisis se divide en varias etapas, cada una de ellas implementada con un módulo de reglas contextuales. La ejecución de los módulos se realiza en forma secuencial, comenzando por aquellos que detectan las estructuras locales y con baja «recursividad» —series simples, incisos apositivos, etc.—, y finalizando con las estructuras más complejas, que, por lo general, abarcan o afectan a toda la oración —series proposicionales, incisos iniciales, etc.—.

(4.5) Griffith,CMII vilipendiado por glorificar al Ku-Klux-Klan en El nacimiento de una nación,CMIF contesta a sus «intolerantes» críticos con las cuatro historias simultáneas (la caída de Babilonia,CMSO la Pasión de Cristo,CMSO la Noche de San Bartolomé y una huelga de trabajadores contemporánea),CMSO los sets babilónicos y las escenas de masas de Intolerancia.

A continuación se presentan las distintas etapas en que se divide el análisis. Para ejemplificar su salida, se utiliza el ejemplo 4.5 extraído del corpus.

1. **Series simples**. Este conjunto de 33 reglas detecta la ocurrencia de series simples (nominales, adjetivales, etc.) presentes en el texto. Para realizar el etiquetado, se basan fuertemente en la presencia de comas etiquetadas como CMSO, nombres propios o comunes, etc. Sin embargo, en esta etapa no se tiene en cuenta la ocurrencia de series anidadas[5] u otra estructura que pueden interrumpirla (por ejemplo, incisos).

   Griffith ,CMII vilipendiado por glorificar al Ku-Klux-Klan en El nacimiento de una nación,CMIF contesta a sus «intolerantes» críticos con las cuatro historias simultáneas (
   la caída de Babilonia,CMSO la Pasión de Cristo,CMSO la Noche de San Bartolomé y una huelga de trabajadores contemporánea$_{serie}$
) ,CMSOlos sets babilónicos y las escenas de masas de Intolerancia.

   En el ejemplo, la serie «*las cuatro historias,. . . , las escenas de masas*» no se reconoce por la interrupción producida por el inciso entre paréntesis.

2. **Incisos en texto entre paréntesis o rayas**. El texto que se encuentra entre paréntesis o rayas es marcado como inciso en una etapa posterior. Luego de que esto último ocurre, no se le realiza ningún otro análisis, con lo que cualquier marcado debe realizarse en las primeras etapas.[6].

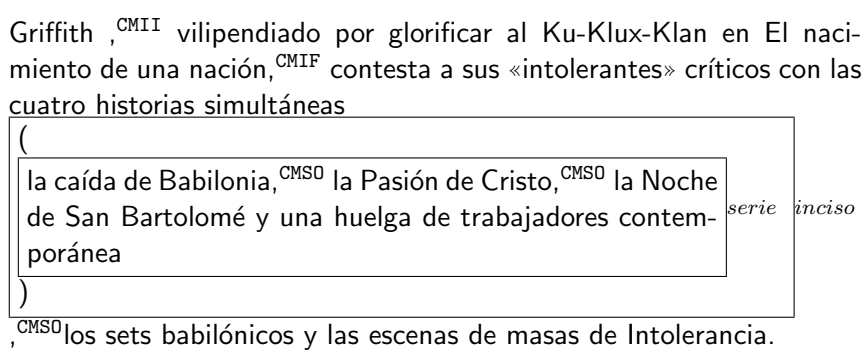
[5] Series que contienen a su vez otras series como elementos.

[6] A modo de ejemplo, el módulo anterior marca las series simples que se encuentran entre paréntesis en el texto 4.5.

Este módulo consiste en un conjunto de 14 reglas que marcan la ocurrencia de incisos dentro de estos textos. Por ejemplo, incisos de la clase «modificadores» delimitados por un paréntesis de apertura y una coma «CMMO».

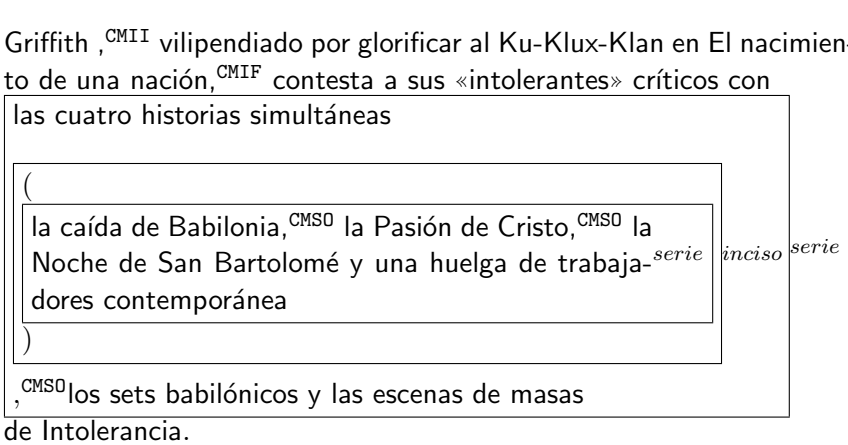
En el ejemplo 4.5 no existe ningún inciso dentro de los paréntesis, con lo que el texto etiquetado permanece incambiado.

3. **Incisos delimitados por paréntesis o rayas**. Durante esta etapa, todo texto entre paréntesis y rayas es etiquetado como inciso utilizando únicamente dos reglas. De esta forma se lo «elimina» del posterior análisis.



Luego, el módulo etiqueta como inciso al texto entre paréntesis en el ejemplo 4.5.

4. **Series simples**. Se ejecuta nuevamente el módulo de series simples, para detectar aquellas series no marcadas en la primera etapa debido a la presencia de interrupciones.



En esta etapa se detecta la serie «*las cuatro historias (. . .) las escenas de masas*», dado que la interrupción del inciso entre paréntesis queda oculta por el etiquetado realizado en la etapa previa.

5. **Estructuras bipolares**. Se marcan las estructuras condicionales de la forma «si. . . , entonces . . . » como dos proposiciones. El módulo consta de tres reglas que buscan subsanar los errores que de forma sistemática comete el evaluador de comas en frases con esta estructura.

   Continuando con el ejemplo, la entrada se ve incambiada al no contener frases bipolares.

6. **Incisos**. Este módulo tienen como objetivo detectar la presencia de incisos en el texto. Se escriben un total de 26 reglas, las que se aplican en tres subetapas: primero se marcan los incisos iniciales (modificadores, conectivos, etc.), luego los discursivos y finalmente el resto de los incisos. Esta decisión se basa en la distinta tasa de error que comete el clasificador y en la facilidad con que se pueden detectar estos incisos, esto último, de forma independiente al valor de la coma.

   Griffith
   ,CMII vilipendiado por glorificar al Ku-Klux-Klan en El nacimiento de una nación,CMIF *inciso*
   contesta a sus «intolerantes» críticos con
   las cuatro historias simultáneas
   (
   la caída de Babilonia,CMSO la Pasión de Cristo,CMSO la Noche de San Bartolomé y una huelga de trabajadores contemporánea *serie* *inciso* *serie*
)
   ,CMSOlos sets babilónicos y las escenas de masas
   de Intolerancia.

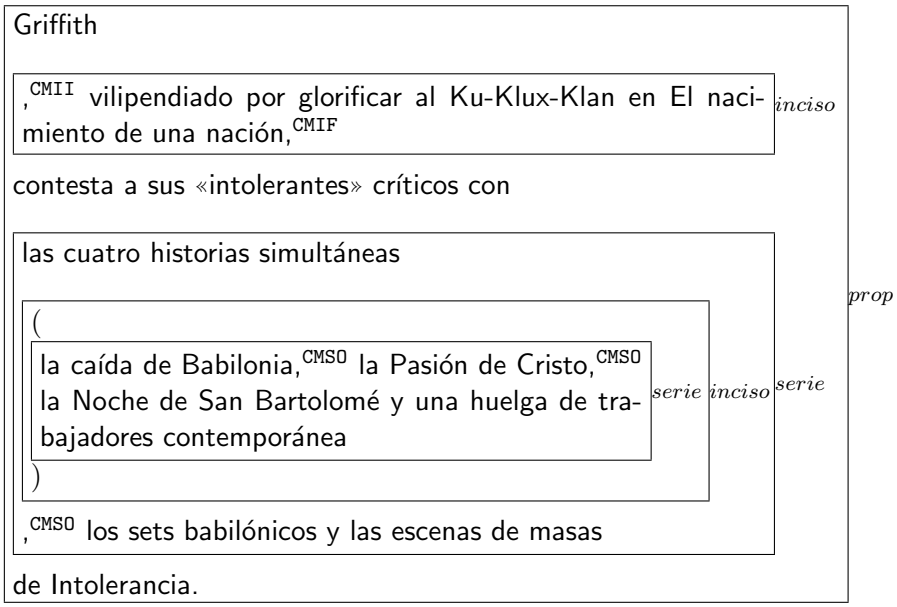
   El texto « *vilipendiado por (. . .) de una nación*» es etiquetado como inciso, debido a que ocurre entre dos comas valuadas como comienzo y final de inciso («CMII» y «CMIF», respectivamente), aunque también se utiliza el hecho que el texto marcado comienza con un participio.

7. **Series proposicionales**. Se escribe un conjunto de 16 reglas con el objetivo detectar proposiciones y las series que forman. Sin embargo, cabe aclarar que no todas las proposiciones son etiquetadas, sino que únicamente se marcan aquellas que ocurren entre signos de puntuación [7].

   En el ejemplo, se termina el análisis de la frase, al detectar el verbo finito «*contesta*» entre el comienzo y el final de la oración, estando todos los signos de puntuación «ocultos» por alguna etiqueta.

[7] A modo de ejemplo, las proposiciones relativas no son detectadas.

8. **Análisis final**. Por último, se ejecuta un módulo con el que se busca completar el análisis del texto con las comas que no fueron utilizadas por los módulos previos. Se intenta, entonces, determinar la posible estructura en base a otros fenómenos, pero tomando en cuenta los posibles errores cometidos por el clasificador. El módulo cuenta con 12 reglas.



En las siguientes secciones se describen las principales reglas que componen cada uno de estos módulos. En las reglas de los ejemplos, para mayor claridad, no se escriben los elementos con la formulación interna del intérprete —una lista de elementos atributo-valor— sino que se utilizan elementos atómicos con sus características relevantes. Así, si el elemento es una coma clasificada como de serie proposicional, se opta por la etiqueta «CMSP» en lugar de *[categoría=coma, tipo=serie, subtipo=proposicional]*.

En el anexo C se encuentra un listado detallado de todas las reglas del sistema, junto con el significado de cada una de las etiquetas mencionadas.

4.2.1. Series simples

Por lo general, las series simples están conformadas por una secuencia de elementos equivalentes, todos éstos separados por comas. La excepción es el último elemento, que no se incorpora a la serie con una coma sino con una conjunción copulativa (*y*, *o*, *ni*, etc.).

(4.6) No se sabe, en realidad, si lo que atrae es *la promesa de galletas tibias, el anisado de unas rosquitas, la hogaza de pan dorada en su punto justo* **o** *todo eso junto*.

En el ejemplo 4.6, se encuentra la serie «*la promesa... o todo eso junto*»; en este caso el último elemento, «*todo eso junto*», se encuentra separado por la conjunción *o*.

Se observa, entonces, que se puede considerar la aparición de una coma valorada como «serie otras» («CMSO») junto a una conjunción en su «cercanía» como un buen indicador del final de una serie simple.

(4.7) a. $\text{serie}_{ac} \rightarrow \text{CMSO} \setminus *(S, 5)$ NOM CONJ NOM /;
$S = \{\text{CONJ, CM, VERBFIN}\}$

b. $\text{serie}_{ac} \rightarrow \text{CMSO} \setminus \text{ETC}/;$

Otra opción es que el final de la serie esté determinado por «etcétera», y a diferencia del caso anterior, la serie no cuenta con una conjunción separando sus últimos elementos.

La reglas en el ejemplo 4.7 se muestra un caso muy simple, en el cual se detectan los «finales» de series de nombres [8], como ser «María, Juan y Pedro», con la regla (a) y «María, Juan, etc.» con la regla (b). Las zonas de exclusión permiten reconocer elementos opcionalmente antepuestos al nombre, por ejemplo adjetivos, que son parte del grupo nominal.

Luego de «detectado» el final, se construye la serie de derecha a izquierda [9]: se agregan otros elementos si se encuentran antes del sector etiquetado y separado por comas de serie.

Continuando con el ejemplo de series de nombres, se crean las siguientes reglas contextuales para «acumular» elementos en la serie:

(4.8) $\text{serie}_{ac} \rightarrow \text{CMSO} \setminus *(S, 5)$ NOM CMSO serie_{ac} /;
$S = \{\text{CONJ, CM, VERBFIN}\}$

Finalmente, cuando no hay más comas de series, se debe incorporar al primer elemento:

(4.9) $\text{serie} \rightarrow \setminus$ NOM CMSO serie_{ac} /;
$S = \{\text{CONJ, CM, VERBFIN}\}$

Se observa que siempre que puede aplicarse la regla 4.8 sobre una porción de texto, también es aplicable la regla 4.9. Sin embargo, sólo debería utilizarse la segunda regla cuando la primera no es exitosa. Este comportamiento se logra asignando mayor prioridad a la regla 4.8 sobre la regla 4.9.

Determinar el comienzo de una serie no es algo simple, como podría deducirse de la regla del ejemplo 4.9. Los elementos medios de estas series están delimitados por las comas, y en esto se basan las reglas de los ejemplos 4.7 y 4.8, las cuales

[8] Para el ejemplo, los nombres se etiquetan con «NOM».

[9] Esta forma de procesar el texto se debe, principalmente, a la forma en que está construido el intérprete de reglas contextuales.

utilizan las comas como contexto izquierdo. Pero ¿hasta dónde se extiende el comienzo o el final de la serie? En estos casos no se cuenta con una categoría que actúe como un límite claro de la estructura.

Una posibilidad es utilizar el criterio de «uniformidad sintáctica» de las series. Por ejemplo, en muchas de las series nominales, o bien todos, o bien ninguno de los elementos comienza con un determinante o cuantificador «***el*** *anisado (...),* ***las*** *rosquitas, (...)* », «*para* ␣ *detectar,* ␣ *denunciar y* ␣ *ejercer...*»). Se agrega, entonces, un atributo a las etiquetas de series que indican la presencia o ausencia de determinantes para luego permitir su incorporación al primer elemento de la serie.

Por otra parte, en el corpus es común encontrar series simples encerradas dentro de un par de paréntesis o rayas. Estos casos muchas veces no presentan una conjunción al final de la serie, sino que son las comas las encargadas de separar a todos los elementos. Por ejemplo:

(4.10) a. Más allá de los compartimientos estancos en que las corporaciones (del tango, del folklore, del rock) gustan dividir a la cultura (...).

b. Los nombres ya clásicos de la última década de música uruguaya —Los Terapeutas, El Cuarteto de Nos, La Trampa o La Tabaré (devenida en milongón banda)— no defraudaron a su público entusiasta.

En el ejemplo 4.11 se detallan las reglas para el marcado de series sin conjunción final y delimitadas (a) por rayas y (b) por paréntesis. Estas series, junto con sus delimitadores, son luego etiquetadas como incisos por el siguiente módulo.

(4.11) a. **series entre rayas**:

$$\text{serie}_{\_} \rightarrow \text{CMSO} \setminus * (S, 10)\ \text{NOMP}\ / — ;\quad S = \{\text{CONJ, NOM, CM, SENT}\}$$

$$\text{serie}_{\_} \rightarrow \text{CMSO} \setminus * (S, 10)\ \text{NOM CMSO serie}_{\_}\ /;\quad S = \{\text{CONJ, NOM, CM, SENT}\}$$

$$\text{serie} \rightarrow — \setminus * (S, 10)\ \text{NOM CMSO serie}_{\_}\ /;\quad S = \{\text{CONJ, NOM, CM, SENT}\}$$

b. **series entre paréntesis**:

$$\text{serie}_{()} \rightarrow \text{CMSO} \setminus * (S, 10)\ \text{NOM}\ /\ ) ;\quad S = \{\text{CONJ, NOM, CM, SENT}\}$$

$$\text{serie}_{()} \rightarrow \text{CMSO} \setminus * (S, 10)\ \text{NOM CMSO serie}_{()}\ /;\quad S = \{\text{CONJ, NOM, CM, SENT}\}$$

$$\text{serie} \rightarrow (\ \setminus * (S, 10)\ \text{NOM CMSO serie}_{()}\ /;$$

$$S = \{\text{CONJ, NOM, CM, SENT}\}$$

c. **series marcadas por dos puntos**:

$$\text{serie}_{:} \rightarrow \text{CMSO} \setminus * (S, 10)\ \text{NOM} \ /\ \text{SENT}\ ; \quad S = \{\text{CONJ, NOM, CM, SENT}\}$$

$$\text{serie}_{:} \rightarrow \text{CMSO} \setminus * (S, 10)\ \text{NOM CMSO serie}_{:}\ /; \quad S = \{\text{CONJ, NOM, CM, SENT}\}$$

$$\text{serie} \rightarrow : \setminus * (S, 10)\ \text{NOM CMSO serie}_{:}\ /; \quad S = \{\text{CONJ, NOM, CM, SENT}\}$$

En ambos casos, y al igual que las series cerradas con una conjunción, se utilizan tres reglas: una para marcar el final, otra para «acumular» elementos y otra para determinar el comienzo de la serie. En este caso, y a diferencia de los casos anteriores, determinar dónde comienza la serie es sencillo: la raya o el paréntesis marcan el límite buscado.

Un caso análogo a los anteriores es el de las series que comienzan por dos puntos y que, por lo observado en el corpus, se extienden hasta el final de la oración:

(4.12) Hay algo que Helena, ángel o demonio, siempre fue, en todas las versiones hasta ésta: una semidiosa, una diva, la primera estrella de Hollywood de la literatura.

Nuevamente, el problema de dónde comienza la serie se encuentra resuelto: existe un marcador claro, el símbolo dos puntos, que permite determinarlo. El marcador de fin, en este caso, se encuentra dado por el límite de la frase y no por otro símbolo correlativo, como en las series delimitadas por rayas y paréntesis.

Las reglas para el caso anterior se encuentran detalladas en el ejemplo 4.11(c). Al igual que para las series de paréntesis y rayas, las reglas utilizan únicamente el que haya una coma etiquetada como «CMSO» y un nombre para incorporarla a la serie.

Se propone, en un principio, un total de 22 reglas para el caso de las series simples. Sin embargo, aunque la mayoría de las series son detectadas, el marcado no se realiza correctamente.

Uno de los principales problemas es el error acumulado por el error cometido por el clasificador en la etapa previa. Así, muchos incisos son incluidos dentro de una serie, cuando ésta se encuentra formada por nombres propios y comunes. Por ejemplo, en el enunciado 4.12, se determina que «*Helena, ángel o demonio*» es una serie, debido a que el clasificador de comas erróneamente asigna el valor «CMSO» a la coma.

En consecuencia, se escriben reglas que especifican las series de acuerdo a sus componentes, únicamente «acumulando» elementos que sean de igual tipo: sólo nombres propios, sólo nombres comunes, etc. Si bien esta especificación de las

series no permite reconocer aquellas en las cuales sí se mezclan nombres propios y comunes, el comportamiento general del sistema mejora.

Las reglas esquematizadas en los ejemplos previos permiten reconocer series cortas, de pocas palabras, pero no series más largas o complejas como la del ejemplo 4.13:

(4.13) Star Wars era, sí, como volver al más torpe Big Bang del género: espadas zumbantes, princesas, compulsión tecnológica, nombres absurdos, extraterrestres muy raros con el cutis de los que sólo consumen comida chatarra, y slogans supuestamente místicos.

La forma en que se encuentra construido el intérprete a veces complica la formulación de las reglas para el análisis de series simples: dado que el reconocimiento se realiza de derecha a izquierda, la categoría que precede a una zona de exclusión debe además pertenecer a ésta. En consecuencia, si se desea reconocer un segmento de la forma «CMSO NOM. . . CMSO», se debe incluir la categoría «NOM» dentro de la zona de exclusión. Luego, si existe otra categoría «NOM» en el texto a etiquetar, el reconocimiento falla. Esto ocurre, por ejemplo, cuando el grupo nominal es relativamente complejo.

Una posible solución es crear varias reglas contemplando el caso en que la etiqueta ocurre una vez, dos veces, etc.; otra es realizar el análisis completo de los grupos nominales. Sin embargo, utilizando la información adicional provista por el clasificador de comas, alcanza con analizar las porciones de texto que aparecen inmediatamente luego de cada coma.

Para atacar este problema, se crean dos módulos que, en sucesivas pasadas, primero marcan la ocurrencia de nombres (con o sin determinantes) a continuación de una coma de serie, y luego, construyen con estos elementos a la serie en sí.

En el ejemplo 4.14 se muestra el proceso que realizan las reglas sobre el texto 4.13; en (a) se encuentran marcados los nombres que ocurren luego de una coma de serie simple, y en (b) el resultado final.

(4.14) a. Star Wars era, sí, como volver al más torpe Big Bang del género: [espadas] zumbantes,[CMSO] [princesas],[CMSO] [compulsión] tecnológica,[CMSO] [nombres] absurdos,[CMSO] [extraterrestres] muy raros con el cutis de los que sólo consumen comida chatarra,[CMSO] y [slogans] supuestamente místicos.

b. Star Wars era, sí, como volver al más torpe Big Bang del género:

espadas zumbantes,CMSO princesas,CMSO compulsión tecnológica,CMSO nombres absurdos,CMSO extraterrestres muy raros con el cutis de los que sólo consumen comida chatarra,CMSO y slogans supuestamente místicos. SERIE

Esta aproximación, aunque sencilla, permite obtener muy buenos resultados sin realizar mayores análisis, «confiando» básicamente en la información que proviene del clasificador.

En el siguiente ejemplo, se reconoce a la serie «de valor (...), de virtud (...), del honor(...), de rapiña(...)»:

(4.15) a. (...) la Guerra de Troya y los héroes que tomaron parte en ella han tenido un sentido distinto: modelo de valor guerrero y sabiduría para Homero, de virtud moral para Virgilio, del honor caballeresco y amor cortés en las versiones medievales y en Chaucer, de rapiña destructiva y ausencia de toda honra en la nihilista obra de Shakespeare ...

b. (...) la Guerra de Troya y los héroes que tomaron parte en ella han tenido un sentido distinto: modelo
de valor guerrero **y** sabiduría para Homero,
de virtud moral para Virgilio,
del honor caballeresco **y** amor cortés en las versiones medievales **y** en Chaucer,
de rapiña destructiva **y** ausencia de toda honra en la nihilista obra de Shakespeare ...

Gracias al marcado previo de nombres, la serie simple se detecta correctamente, a pesar de la presencia de conjunciones entre, por ejemplo, «*rapiña destructiva* » y «*ausencia de toda honra...* » o entre «*honor caballeresco*» y «*amor cortés*», que podrían haber disparado las otras reglas de reconocimiento de series simples. Estas últimas delimitarían mal los elementos de la serie, haciendo que, por ejemplo «rapiña» y «ausencia» figuraran como dos elementos distintos, cuando no lo son (a pesar de que sí son, a su vez, los únicos elementos de otra serie que no tiene ninguna coma).

Por otra parte, un problema similar ocurre cuando otras estructuras, principalmente incisos, interrumpen la serie. Este problema es «recursivo»: es posible que una serie sea parte de un inciso, que a su vez interrumpe una serie dentro de

un inciso... Sin embargo, por lo que se observa en el corpus, el nivel de anidamiento de estas estructuras —al menos entre los incisos y las series— no supera los dos niveles.

En estos casos, se marcan primero los incisos más simples, como se explica en la sección 4.2.3, y luego se aplican el módulo de series simples. Los incisos marcados «ocultan» las comas al módulo de series, lo que permite utilizar las mismas reglas para el marcado, que «ignoran» los incisos gracias a la subespecificación de las zonas de exclusión. Finalmente, se vuelve a correr el módulo de incisos, ahora, con las comas de series «ocultas» por el marcado previo.

4.2.2. Estructuras bipolares

El objetivo del módulo de «estructuras bipolares» es solucionar errores que el clasificador comete debido con mucha frecuencia: clasificar erróneamente las comas que se clasifican como bipolares.[10] Como se menciona en la sección 2.3.2, se presentan en estructuras en donde una de las componentes es «subordinada» a la otra, por ejemplo, frases condicionales, concesivas, etc.

(4.16) a. $\text{prop}_{bip} \rightarrow$ SENT\ CONJBIP *(S,50) VERBFIN / CM *(S,50) VERBFIN *(S,50) SENT;
$S = \{$CONJBIP, CM, VERBFIN, SENT$\}$

b. $\text{prop}_{bip} \rightarrow \text{prop}_{bip}$\ CM *(S,50) VERBFIN *(S,50) / SENT
$S = \{$CONJBIP, CM, VERBFIN, SENT$\}$

c. $\text{prop}_{bip} \rightarrow$ CMSP\ op(CONJ) CONJBIP *(S,50) VERBFIN *(S,50)/ CM
$S = \{$CONJBIP, CM, VERBFIN, SENT$\}$

Este módulo cuenta únicamente con tres reglas, descritas en 4.16. Las reglas (a) y (b) detectan la ocurrencia de este tipo de estructuras basándose en la ocurrencia de una conjunción *si*, *aunque*, etc. y una coma —sin importar el valor que le asigne el clasificador— separando dos segmentos de texto con un verbo conjugado. En caso de detectar esta estructura, se marca a cada segmento de texto que la compone con la etiqueta «proposición».

La regla (c) se basa también en la presencia de una conjunción y un verbo conjugado, pero exige la presencia de una coma marcada como de serie proposicional como contexto izquierdo. Esto le permite marcar una proposición que se encuentra en el medio de una oración, a diferencia de las otras dos reglas que solo marcan estructuras bipolares con dos componentes.

(4.17) a. Si$^{\text{CONJ}}$ se sacrifica$^{\text{VERBFIN}}$ un peón a cambio de mejor posición,$^{\text{CMMO}}$ se llama$^{\text{VERBFIN}}$ gámbito.

[10] Se recuerda que, debido a su baja frecuencia de aparición en el conjunto de entrenamiento, estas comas fueron eliminadas, con lo que el clasificador no puede aprenderlas.

b. Si se sacrifica un peón a cambio de mejor posición *prop*
,CMMO
se llama gámbito *prop*

Por ejemplo, en el texto 4.17, se detecta una estructura bipolar debido a la presencia de la conjunción *si* al comienzo de la frase y los verbos *sacrifica* y *llama* en dos segmentos de texto separados por una coma. A pesar que en 4.17(b) la etiqueta utilizada en el etiquetado es *prop*, en el marcado se distingue entre la proposición principal y la subordinada.

4.2.3. Incisos

Las comas etiquetadas como incisos son, en principio, sobre las que el etiquetador comete menos errores. Por ejemplo, se observa que un par de comas etiquetadas como comienzo y fin de un inciso, cuando no ocurre un verbo conjugado, marcan de forma sistemática la ocurrencia de un inciso. Esto permite la escritura de reglas sencillas, basadas principalmente en la valoración de las comas:

(4.18) a. inciso → \ CMII *(S,10) CMIF/;
$S = \{$inciso, PUNT, CM, VERBFIN, SENT$\}$

b. inciso → \ CMII *(S,10) / SENT;
$S = \{$inciso, PUNT, CM, VERBFIN, SENT$\}$

Las comas que marcan la presencia de incisos iniciales —tanto modificadores como conectivos— también son, en su mayoría, correctamente etiquetadas. Estas comas no tienen, en principio, correlación con otras: el comienzo de la oración marca el inicio del inciso, y la coma marca su final. En consecuencia, las reglas para estos casos únicamente utilizan las comas valuadas, la ausencia de un verbo conjugado y el comienzo de la oración para detectar el inciso y realizar el etiquetado.

(4.19) a. inciso_{ini} → SENT \ *(S,20) CMCO / ;
$S = \{$CM, SENT$\}$

b. inciso_{ini} → SENT \ *(S,20) CMMO / ;
$S = \{$CM, SENT$\}$

En el ejemplo 4.20, se puede apreciar el resultado de aplicar las reglas de 4.18 y 4.19 en el texto. Los dos incisos detectados «ocultan» las comas, lo que permite, durante la ejecución del módulo de series proposicionales, terminar el análisis de la oración.

(4.20) a. Por supuesto,CMMO aumentar los niveles de serotonina en el cerebro desencadena un proceso que,CMII con el tiempo,CMIF puede ayudar a personas deprimidas a sentirse mejor.

b. [Por supuesto,]$^{\texttt{CMMO}}$ *inciso*
aumentar los niveles de serotonina en el cerebro desencadena un proceso que
[, $^{\texttt{CMII}}$ con el tiempo,]$^{\texttt{CMIF}}$ *inciso*
puede ayudar a personas deprimidas a sentirse mejor.

Por otra parte, como se comenta en la sección 2.3, los modificadores y conectores usualmente se «acumulan» al comienzo de la oración. Este hecho se refleja en las reglas de incisos: no sólo se determina un inciso inicial con una coma y el comienzo de la oración (reglas 4.19), sino también con una coma y un inciso previamente etiquetado (reglas 4.21).

(4.21) a. $\text{inciso}_{ini} \rightarrow \text{inciso}_{ini}\backslash$ *(S,20) CMMO / ;
$S = \{\text{inciso}_{ini}, \text{CM}, \text{SENT}\}$

b. $\text{inciso}_{ini} \rightarrow \text{inciso}_{ini}\backslash$ *(S,20) CMCO / ;
$S = \{\text{inciso}_{ini}, \text{CM}, \text{SENT}\}$

Estas reglas permiten, entonces, reconocer varios incisos al comienzo de la oración, como se muestra en el ejemplo 4.22; en este caso, primero se reconoce al conectivo «*Por eso*», y luego, utilizando a éste como límite izquierdo, al formado por «*desde el arranque*».

(4.22) a. Por eso,$^{\texttt{CMCO}}$ desde el arranque,$^{\texttt{CMIF}}$ El Deseo es un pueblo con una ley propia y donde la verosimilitud sui generis fue instalada con mucha autoridad por los guionistas.

b. [Por eso,]$^{\texttt{CMCO}}$ *inciso* [desde el arranque,]$^{\texttt{CMIF}}$ *inciso*
El Deseo es un pueblo con una ley propia y donde la verosimilitud sui generis fue instalada con mucha autoridad por los guionistas.

El otro tipo de incisos delimitados por una única coma son los que ocurren al final de las oraciones. En estos casos, determinar si efectivamente son incisos no es tan sencillo como en el caso anterior, sobre todo debido a posibles errores del analizador morfosintáctico y del evaluador de comas. El mismo problema se repite en los incisos se encuentran delimitados por dos comas en cualquier lugar de la oración.

Para contemplar todos estos casos, se decide utilizar otras marcas presentes en el texto, pronombres relativos y participios de verbos, como se muestra en las reglas del ejemplo 4.23.

(4.23) a. inciso → \ CMII op(DET) op(NOM) PRON *(S,20)
CMIF /;

$$S = \{\text{inciso, CM, SENT, PRON, PUNT}\}$$

b. inciso → \ CMII op(DET) op(NOM) PRON *(S,20) / SENT;
$S = \{\text{inciso, CM, SENT, PRON, PUNT}\}$

c. inciso → \ CMII VPART *(S,50) CMIF / ;
$S = \{\text{inciso, VERB, CM, SENT, PUNT}\}$

d. inciso → VERBFIN *(S_1,50) \ CMII VPART *(S_2,50) / SENT;
$S_1 = \{\text{VERB, SENT}\}$
$S_2 = \{\text{inciso, VERB, CM, SENT, PUNT}\}$

En el ejemplo 4.24 se observan dos incisos; el primero, «además... homérico», es marcado por las regla 4.18(a) [11], mientras que el segundo, al final de la oración, queda determinado por la coma de inciso inicial y el pronombre relativo *quien*.

(4.24) a. Este duelo final entre dos héroes «buenos»,CMII además de genuinamente homérico,CMIF tiene una fuente inesperada en el proyecto anterior de Petersen ,CMII quien PRONREL se abocó a Troya tras abandonar la épica Superman contra Batman.

b. Este duelo final entre dos héroes «buenos» ,CMII además de genuinamente homérico,CMIF *inciso* tiene una fuente inesperada en el proyecto anterior de Petersen ,CMII quien se abocó a Troya tras abandonar la épica Superman contra Batman *inciso*.

Otro ejemplo sobre marcado de incisos, esta vez basándose en la presencia de un participio, se muestra en el ejemplo 4.25: «*ofreciéndose*» sirve de marca confirmatoria del inciso final.

(4.25) a. Quizás Troya hubiera alcanzado el sabor de lo épico si dejaban que el incendio arrasara con todo,CMSP y en una gran apoteosis final el director se inmolara en su propia película,CMII ofreciéndose PART a ser pasto de las llamas.

b. Quizás Troya hubiera alcanzado el sabor de lo épico si dejaban que el incendio arrasara con todo,CMSP y en una gran apoteosis final el director se inmolara en su propia película ,CMII ofreciéndose a ser pasto de las llamas *inciso*.

[11] El texto se encuentra entre dos comas y no hay ocurrencia de verbos.

Para marcar los incisos de discurso se opta por exigir, además de al menos una coma etiquetada como «CMID», la existencia de un verbo discursivo dentro del segmento. Se aprovecha, entonces, la tabla de verbos de este tipo realizada para el clasificador de comas.

Así, las reglas del ejemplo 4.26 marcan un segmento como inciso discursivo cuando contiene un verbo de discurso entre una coma etiquetada de inciso discursivo y el final de la oración (regla a) o entre una coma de inciso discursivo y otra coma cualquiera (regla b).

(4.26) a. inciso → \ CMID $*(S, 20)$ VERBFINDISC $*(S, 20)$ / SENT;
$S = \{$VERBFIN, CM, SENT$\}$

b. inciso → \ CMID $*(S, 20)$ VERBFINDISC $*(S, 20)$ / CM ;
$S = \{$VERBFIN, CM, SENT$\}$

Finalmente, se implementan reglas para los incisos que se encuentran delimitados por paréntesis y rayas. Al texto presente en este de tipo incisos se lo analiza en búsqueda de series simples u otros incisos, lo que lleva a la creación de reglas similares a las expuestas en los ejemplos anteriores, en donde el límite considerado ya no es el comienzo o el final de la oración, sino la raya o el paréntesis según corresponda.

En el ejemplo 4.27, se marcan dos incisos: el primero utilizando las reglas expuestas en 4.21[12], otro inciso delimitado por los paréntesis al final de la oración, y dentro de éste, el inciso «de hecho», marcado por una coma de inciso conector y el paréntesis de apertura.

(4.27) a. Una vez hechas las cuentas,$^{\texttt{CMMO}}$ resultó que el número de granos era tan elevado que con los graneros de todo el imperio no había suficiente trigo para pagar la recompensa (de hecho,$^{\texttt{CMCO}}$ se necesitaría la producción mundial actual de trigo de una docena de años).

b. [Una vez hechas las cuentas,$^{\texttt{CMMO}}$ $_{inciso}$
resultó que el número de granos era tan elevado que con los graneros de todo el imperio no había suficiente trigo para pagar la recompensa
([de hecho,$^{\texttt{CMCO}}$ $_{inciso}$ se necesitaría la producción mundial actual de trigo de una docena de años)]$_{inciso}$.

A pesar que, como se muestra en el ejemplo anterior, se realiza un análisis en las porciones de texto entre paréntesis o guiones, éste es mínimo: sólo se intenta identificar series simples e incisos. Luego no se le realiza ningún otro análisis, debido a que al ser marcado como inciso se impide su posterior tratamiento.

[12] Hay una coma marcada como inciso modificador al comienzo de la oración.

Este proceso tiene como principal ventaja la simplificación de las reglas para las sucesivas etapas[13].

4.2.4. Series proposicionales

Las reglas de series proposicionales se basan fuertemente en la presencia de tres marcadores en el texto: una coma evaluada como de serie proposicional («CMSP»), la presencia de una conjunción (*y*, *o*, *ni*, *pero*, etc.) luego de la coma y la ocurrencia de un verbo conjugado.

(4.28) a. Ensayamos$^{\texttt{VERBFIN}}$ el incendio durante varias semanas antes de filmar,$^{\texttt{CMSP}}$ porque$^{\texttt{CONJ}}$ algo siempre puede$^{\texttt{VERBFIN}}$ salir mal,$^{\texttt{CMSP}}$ y$^{\texttt{CONJ}}$ por eso no hubo$^{\texttt{VERBFIN}}$ heridos durante la filmación de la escena.

b. [Ensayamos el incendio durante varias semanas antes de filmar]$_{prop}$,$^{\texttt{CMSP}}$ [porque algo siempre puede salir mal]$_{prop}$,$^{\texttt{CMSP}}$ [y por eso no hubo heridos durante la filmación de la escena.]$_{prop}$

Las reglas son, en cierta forma, análogas a las de series simples: primero se marca el final de la serie, en este caso dado por el límite de una frase, y luego se construye la serie de derecha a izquierda, exigiendo la ocurrencia de las marcas mencionadas.

(4.29) a. prop → CMSP \ CONJ *(S_1,50) VERBFIN*(S_2,50) / SENT;
S_1 = {prop, conj, CMSP, SENT}
S_2 = {prop, CMSP, VERBFIN, SENT}

b. prop → CMSP \ CONJ *(S_1,50) VERBFIN *(S_2,50)/ CMSP;
S_1 = {prop, conj, CMSP, SENT}
S_2 = {prop, CMSP, VERBFIN, SENT}

c. prop → CMSP \ CONJ *(S_1,50) VERBFIN *(S_2,50)/ SENT;
S_1 = {prop, conj, CMSP, SENT}
S_2 = {prop, CMSP, VERBFIN, SENT}

d. prop → SENT \ *(S_1,20) VERBFIN *(S_2,20)/ SENT;
S_1 = {prop, inciso$_{ini}$, CM, SENT}
S_2 = {prop, CM, VERBFIN, SENT}

[13]Por ejemplo, en el proyecto Clatex no se realiza ningún análisis del texto entre guiones o paréntesis, quedando su análisis para una futura etapa.

Sin embargo, a diferencia de las series simples, este tipo de serie no se marca como un bloque con etiqueta *serie*, sino que sus elementos quedan «separados», aunque rotulados como proposiciones. Las comas «CMSP» indican los límites de los elementos, y no son incorporadas a ninguno de ellos, como sucedía en las series simples.

Como es de esperar, las reglas anteriores no permiten el reconocimiento de todos los elementos que conforman una serie de proposiciones. Por ejemplo, no siempre existe un verbo conjugado dentro del segmento de texto a marcar, habiendo oraciones en el corpus en donde los verbos se encuentran elididos. La detección en estos casos no es sencilla, y se debe tomar en cuenta la presencia de otros fenómenos: la ocurrencia de una conjunción, otros elementos de la serie de proposiciones previamente marcados rodeando a la porción de texto sin marcar, etc.

(4.30) a. No son$^{\texttt{VERBFIN}}$ fieritas como los vecinos de los Roldán,$^{\texttt{CMSP}}$ no son$^{\texttt{VERBFIN}}$ comunes y corrientes como los chicos de Los pensionados,$^{\texttt{CMSP}}$ ni$^{\texttt{CONJ}}$ siquiera buena gente como la familia de Pablito Echarri en los comienzos de Resistiré.

b. [No son fieritas como los vecinos de los Roldán]$_{prop}$ CMSP , [no son comunes y corrientes como los chicos de Los pensionados]$_{prop}$ CMSP , [ni siquiera buena gente como la familia de Pablito Echarri en los comienzos de Resistiré.]$_{prop}$

En el ejemplo 4.30, la última proposición («*ni siquiera...*») tiene el verbo *ser* elidido. La presencia antepuesta de texto etiquetado como proposición, seguida de una coma «CMSP» y una conjunción permite etiquetarla como proposición.

Dentro de las series proposicionales, los dos puntos también juegan un rol importante, dado que muchas veces actúan como un separador de proposiciones. Además, y a diferencia de las series simples, se toma en cuenta al resto de los signos de puntuación —punto y comas, puntos suspensivos y signos de exclamación e interrogación— como separadores de proposiciones. [14]

(4.31) a. En la declaración que efectuó en 1996 ante la policía alemana,$^{\texttt{CMMO}}$ Mesbahi aseguró$^{\texttt{VERBFIN}}$ que había llegado a un acuerdo con la policía suiza: ésta lo dejaba$^{\texttt{VERBFIN}}$ «tranquilo para llevar a cabo mis actividades políticas en el país».

[14] Este criterio es similar al adoptado en el segmentador de proposiciones del proyecto Clatex.

b. [En la declaración que efectuó en 1996 ante la policía alemana,]$_{inciso}$
[Mesbahi aseguró que había llegado a un acuerdo con la policía suiza]$_{prop}$
:
[ésta lo dejaba «tranquilo para llevar a cabo mis actividades políticas en el país».]$_{prop}$

En el ejemplo 4.31, el módulo de incisos marca el comienzo de la frase como un inciso inicial. Luego, el dos puntos surge como un separador de proposiciones, las cuales son «confirmadas» por la presencia de los verbos finitos «*aseguró*» y «*dejaba*».

Se destaca que todas los reglas de reconocimiento de proposiciones en este módulo se basan en la presencia de comas etiquetadas como de serie proposicional.

4.2.5. Análisis final

El último módulo consta de reglas que intentan terminar el análisis del texto en caso que no se haya logrado en alguno de los módulos previos debido a la presencia de comas mal clasificadas, comas incorrectamente utilizadas en el texto o la falta de marcas «confirmatorias» (en las reglas que así lo requerían).

Las reglas sólo funcionan con frases en las cuales quedan sin utilizar una única coma e intentan detectar incisos iniciales, finales y proposiciones. Ninguna de éstas toma en cuenta el valor de las comas asignado por el clasificador y, por ende, requieren la presencia o ausencia de otros elementos en el texto.

(4.32) a. Sin la exacta explicación de Borges («el pudor estoico no había sido aún inventado y Héctor podía huir sin desmedro»),$^{\texttt{CMII}}$ quizá no hubiera resultado aceptable.

b. Sin la exacta explicación de Borges
[(«el pudor estoico no había sido aún inventado y Héctor podía huir sin desmedro»)]$_{inciso}$
,$^{\texttt{CMII}}$ quizá no [hubiera]$^{\texttt{VERBFIN}}$ resultado aceptable.

c. [Sin la exacta explicación de Borges
[(«el pudor estoico no había sido aún inventado y Héctor podía huir sin desmedro»)]$_{inciso}$
,$^{\texttt{CMII}}$]$_{inciso}$
[quizá no hubiera resultado aceptable.]$_{prop}$

Por ejemplo, la única coma de la frase 4.32(a) se encuentra erróneamente clasificada —debiera ser CMIF en vez de CMII—, y los módulos anteriores sólo marcan el inciso entre paréntesis, como se muestra en 4.32(b). En este último módulo, la ausencia de un verbo en el segmento de texto anterior a la coma, y su presencia en el posterior, permiten etiquetarlos como inciso inicial y proposición respectivamente.

Las reglas en 4.33 son algunas de las que permiten el marcado de estructuras inciso/proposición: las dos primeras detectan el inciso al comienzo de la oración y las dos últimas la proposición que le sigue. En particular, las reglas (a) y (c) son las que realizan el marcado del ejemplo 4.32.

(4.33) a. $\text{inciso}_{est} \rightarrow$ SENT\ *(S$_1$,50) CM / *(S$_1$,50) VERBFIN *(S$_2$,50) SENT;
$S_1 = \{$prop, CM, VERBFIN, SENT$\}$
$S_2 = \{$CM, SENT$\}$

b. $\text{inciso}_{est} \rightarrow$ SENT OP(inciso) OP(inciso) \ PREP *(S$_1$,50) CM / *(S$_2$,50) VERBFIN *(S$_3$,50) SENT;
$S_1 = \{$PREP, CM, prop, SENT$\}$
$S_2 = \{$VERBFIN, CM, prop, SENT$\}$
$S_3 = \{$CM, SENT$\}$

c. $\text{prop}_{est} \rightarrow \text{inciso}_{est}$\ *(S$_1$,50) VERBFIN *(S$_2$,50) / SENT;
$S_1 = \{\text{inciso}_{est}$, CM, VERBFIN, SENT$\}$
$S_2 = \{$CM, SENT$\}$

d. $\text{prop}_{est} \rightarrow \text{inciso}_{est}$\ *(S$_1$,50) VERBFIN *(S$_2$,50) / prop;
$S_1 = \{\text{inciso}_{est}$, CM, prop, SENT$\}$
$S_2 = \{$CM, prop, VERBFIN, SENT$\}$

Por otra parte, se crean reglas análogas a las anteriores para detectar estructura de la forma proposición/inciso final, como las descriptas en el ejemplo 4.34.

(4.34) a. $\text{inciso}_{est} \rightarrow$ SENT *(S$_1$,50) VERBFIN *(S$_2$,50) \ CM *(S$_2$,50) / SENT;
$S_1 = \{$CM, SENT, $\text{inciso}_{est}\}$
$S_2 = \{$prop, inciso_{est}, CM, VERBFIN, SENT$\}$

b. $\text{prop}_{est} \rightarrow$ SENT\ *(S$_1$,50) VERBFIN *(S$_2$,50) /inciso_{est};
$S_1 = \{$CM, SENT, prop$\}$
$S_2 = \{$prop, CM, VERBFIN, SENT$\}$

Finalmente, se intenta detectar estructuras proposición/proposición: el marcado se realiza cuando en ambos segmentos se detectan verbos finitos.

(4.35) a. $\text{prop}_{est} \rightarrow$ SENT \ *(S$_1$,50) VERBFIN *(S$_2$,50) CM/ *(S$_1$,50) VERBFIN *(S$_2$,50) SENT;

$$S_1 = \{\text{CM, SENT, prop }\}$$
$$S_2 = \{\text{prop, CM, VERBFIN, SENT}\}$$

b. $\text{prop}_{est} \rightarrow \text{prop}_{est}$ CM \ *(S_1,50) VERBFIN *(S_2,50) / SENT;

$$S_1 = \{\text{CM, SENT, prop }\}$$
$$S_2 = \{\text{prop, CM, VERBFIN, SENT}\}$$

En el siguiente ejemplo, se aplican las reglas 4.35 para reconocer dos proposiciones, a pesar que la coma se encuentra erróneamente clasificada como de inciso inicial:

(4.36) a. Shakespeare se embarca$^{\text{VERBFIN}}$ en una inversión radical del espíritu épico-heroico que la Ilíada representaba,$^{\text{CMII}}$ y su perversión de ella está$^{\text{VERBFIN}}$ animada por un asco sincero y sistemático.

b. [Shakespeare se embarca en una inversión radical del espíritu épico-heroico que la Ilíada representaba]$_{prop}$ $^{\text{CMII}}$, [y su perversión de ella está animada por un asco sincero y sistemático.]$_{prop}$

Se destaca que todas las reglas de este módulo utilizan todas las categorías de la oración (como zona a etiquetar o como contexto) para lograr el marcado, a diferencia de lo que sucede en los módulos anteriores, en donde, por lo general, se utiliza un contexto mucho más reducido.

4.3. Resultados

Para realizar la evaluación de los módulos implementados, se seleccionaron textos que no habían sido utilizados ni en la fase de entrenamiento del clasificador de comas, ni en el análisis para la escritura de las reglas.

El texto de evaluación se compone de dos artículos, con un total de 153 oraciones. De este total, se descartan 52 oraciones por no presentar signos de puntuación, quedando, entonces, 102 oraciones a procesar con un largo promedio de 25 palabras.

En primer lugar, se le aplica al texto el proceso de análisis —etiquetado morfosintáctico, evaluación de comas y análisis sintáctico—, para luego, a partir del grafo de texto resultante generar código HTML con las estructuras encontradas, al cual se le aplica una hoja de estilo para mejorar su visualización. El texto es resaltado en distintos colores de acuerdo a la etiqueta asignada: tonalidades de azul a los incisos, rojo a las series simples, verde a las series proposicionales, etc. En la figura 4.2 se presenta parte de la salida obtenida para uno de los textos utilizados.

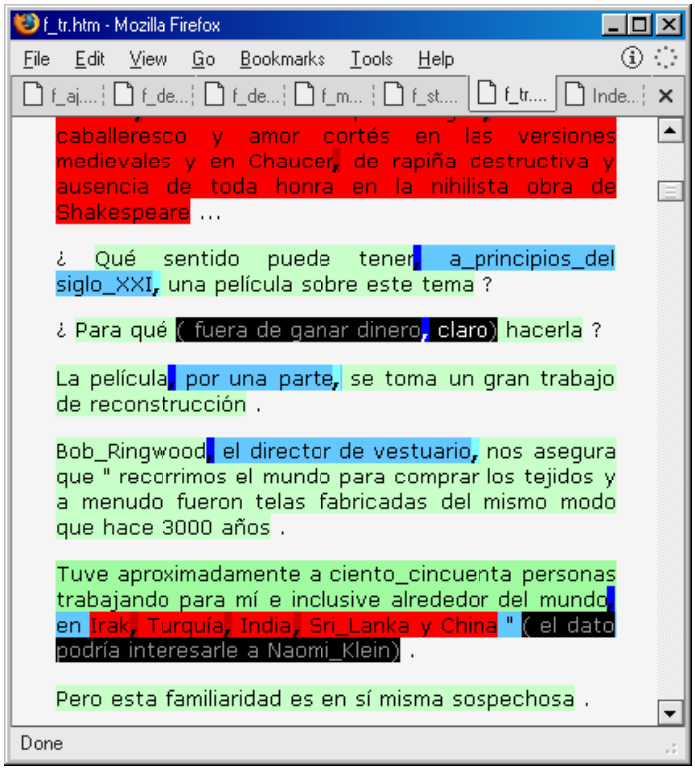


Figura 4.2: Texto coloreado dado como salida por el sistema.

Para realizar una evaluación numérica de los resultados, se utilizan las medidas de precisión, recuperación y medida-f, definidas en la sección 3.3. Cabe señalar que en los resultados se reflejan tres clases de errores, que se acumulan durante el proceso: (a) el error del analizador morfosintáctico, (b) el del clasificador de comas y (c) el de las reglas propuestas.

A los efectos de realizar los cálculos, se considera incorrecto:

- Límites mal establecidos: la zona segmentada no abarca todo el texto que debiera. Esto ocurre con las series simples y las proposiciones.
- Etiquetado incorrecto: el segmento se encuentra delimitado correctamente, pero la etiqueta asignada es incorrecta.

Además, se toma como válido el etiquetado, o bien con «inciso», o bien con «proposición», de segmentos de la forma «cuando...», «como...», etc., del estilo del ejemplo 4.37, debido a que ambas interpretaciones —inciso modificador o serie proposicional— resultan correctas. Por otra parte, dentro de la categoría incisos no se toman en cuenta aquellos delimitados por paréntesis o rayas.

(4.37) a. [Cuando Byrnes le informó de los planes del gobierno,]^{inciso} Szilard objetó que, derrotada Alemania, no tenía objeto seguir con la bomba.

b. Por último, [cuando va a ser usado para reprimir una sublevación en las colonias africanas]^{prop} ,los nativos destruyen accidentalmente la bomba antes de que pueda ser usada.

Del total de 102 oraciones procesadas, 68 fueron estructuradas correctamente de forma completa: todo el texto es etiquetado y a las estructuras se les asigna la etiqueta correcta, y no hay errores en sus límites. Se reconocen en estas oraciones un total de 90 segmentos de proposiciones, 56 de incisos y 5 series simples.

Las 34 oraciones restantes presentan distintos problemas. Algunas cuentan con estructuras reconocidas parcialmente, cuyos límites son incorrectos; otras presentan segmentos correctamente delimitados pero con etiquetadas mal asignadas (proposición por inciso, inciso por serie, etc.). Sin embargo, dentro de estas oraciones, hay porciones de texto correctamente analizadas: se detectan 15 segmentos de proposición, 13 incisos y 1 serie simple.

El número de elementos reconocidos junto con las medidas de precisión, recuperación y medida-f (discriminadas por etiqueta) se presentan en el cuadro 4.1. El detalle de los errores cometidos se encuentra en el cuadro 4.2.

Los resultados son buenos para el etiquetado de proposiciones e incisos —84 % y 85 % de medida-f, respectivamente— y no muy buenos para las series simples —únicamente un 50 %— .

	Correctas	Incorr.	No rec.	Prec.	Recup.	Medida-f
Series simples	6	2	8	75	38	**50**
Incisos	69	4	21	95	77	**85**
Proposiciones	105	14	25	88	81	**84**

Cuadro 4.1: Precisión, recuperación y medida-f para el analizador

	Límite	Serie simple	Inciso	Proposición	Total
Series simple	2		0	0	**2**
Inciso	0	3		1	**4**
Proposición	7	2	5		**14**

Cuadro 4.2: Segmentos mal etiquetados: en las filas, se encuentra la etiqueta asignada; en las columnas, si el error es en la delimitación del segmento o en la etiqueta (discriminado por etiqueta correcta).

En el caso de las series simples, los errores se deben exclusivamente a series reconocidas parcialmente, y no a la existencia de texto marcado como serie sin ser tal. Este resultado se ve influenciado por dos factores importantes: primero, hay errores cometidos por el analizador morfosintáctico, el cual marca de forma sistemática como adverbio las palabras que no tiene en el diccionario; segundo, no todas las comas de estas series son catalogadas con CMSO por el clasificador.

En el ejemplo 4.38, el reconocimiento de la serie termina en «*autorreproches*», al ser esta palabra clasificada erróneamente como adverbio («ADV»). Inclusive, aunque estuviese etiquetada como nombre común, la serie no sería reconocida por la presencia de una coma de inciso.[15]

(4.38) a. Estados de ánimo y afectividad: tristeza,$^{\texttt{CMII}}$ baja autoestima,$^{\texttt{CMSO}}$ autorreproches$^{\texttt{ADV}}$,$^{\texttt{CMSO}}$ pérdida de placer e interés,$^{\texttt{CMSO}}$ sensación de vacío,$^{\texttt{CMSO}}$ apatía,$^{\texttt{CMSO}}$ ansiedad,$^{\texttt{CMSO}}$ tensión,$^{\texttt{CMSO}}$ irritabilidad,$^{\texttt{CMSO}}$ inhibiciones varias.

b. Estados de ánimo y afectividad: tristeza,$^{\texttt{CMII}}$ baja autoestima,$^{\texttt{CMSO}}$ autorreproches$^{\texttt{ADV}}$,$^{\texttt{CMSO}}$ pérdida de placer e interés, sensación de vacío, apatía, ansiedad, tensión, irritabilidad, inhibiciones varias.^{serie}

La falta de reconocimiento también se ve influida por los errores cometidos durante el análisis morfosintáctico, a lo que se agregan las series separadas por puntos y comas, no consideradas en este trabajo.

[15] Esto es discutible, dado que el evaluador podría cambiar su clasificación al cambiar el análisis morfosintáctico de la oración.

Los incisos tienen un alto grado de precisión (95 %), esto se debe principalmente a que los incisos apositivos y los iniciales (modificadores y conectivos) son reconocidos prácticamente sin errores. Tres de los cuatro incisos etiquetados incorrectamente son en realidad series simples, con una coma etiquetada como delimitador de inciso (como en el ejemplo 4.39). El cuarto segmento erróneamente etiquetado como inciso es en realidad una serie de tres nombres propios, no reconocida por estar mal valuada la única coma presente en la serie.

(4.39) a. El agobio se expresa en la temporalidad («no tengo futuro»),$^{\texttt{CMII}}$ en la motivación («no tengo fuerzas») y en el valor («no valgo nada»).

b. El agobio se expresa en la temporalidad
(«no tengo futuro»)^{inciso}
,$^{\texttt{CMII}}$ en la motivación («no tengo fuerzas»)^{inciso} y
en el valor («no valgo nada»)^{inciso}.^{inciso}

El nivel de recuperación es en los incisos significativamente más bajo que en el resto de las clases. Esto se explica por la presencia de comas erróneamente valuadas o estructuras complejas que no son capturadas por las reglas, así como la falta de reglas para cubrir comas con múltiples funciones o incisos delimitados por otros signos de puntuación (dos puntos, punto y coma, etc.).

Por ejemplo, en 4.40, la primera coma se encuentra mal clasificada como límite de inciso inicial. Ninguna regla puede aplicarse en este caso: por un lado, no existe luego una coma de fin de inciso, lo que permitiría etiquetar un inciso apositivo, ni tampoco una proposición al comienzo de la oración, lo cual permitiría etiquetar un inciso «modificador».

(4.40) a. El 28 de mayo,$^{\texttt{CMII}}$ Leo Szilard,$^{\texttt{CMII}}$ que trabajaba para el Proyecto Manhattan,$^{\texttt{CMIF}}$ se entrevistó con el secretario de Estado James F. Byrnes,$^{\texttt{CMII}}$ quien volvía de la conferencia de Yalta.

b. El 28 de mayo,$^{\texttt{CMII}}$ Leo Szilard
,$^{\texttt{CMII}}$ que trabajaba para el Proyecto Manhattan,$^{\texttt{CMIF}}$ ^{inciso}
se entrevistó con el secretario de Estado James F. Byrnes
,$^{\texttt{CMII}}$ quien volvía de la conferencia de Yalta. ^{inciso}

En la frase 4.41, el inciso no se detecta por no contemplarse correctamente el caso en que una coma cumple más de una función en la oración: la segunda coma actúa como límite de inciso y como separador de proposiciones al mismo tiempo.

(4.41) a. Allí se narraba cómo en respuesta a un ataque japonés el presidente de los Estados Unidos autoriza desarrollar «la mayor arma que jamás ha conocido la ciencia»,$^{\texttt{CMII}}$ que combina el poder del átomo con el poder aéreo,$^{\texttt{CMSP}}$ y decide usarla «para terminar para siempre con las guerras».

b. Allí se narraba cómo en respuesta a un ataque japonés el presidente de los Estados Unidos autoriza desarrollar «la mayor arma que jamás ha conocido la ciencia»,$^{\texttt{CMII}}$ que combina el poder del átomo con el poder aéreo

$\boxed{,^{\texttt{CMSP}} \text{ y decide usarla «para terminar para siempre con las guerras».}}_{prop}$

La precisión de las proposiciones respecto a los incisos baja significativamente (87 % *versus* 95 %), debido en iguales proporciones a los errores al establecer sus límites, como a las etiquetas mal asignadas. El primero de los errores se origina en que no siempre se cuenta con límites claros en las proposiciones. Por ejemplo, no siempre se tiene una marca de puntuación al comienzo o al final de la proposición, y por otro lado, una misma marca puede servir tanto de límite a una proposición, como formar parte de ella. Esto no sucede con los incisos ya que sus límites siempre son claros: dos comas, una coma y el fin de la oración, etc.

Por otra parte, el reconocimiento de incisos y series influye de forma significativa, tanto en la precisión, como en la recuperación de proposiciones. En la precisión, porque el último módulo etiqueta proposiciones en base a la presencia de verbos conjugados y una coma en el texto, la cual puede ser el límite de un inciso no reconocido en un paso previo. En la recuperación, porque para el etiquetado de una oración como proposición, no pueden quedar segmentos con comas sin «ocultar» por marcas de series e incisos. Se estima, entonces, que una mejora en la recuperación de series simples e incisos trae aparejada también una mejora en los niveles de recuperación de las proposiciones.

Nuevamente, los resultados parecen confirmar, en cierta medida, las observaciones realizadas para el idioma inglés por Bayraktar, Say y Akman sobre la estabilidad de las clases de comas [16] respecto a los patrones sintácticos en los cuales aparecen: los incisos apositivos e iniciales son fácilmente capturables mediante pocos patrones sencillos, mientras que las series y los incisos «irrestrictos» precisan más reglas de una mayor complejidad.

De alguna forma, esto también se puede haber reflejado en el aprendizaje de la evaluación de comas: en la medida que una clase presenta mayor diversidad en su uso, más difícil resulta aprender cómo detectar su ocurrencia.

[16] La definición de *estabilidad* se encuentra en la sección 2.2.2

Resumen

A pesar que el clasificador de comas tiene una tasa de error relativamente alta, se plantea la hipótesis que su evaluación de las comas puede ser útil para el análisis de superficie de texto.

Se experimenta, entonces, con la construcción de un analizador sintáctico de superficie utilizando el formalismo de reglas contextuales. Este analizador se basa principalmente en la puntuación presente en el texto para segmentarlo y etiquetarlo.

Los resultados obtenidos confirman que, con reglas poco complejas, se pueden capturar razonablemente distintos fenómenos textuales.

Capítulo 5

Conclusiones

El presente trabajo se plantea como objetivo el estudio de la relevancia de los signos de puntuación en la tarea del análisis sintáctico, centrándose en la coma como caso particular de estudio. La tarea se lleva a cabo en tres etapas bien diferenciadas:

1. Elaboración de una clasificación de los usos de la coma.
2. Construcción de un evaluador de comas utilizando métodos de aprendizaje automático.
3. Construcción un analizador sintáctico de superficie.

Cada una de éstas tiene un aporte concreto, los cuales se presentan a continuación.

Clasificación

La primera tarea consiste en la creación de una categorización de las comas según su uso, que se adapte a las tareas de procesamiento automático de texto.

Para esto, se estudia la normativa de la Real Academia Española para la puntuación en el idioma español y clasificaciones para el inglés que, si bien no son «oficiales», se crean para el análisis automático. El estudio de estas clasificaciones inspira, entonces, la creación de una nueva categorización.

La nueva clasificación es corroborada mediante su aplicación a un conjunto de artículos periodísticos. La experimentación directa con casos reales permite refinar y ajustar definitivamente el conjunto de categorías, obteniendo así una clasificación que cumple con los objetivos planteados.

Clasificador de comas

Contando con una clasificación definida, se procede a la construcción del evaluador de comas utilizando técnicas de aprendizaje automático.

En primer lugar, debido a que ambas técnicas son supervisadas, se debe construir un conjunto de entrenamiento con ejemplos en los cuales las comas se encuentren etiquetadas. Esto implica el análisis de textos y la clasificación manual de aproximadamente 5200 comas presentes en ellos.

Luego, se realizan experimentos con dos técnicas distintas, árboles de decisión y *boosting*, variando el conjunto de atributos utilizados para describir las instancias del problema. Por un lado, los árboles permiten obtener un clasificador interpretable, transformándolo a un conjunto de reglas; por otro, *Boostexter* da mejores resultados a costa de la legibilidad del resultado.

Los clasificadores resultantes de ambos algoritmos no presentan un gran nivel de precisión. Esto puede deberse a que el conjunto de entrenamiento no es lo suficientemente grande, a que el conjunto de características para representar el problema no es el adecuado, a que los algoritmos no son los adecuados para capturar una solución o a que el problema es difícil de aprender por su naturaleza.

Un punto a favor de la última hipótesis son los resultados dados por Bayraktar et al. [4]: la precisión que se obtiene por categoría es inversamente proporcional a la *estabilidad*, o sea, a la diversidad de patrones sintácticos que presenta cada uno de los distintos usos. Además, cabe destacar que determinar exactamente qué función cumple una coma en el texto es muchas veces complejo inclusive para un evaluador humano. Sin embargo, es imprescindible realizar nuevos experimentos variando los otros tres parámetros para confirmar o desmentir esta hipótesis.

Lamentablemente, no existen trabajos similares sobre la coma en el área del aprendizaje automático para el español con los cuales comparar resultados. El único trabajo reportado sobre clasificación de comas (para el inglés) es el de van Delden y Gómez [62], el cual se basa en un conjunto de reglas determinadas de forma manual —no inducidas a partir del corpus— y codificadas con autómatas finitos.

Si bien los resultados presentados por van Delden y Gómez, más del 90 % de precisión, superan ampliamente los obtenidos con las técnicas aplicadas en este trabajo, se debe tener en cuenta que: (a) las reglas fueron escritas de forma manual, varias de ellas para capturar fenómenos muy específicos (según los propios autores); y (b) se realiza un preprocesamiento de la entrada reconociendo, por ejemplo, cláusulas relativas, algo descartado en este trabajo porque el objetivo final es realizar el análisis sintáctico a partir del etiquetado.

Por otro lado, los autores entrenan al etiquetador de Brill [7] para que tenga en cuenta distintas etiquetas para las comas, logrando una precisión del 55 %, diez puntos por debajo de la del clasificador basado en *boosting*.

Ambos resultados parecen confirmar los obtenidos por García y González [23] para el problema de reconocimiento y clasificación de entidades con nombre: las soluciones con técnicas de aprendizaje automático logran una menor tasa de precisión que aquellas construidas manualmente mediante reglas.

Analizador sintáctico

Finalmente, se desarrolla un conjunto de reglas contextuales para el análisis de superficie de texto. Estas reglas se basan en la valuación hecha por el clasificador obtenido con *Boostexter* y en otros fenómenos lingüísticos que se detectan en los textos del corpus, pero utilizando a estos últimos lo menos posible, con el objetivo de estimar cuán aprovechable es contar con una estimación de la función que cumple cada una de las comas en el texto.

Si bien la información sobre las comas es imperfecta, se logran niveles de precisión y recuperación entre buenos y muy buenos en las proposiciones e incisos, y buenos niveles de precisión en el marcado de series simples. Los resultados obtenidos por tipo de estructuras parecen confirman los resultados primarios: en el español también hay estructuras más «estables» que otras. Estas estructuras son fácilmente capturables con reglas sencillas, aprovechando el valor de las comas presentes en el texto.

En definitiva, el tratamiento de la puntuación ayuda en el análisis sintáctico de textos en español y, en particular, contar con información adicional de qué función cumplen las comas en un texto, permite escribir analizadores concisos y con buenos resultados. Se concluye que las aproximaciones estadísticas (el clasificador) y los enfoques simbólicos (las reglas contextuales) se complementan, permitiendo obtener buenos resultados en el tratamiento automático de texto.

5.1. Trabajos a futuro

Cada una de las tres etapas en que se divide el trabajo presenta puntos a ser mejorados o a ser tratados en mayor profundidad.

La clasificación sobre usos de coma debería ser validada contra textos de un origen diferente al periodístico, lo cual podría acarrear nuevas modificaciones y ajustes en la cantidad de categorías y subcategorías (o en los alcances de éstas).

En cuanto a la implementación de los clasificadores, es indispensable realizar nuevos experimentos para mejorar o validar los resultados obtenidos. Dentro de las tareas a realizar se encuentran:

- Cambiar los atributos que describen las instancias del conjunto de entrenamiento. Esto implica reformular las existentes (sus valores, la forma de calcularlos, etc.) y estudiar el uso de nuevos atributos en conjunto o sustituyendo algunos de los actuales.

- Aumentar el conjunto de entrenamiento. Esto permitiría medir cuánto y de qué forma influye la cantidad de ejemplos en este problema en particular, determinando así, si los resultados obtenidos son causa directa de la cantidad de ejemplos generados o de la dificultad del problema en sí mismo.

- Probar nuevos algoritmos de aprendizaje. Dado que los mejores resultados

no se obtienen con métodos que construyen clasificadores «legibles», y luego de levantada esta restricción, hay otros métodos y técnicas distintas a *boosting* que podrían dar mejores resultados.

Inclusive, la técnica de *boosting* podría ser implementada con clasificadores débiles más complejos que los *decision stumps* —por ejemplo, árboles de decisión de profundidad fija—, que han dado buenos resultados en otras tareas de procesamiento automático de texto.

Otra opción es la construcción de clasificadores combinando los resultados de distintas algoritmos. Esta aproximación también ha resultado exitosa en la resolución de problemas de procesamiento del lenguaje natural.

- Construir un clasificador para el idioma inglés. Esto permitiría comparar resultados con otros trabajos que, si bien no implementaron un evaluador de comas, analizaron y cuantificaron su uso en este idioma.

  A modo de ejemplo, se puede replicar la solución construida utilizando la clasificación de Bayraktar et al., y comparar los resultados del clasificador con su medida de estabilidad y con los patrones detectados y clasificados de forma manual por los autores.

Finalmente, el analizador sintáctico implementado con reglas contextuales admite varias mejoras.

Como casi todo sistema constituido por reglas escritas a partir del análisis de un corpus y la generalización de los fenómenos en él encontrados, es posible aumentar el número de reglas para contemplar fenómenos no considerados y, al mismo tiempo, afinar las reglas existentes para aumentar la precisión general del sistema.

Además, dado que sólo se elige la categoría más probable asociada a cada coma, no se explota que el clasificador devuelve un vector de «confianza» en la predicción de cada etiqueta. Luego, es posible implementar un analizador probabilístico que tome en cuenta el vector de predicción y arme en consecuencia los posibles análisis.

Por otra parte, queda por integrar el clasificador y los módulos de reglas escritos con el analizador de proposiciones del proyecto Clatex, evaluando cuáles reglas permiten mejorar el rendimiento de ambas aplicaciones. Esta tarea no es sencilla: la interacción entre las distintas reglas debe ser cuidadosamente evaluada, dado que no todas las estructuras fueron contempladas de igual forma en ambos trabajos.

Apéndice A

Marcadores y verbos discursivos

A continuación se listan los marcadores y verbos discursivos utilizados para la generación del conjunto de entrenamiento del clasificador (sección 3.2).

A.1. Marcadores discursivos

además	así	después	después de todo
más aún	al contrario	a su vez	ante todo
ciertamente	en cambio	en concreto	en consecuencia
en definitiva	en ese sentido	en fin	en general
en particular	en primer lugar	en tanto	en todo caso
es decir	es cierto	es más	de este manera
del contrario	de hecho	de todos modos	de todas maneras
de todas formas	en caso contrario	en síntesis	entonces
ergo	finalmente	hasta aquí	luego
no obstante	mientras tanto	para colmo	para ello
para redondear	para entonces	pero	pero no
por contraste	por ejemplo	por el contrario	por eso
por lo demás	por otra parte	por si fuera poco	por supuesto
por tanto	por último	pues bien	sin embargo
y por último	y finalmente		

A.2. Verbos discursivos

aclarar	afirmar	agregar	anunciar	aportar
apuntar	asegurar	comentar	concluir	confirmar
considerar	contar	contestar	creer	decir
declarar	denunciar	destacar	escribir	estimar
explicar	expresar	indicar	ironizar	observar
pensar	preguntar	recordar	relatar	

Apéndice B

Elección de atributos

El conjunto de características que conforma cada uno de los ejemplos de entrenamiento presentado en el capítulo 3 se construye de forma progresiva, agregando o transformando los atributos existentes en un paso previo y evaluando su desempeño en el aprendizaje.

El proceso se divide en diez experimentos que utilizan desde un conjunto de atributos extremadamente sencillos hasta los atributos utilizados por el clasificador final, presentado en el capítulo 3.

Los experimentos se dividen en dos etapas:

- Una etapa inicial —los ocho primeros experimentos— en donde se utilizan los 1174 ejemplos extraídos del corpus CORIN, previa clasificación manual de sus comas. Las categorías consideradas son las de la versión preliminar propuesta en la sección 2.3.1.
- Una segunda etapa —noveno y décimo experimento— en donde se utiliza la categorización definitiva de la sección 2.3.2, junto con los 5208 ejemplos del conjunto final de entrenamiento.

En todos los experimentos salvo el último se emplea el etiquetador morfosintáctico *Xelda*[1] [70] y se entrena únicamente con el algoritmo C4.5.

La evaluación de los clasificadores obtenidos se realiza utilizando validación cruzada de tamaño diez. Lamentablemente, los resultados obtenidos en estas pruebas no siempre son comparables entre sí debido a una o varias de las siguientes razones:

1. Las primeras pruebas se realizan con la categorización preliminar, sensiblemente distinta a la final, utilizada en los últimos dos experimentos.
2. El conjunto de entrenamiento en la primera etapa es distinto y significativamente menor —es cerca del 22 % del utilizado en la segunda etapa—.

[1] Se utiliza la versión 1.7.12 provista por Xerox. Actualmente, Xelda es un producto del grupo TEMIS.

3. El analizador morfosintáctico utilizado en todas las pruebas salvo la última no es *Freeling* [10] sino *Xelda*; en consecuencia, el conjunto de etiquetas no es el mismo. Debido a las características que diferencia cada etiquetador, no resulta factible traducir automáticamente las etiquetas de uno al otro. Además, y en parte debido a esto último, los errores cometidos por cada etiquetador —que afectan la calidad de los atributos del conjunto de entrenamiento— son de origen diferente.

Sin embargo, se considera relevante detallar estos resultados para ejemplificar cómo la variación en los atributos considerados influye en la resolución del problema de clasificación de comas.

A continuación se presentan los atributos utilizados en cada una de los experimentos y los resultados con ellos obtenidos. En el cuadro B.1 se encuentran las tasas de precisión de los árboles y reglas resultantes.

B.1. Etapa I

En el primer experimento se decide utilizar únicamente una ventana de etiquetas morfosintácticas por ser la única información con la que se cuenta de forma directa. Sin embargo, dados los resultados obtenidos, se estima que no se pueden lograr niveles aceptables de clasificación si no se agregan nuevas características. Éstas se calculan a partir del texto de entrada o de las etiquetas morfosintácticas. Se utiliza, además, una tabla de verbos discursivos para los atributos que así lo requieren (ver apéndice A).

B.1.1. Experimento 1

El tamaño de la ventana se elige arbitrariamente de tamaño nueve, las seis etiquetas anteriores y las tres posteriores a la coma, siempre y cuando ocurran dentro de la misma oración. Como se explica en la sección 3.2, en caso de no haber suficientes etiquetas dentro de la oración, se completa la ventana con la etiqueta de los delimitadores de oración («SENT»).

En este experimento se obtienen muy bajas tasas de precisión: 53,7 % para el árbol de decisión y 51,6 % para las reglas creadas a partir del árbol.

B.1.2. Experimento 2

Como primera prueba de atributos distintos a etiquetas morfosintácticas, se decide utilizar tres atributos *booleanos*, fácilmente calculables:

- *primera*: indica si la coma es la primera de la oración.
- *ultima*: indica si la coma es la última de la oración.
- *hay conjunción*: marca la presencia de una conjunción luego de la coma, sin intermediar ningún otro signo de puntuación entre ellas.

# Exp.	Árbol	Reglas
1	53,7	51,6
2	57,1	55,9
3	58,5	57,2
4	57,4	55,9
5	57,3	55,9
6	58,9	58,5
7	59,3	58,2
8	58,9	58,8
9	58,8	58,7
10	61,4	60,9

Cuadro B.1: Precisión para el árbol y las reglas discriminada por experimento

Los experimentos muestran una mejora de más de tres puntos en la precisión respecto a los resultados previos: 57,1 % y 55,9 % para el árbol y el conjunto de reglas respectivamente.

B.1.3. Experimento 3

Se agregan, entonces, nuevos atributos a los ejemplos de entrenamiento:

- *verbo discursivo antes*: atributo *booleano* que indica la presencia de un verbo antes que la coma.
- *verbo discursivo después*: análogo al atributo anterior, pero si el verbo se encuentra luego de la coma.
- *patrón*: es un atributo numérico que toma el valor de la cantidad de repeticiones de una misma etiqueta morfosintáctica alrededor de la coma.

Además, se amplía el tamaño de la ventana de nueve a diez categorías, dejándola simétrica respecto a la coma (cinco etiquetas anteriores y cinco posteriores). Esta decisión se toma en base a los árboles y reglas obtenidos en los experimentos previos, en los cuales la sexta categoría anterior a la coma nunca es evaluada para realizar la clasificación, y rara vez lo es la quinta.

Los resultados son levemente mejores a los obtenidos con los atributos anteriores, con una precisión del 58,5 % para el árbol y del 57,2 % para las reglas.

B.1.4. Experimento 4

En el siguiente experimento se colapsan las categorías «*verbo discursivo antes*» y «*verbo discursivo después*» en un único atributo «*verbo discursivo*», que indica la ocurrencia de este tipo de verbo, ya sea antes o después de la coma. Por otra

parte, se agrega el atributo «*próxima*», que mide en unidades léxicas la distancia hasta la coma más cercana (anterior o posterior).

La precisión en este caso empeora levemente (cerca de un punto porcentual), con tasas del 57,4 % para el árbol y del 55,9 % para las reglas.

B.1.5. Experimento 5

Se decide modificar el atributo «*patrón*», el cual deja de ser un valor numérico (la cantidad que se repite un patrón) a ser una categoría morfosintáctica (la etiqueta que se repite).

La precisión vuelve a los niveles previos: en este caso son 57,3 % para el árbol de decisión y 55,9 % para las reglas.

B.1.6. Experimento 6

Luego, se introduce una nueva modificación en el atributo «*patrón*». El valor deja de ser cualquier etiqueta morfosintáctica, y pasa a tomar un valor de un conjunto más reducido: adjetivo, nombre común, nombre propio, otros y ninguno.

Los experimentos muestran un aumento de las tasas de precisión que pasan a 58.9 % y 58.5 % para el árbol y las reglas respectivamente. Se observa que las tasas en el árbol y en las reglas están en niveles prácticamente iguales, mientras que en todos los experimentos anteriores el árbol siempre es superior.

B.1.7. Experimento 7

El atributo «*verbo discursivo*» es sustituido por el atributo «*verbo*», que toma por valor «DISCURSO», «OTRO» o «NINGUNO» de acuerdo a si existe un verbo discursivo, un verbo no discursivo o ningún verbo, antes o después de la coma.

Los resultados obtenidos no muestran significativas variaciones respecto al experimento anterior: 59,3 % para el árbol y 58,2 % para las reglas.

B.1.8. Experimento 8

Finalmente, se discrimina entre la ocurrencia de un verbo antes o luego de la coma sustituyendo al atributo «*verbo*» por «*verbo anterior*» y «*verbo posterior*». Los nuevos atributos toman los mismos valores que el atributo sustituido.

Nuevamente, las reglas y el árbol muestran un nivel de precisión muy parecido: 58.9 %, y 58.8 % respectivamente.

B.2. Etapa II

En estas pruebas se utiliza la categoría definitiva de las comas, con lo cual los resultados no son directamente comparables a los anteriores. Por otra parte, el

etiquetador morfosintáctico difiere entre los dos experimentos de esta etapa: en el primero sigue siendo *Xelda*, mientras que el segundo es *Freeling*[2].

B.2.1. Experimento 9

Se agregan los atributos «*categoría anterior*» y «*categoría posterior*», que toman por valor la etiqueta morfosintáctica de la unidad que se encuentra inmediatamente antes de la coma anterior e inmediatamente posterior a la categoría que se encuentra en la coma posterior[3]

Los resultados son del 58,8 % para el árbol de decisión y del 58,7 % para las reglas.

B.2.2. Experimento 10

Este experimento corresponde con la versión final del clasificador, hecha con la nueva clasificación de las comas y las etiquetas de *Freeling*. Se diferencia del experimento anterior en que se agrega el atributo «*conector*», el cual indica la presencia de un conector discursivo precediendo a la coma.

El árbol resultante tiene una tasa de precisión del 61,4 %, mientras que para las reglas es del 60,9 %.

[2]El cambio de analizador se debe al vencimiento de la licencia de *Xelda*.
[3]El por qué de este atributo se explica en la sección 3.2

Apéndice C

Reglas contextuales

En este apéndice se listan las reglas contextuales que conforman el analizador sintáctico de superficie.

El formato de las reglas impuesto por el intérprete no es claro: es una lista de pares atributo–valor, donde el atributo queda determinado por su posición en la lista[1]. Por esto, al igual que en el capítulo 4, se opta por utilizar etiquetas para la especificación de las reglas.

A modo de ejemplo:

- una coma de inciso inicial se representa en el intérprete como:

  [limFlex, ',', ',', punt, inciso, inicial | _]

  mientras que en las reglas se la denota con la etiqueta «CMII».

- En caso que se admita en la regla cualquier tipo de coma —sin importar si es de inciso, serie, etc.— se utiliza «CM», lo que en el intérprete se traduce como:

  [limFlex, ',', ',', punt | _]

En la sección C.1 se listan las categorías utilizadas y en la sección C.2 las reglas agrupadas por módulos.

C.1. Etiquetas

Aunque todas las etiquetas son tratadas de igual forma por el intérprete, es posible distinguir entre dos clases: las etiquetas morfosintácticas y las etiquetas de segmentos.

[1] Se representa como una lista incompleta de Prolog.

ADJ	adjetivo
ADV	adverbio
CM	coma
CMCO	coma de inciso «conectivo»
CMI	coma de inciso
CMII	coma de inciso «inicial»
CMID	coma de inciso «de discurso»
CMIF	coma de inciso «final»
CMMO	coma de inciso «modificador»
CMSO	coma de series simples
CMSP	coma de serie proposicional
CONJ	conjunción: *y*, *o*, etc.
CONJBIP	conjunción: *si*, *aunque*, etc.
DET	determinante
ETC	etcétera
NOM	nombre común o propio
NOMC	nombre común
NOMP	nombre propio
PRON	pronombre
PREP	preposición
PUNT	signo de puntuación (comillas, rayas, etc)
SENT	límite de oración: punto, punto y coma, etc.
VERB	verbo
VERBFIN	verbo conjugado
VERBFINDISC	verbo discursivo conjugado
VPART	participio de verbo

Cuadro C.1: Etiquetas morfosintácticas

Las primeras se obtienen a partir del analizador morfosintáctico, *Freeling*, o del clasificador de comas, y se convierten al formato interno del intérprete durante la lectura del archivo de entrada (cuadro C.1). En las reglas se denotan en mayúsculas.

Las segundas son el resultado de la aplicación exitosa de alguna regla contextual, esto es, son el lado izquierdo de alguna de éstas[2] (cuadro C.2). Aparecen en minúsculas en la formulación de las reglas.

C.2. Reglas

A continuación se especifican las reglas contextuales, agrupadas según los módulos a los que pertenecen; estos últimos se encuentran ordenados por su

[2]Esto no quita que puedan ocurrir además en el lado derecho de esa u otra regla.

inciso	inciso
inciso$_{est}$	inciso marcado durante el módulo de análisis final
inciso$_{ini}$	inciso inicial
inciso$_{ini-con}$	inciso inicial «conectivo»
inciso$_{ini-mod}$	inciso inicial «modificador»
inciso$_{punt-con}$	inciso «conectivo» dentro de rayas o paréntesis
inciso$_{punt-mod}$	inciso «modificador» dentro de rayas o paréntesis
inciso$_{punt}$	inciso dentro de rayas o paréntesis
inciso$_{punt-con}$	inciso «conectivo» dentro de rayas o paréntesis
inciso$_{punt-mod}$	inciso «modificador» dentro de rayas o paréntesis
nom$_{dp}$	nombre, posible elemento de una serie delimitada por dos puntos
nom$_{dp-det}$	nombre, posible elemento con determinante de serie delimitada por dos puntos
nom$_{dp-nodet}$	nombre, posible elemento sin determinante de serie delimitada por dos puntos
prop	segmento proposicional
prop$_{bip}$	segmento proposicional perteneciente a una estructura «bipolar»
prop$_{est}$	segmento proposicional marcado durante el módulo de análisis final
serie	serie
serie$_{-}$	serie simple entre rayas
serie$_{()}$	serie simple entre paréntesis
serie$_{:-det}$	serie simple con elementos con determinante, delimitada por un dos puntos
serie$_{:-nodet}$	serie simple con elementos sin determinante, delimitada por un dos puntos
serie$_{ac-det}$	serie simple con elementos con determinante (el auto, la casa...)
serie$_{nomc-nodet}$	serie de nombres comunes con elementos sin determinante
serie$_{nomp-nodet}$	serie de nombres propios con elementos sin determinante (juan, maría...)
serie$_{verb-det}$	serie de verbos con determinante
serie$_{verb-nodet}$	serie de verbos sin determinante

Cuadro C.2: Etiquetas de segmentos

orden de ejecución.

La separación de las reglas que se realiza en cada uno de los módulos puede resultar arbitraria, pero su único fin es el de facilitar la lectura agrupando reglas que se comportan de forma parecida o contemplan un mismo fenómeno.

C.2.1. Series simples

Nombres

$$\text{serie}_{ac-det} \rightarrow \text{CMSO} \setminus \text{DET} *(S_1, 5)\ \text{OP(inciso)}\ \text{CONJ}\ \text{DET} *(S_2, 5)\ \text{NOM} *(S_2, 5)\ \text{NOM}\ \text{OP(NOM)}\ \text{OP(inciso)}\ /;$$
$$S_1 = \{\text{CONJ, SENT, CM, VERBFIN, DET, serie}\}$$
$$S_2 = \{\text{DET, CONJ, CM, SENT, VERBFIN, NOM, serie }\}$$

$$\text{serie}_{ac-det} \rightarrow \text{CMSO} \setminus \text{DET} *(S_1, 5)\ \text{OP(inciso)}\ \text{CONJ}\ \text{DET} *(S_2, 5)\ \text{NOM}\ \text{OP(NOM)}\ \text{OP(inciso)}\ /;$$
$$S_1 = \{\text{CONJ, SENT, CM, VERBFIN, DET, serie}\}$$
$$S_2 = \{\text{DET, CONJ, CM, SENT, VERBFIN, NOM, serie }\}$$

$$\text{serie}_{ac-det} \rightarrow \setminus\ \text{DET} *(S, 5)\ \text{OP(inciso)}\ \text{DET} *(S, 5)\ \text{OP(inciso)}\ \text{CMSO}\ \text{serie}_{ac-det}\ /;$$
$$S = \{\text{CONJ, SENT, CM, VERBFIN, DET, serie, PREP, PUNT}\}$$

$$\text{serie}_{ac-det} \rightarrow \setminus\ \text{DET} *(S, 10)\ \text{OP(inciso)}\ \text{CMSO}\ \text{serie}_{ac-det}\ /;$$
$$S = \{\text{CONJ, SENT, CM, VERBFIN, DET, serie, PREP, PUNT}\}$$

$$\text{serie}_{ac-nodet-nomp} \rightarrow \text{CMSO} \setminus *(S_1, 5)\ \text{NOMP} *(S_2, 5)\ \text{OP(inciso)}\ \text{CONJ} *(S_2, 5)\ \text{NOMP} *(S_2, 5)\ \text{NOMP}\ \text{OP(NOM)}\ /;$$
$$S_1 = \{\text{CONJ, SENT, CM, VERBFIN, DET, serie}\}$$
$$S_2 = \{\text{CONJ, CM, SENT, VERBFIN, NOM, serie }\}$$

$$\text{serie}_{ac-nodet-nomc} \rightarrow \text{CMSO} \setminus *(S, 5)\ \text{NOMC} *(S, 5)\ \text{NOMC} *(S, 5)\ \text{OP(inciso)}\ \text{CONJ} *(S, 5)\ \text{NOMC} *(S, 5)\ \text{NOMC}\ \text{OP(NOMC)}\ /;$$
$$S = \{\text{CONJ, CM, SENT, VERBFIN, NOM, serie }\}$$

$$\text{serie}_{ac-nodet-nomc} \rightarrow \text{CMSO} \setminus *(S_1, 5)\ \text{NOMC} *(S_2, 5)\ \text{OP(inciso)}\ \text{CONJ}\ \text{OP(PREP)}\ \text{OP(ADV)}\ \text{NOMC}\ \text{OP(NOMC)}\ \text{OP(inciso)}\ /;$$
$$S_1 = \{\text{CONJ, CM, SENT, VERBFIN, NOM, serie }\}$$
$$S_2 = \{\text{CONJ, CM, SENT, VERBFIN, DET, serie }\}$$

$$\text{serie}_{ac-nodet-nomc} \rightarrow \setminus\ \text{NOMC} *(S, 5)\ \text{OP(NOMC)}\ \text{NOMC} *(S, 5)\ \text{OP(inciso)}\ \text{CMSO}\ \text{serie}_{ac-nodet-nomc}\ /;$$
$$S = \{\text{CONJ, CM, SENT, VERBFIN, NOM, serie }\}$$

$$\text{serie}_{ac-nodet-nomc} \rightarrow \setminus\ \text{OP(NOMC)}\ \text{NOMC} *(S, 5)\ \text{OP(inciso)}\ \text{CMSO}\ \text{serie}_{ac-nodet-nomc}\ /;$$
$$S = \{\text{CONJ, CM, SENT, VERBFIN, NOM, serie }\}$$

$$\text{serie}_{ac-det-nom} \rightarrow \text{CMSO} \setminus \text{OP(ADV)}\ \text{OP(PREP)}\ \text{OP(ADV)}\ \text{DET} *(S, 5)$$

DET $*(S,5)$ CM ETC /;
$S = \{$PUNT, CONJ, CM, SENT, VERBFIN, NOM, serie $\}$

$\text{serie}_{ac-det-nom} \rightarrow$ CMSO $\setminus$ OP(ADV) OP(PREP) OP(ADV) DET $*(S,5)$ CM ETC /;
$S = \{$PUNT, CONJ, CM, SENT, VERBFIN, NOM, serie $\}$

$\text{serie}_{ac-nodet-nom} \rightarrow$ CMSO $\setminus$ OP(ADV) OP(PREP) OP(ADV) OP(NOM) OP(NOM) NOM $*(S,5)$ OP(NOM) OP(NOM) NOM $*(S,5)$ CM ETC /;
$S = \{$PUNT, CONJ, CM, SENT, VERBFIN, NOM, PREP, serie $\}$

$\text{serie}_{ac-nodet-nom} \rightarrow$ CMSO $\setminus$ OP(ADV) OP(PREP) OP(ADV) OP(NOM) OP(NOM) NOM $*(S,5)$ CM ETC /;
$S = \{$PUNT, CONJ, CM, SENT, VERBFIN, NOM, PREP, serie $\}$

Entre rayas

$\text{serie}_{\_} \rightarrow$ — $*(S_1,50)$ CMSO $\setminus *(S_2,5)$ NOM $*(S_3,5)$/ — ;
$S_1 = \{$PUNT, SENT$\}$
$S_2 = \{$PUNT, CONJ, CM, SENT, VERBFIN, NOM, serie$\}$
$S_3 = \{$PUNT, CONJ, CM, SENT, VERBFIN, DET, serie$\}$

$\text{serie}_{\_} \rightarrow$ — $*(S_1,50)$ CMSO $\setminus *(S_2,5)$ NOM $*(S_3,5)$ CONJ $*(S_4,5)$ / —;
$S_1 = \{$PUNT, SENT$\}$
$S_2 = \{$PUNT, CONJ, CM, SENT, VERBFIN, NOM, serie$\}$
$S_3 = \{$PUNT, CONJ, CM, SENT, VERBFIN, DET, serie$\}$
$S_4 = \{$PUNT, CONJ, CM, SENT, serie$\}$

$\text{serie}_{\_} \rightarrow$ CMSO $\setminus *(S_1,5)$ NOM $*(S_2,5)$ CMSO $\text{serie}_{\_}$ /;
$S_1 = \{$CONJ, PUNT, CM, SENT, serie$\}$
$S_2 = \{$NOM, CONJ, PUNT, CM, SENT, serie$\}$
serie $\rightarrow$ — $\setminus *(S_1,5)$ NOM $*(S_2,5)$ CMSO $\text{serie}_{\_}$ /;
$S_1 = \{$CONJ, PUNT, CM, SENT, serie$\}$
$S_2 = \{$NOM, CONJ, PUNT, CM, SENT, serie$\}$

Entre paréntesis

$\text{serie}_{()} \rightarrow$ ($*(S_1,50)$ CMSO $\setminus *(S_2,5)$ NOM $*(S_3,5)$ /) ;
$S_1 = \{$PUNT, SENT$\}$
$S_2 = \{$PUNT, CONJ, CM, SENT, VERBFIN, NOM, serie$\}$
$S_3 = \{$PUNT, CONJ, CM, SENT, VERBFIN, DET, serie$\}$

$\text{serie}_{()} \rightarrow$ ($*(S_1,50)$ CMSO $\setminus *(S_2,5)$ NOM $*(S_3,5)$ CONJ $*(S_4,5)$ /);
$S_1 = \{$PUNT, SENT$\}$
$S_2 = \{$PUNT, CONJ, CM, SENT, VERBFIN, NOM, serie$\}$
$S_3 = \{$PUNT, CONJ, CM, SENT, VERBFIN, DET, serie$\}$
$S_4 = \{$PUNT, CONJ, CM, SENT, serie$\}$

$\text{serie}_{()} \rightarrow$ CMSO $\setminus *(S_1, 5)$ NOM $*(S_2, 5)$ CMSO $\text{serie}_{()}$ /;
$S_1 = \{$CONJ, PUNT, CM, SENT, serie$\}$
$S_2 = \{$NOM, CONJ, PUNT, CM, SENT, serie$\}$

serie $\rightarrow$ ($\setminus *(S_1, 5)$ NOM $*(S_2, 5)$ CMSO $\text{serie}_{()}$ /;
$S_1 = \{$CONJ, PUNT, CM, SENT, serie$\}$
$S_2 = \{$NOM, CONJ, PUNT, CM, SENT, serie$\}$

Dos puntos

$\text{nom}_{dp-nodet} \rightarrow$: $\setminus *(S, 5)$ NOM / ;
$S = \{$NOM, DET, PUNT, CM, SENT, VERBFIN, serie$\}$

$\text{nom}_{dp-nodet} \rightarrow$: $*(S_1, 50) \setminus$ CMSO $*(S_2, 5)$ NOM / ;
$S_1 = \{$PUNT, SENT$\}$
$S_2 = \{$NOM, DET, PUNT, CM, SENT, VERBFIN, serie$\}$

$\text{nom}_{dp-det} \rightarrow$: $\setminus *(S, 5)$ DET $*(S, 5)$ NOM / ;
$S = \{$NOM, DET, PUNT, CM, SENT, VERBFIN, serie$\}$

$\text{nom}_{dp-det} \rightarrow$: $*(S_1, 50) \setminus$ CMSO $*(S_2, 5)$ NOM / ;
$S_2 = \{$PUNT, SENT$\}$
$S_2 = \{$NOM, DET, PUNT, CM, SENT, VERBFIN, serie$\}$

$\text{serie}_{:-det} \rightarrow$ CMSO $\setminus *(S_1, 5)$ NOM_{dp-det} $*(S_2, 20)$ / SENT ;
$S_1 = \{$NOM, CONJ, PUNT, CM, SENT, VERBFIN, serie$\}$
$S_2 = \{\text{NOM}_{dp}$, CONJ, PUNT, CM, SENT, serie$\}$

$\text{serie}_{:-det} \rightarrow \setminus *(S_1, 5)$ NOM_{dp-det} $*(S_2, 20)$ CMSO $\text{serie}_{:-det}$ /;
$S_1 = \{$NOM, CONJ, PUNT, CM, SENT, VERBFIN, serie$\}$
$S_2 = \{\text{NOM}_{dp}$, CONJ, PUNT, CM, SENT, serie$\}$

$\text{serie}_{:-nd} \rightarrow$ CMSO $\setminus *(S_1, 5)$ $\text{NOM}_{dp-nodet}$ $*(S_2, 20)$ / SENT ;
$S_1 = \{$NOM, CONJ, PUNT, CM, SENT, VERBFIN, serie$\}$
$S_2 = \{\text{NOM}_{dp}$, CONJ, PUNT, CM, SENT, serie$\}$

$\text{serie}_{:-nd} \rightarrow \setminus *(S_1, 5)$ $\text{NOM}_{dp-nodet}$ $*(S_2, 20)$ CMSO $\text{serie}_{:-nd}$ /;
$S_1 = \{$NOM, CONJ, PUNT, CM, SENT, VERBFIN, serie$\}$
$S_2 = \{\text{NOM}_{dp}$, CONJ, PUNT, CM, SENT, serie$\}$

Verbos

$\text{serie}_{verb-det} \rightarrow$ CMSO $\setminus$ DET VERB $*(S, 5)$ CONJ DET $*(S, 5)$ VERB /;
$S = \{$DET, CONJ, CM, SENT, VERBFIN, NOM, serie $\}$

$\text{serie}_{verb-det} \rightarrow \setminus$ DET $*(S, 5)$ VERB OP(inciso) CMSO $\text{serie}_{verb-det}$ /;
$S = \{$DET, CONJ, CM, SENT, serie $\}$

$\text{serie}_{verb-nodet} \rightarrow$ CMSO $\setminus$ VERB $*(S, 5)$ CONJ VERB /;

$S = \{$PUNT, CONJ, CM, SENT, VERB, NOM, serie $\}$

serie$_{verb-nodet}$ $\rightarrow$ $\backslash$ VERB $*(S, 5)$ CMSO serie$_{verb-nodet}$ /;
$S = \{$PUNT, CONJ, CM, SENT, serie $\}$

C.2.2. Incisos dentro de paréntesis o rayas

Iniciales

inciso$_{punt-con}$ $\rightarrow$ ($\backslash$ $*(S_1, 20)$ CMCO / $*(S_2, 20)$) ;
$S_1 = \{$ (, CM, SENT $\}$
$S_2 = \{$), SENT $\}$

inciso$_{punt-con}$ $\rightarrow$ — $\backslash$ $*(S_1, 20)$ CMCO / $*(S_2, 20)$ — ;
$S_1 = \{$ —, CM, SENT $\}$
$S_2 = \{$ —, SENT $\}$

inciso$_{punt-mod}$ $\rightarrow$ ($\backslash$ $*(S_1, 20)$ CMMO / $*(S_2, 20)$) ;
$S_1 = \{$ (, CM, SENT $\}$
$S_2 = \{$), SENT $\}$

inciso$_{punt-mod}$ $\rightarrow$ — $\backslash$ $*(S_1, 20)$ CMMO / $*(S_2, 20)$ — ;
$S_1 = \{$ —, CM, SENT $\}$
$S_2 = \{$ —, SENT $\}$

inciso$_{punt-con}$ $\rightarrow$ inciso$_{punt}$ $\backslash$ $*(S_1, 20)$ CMCO / ;
$S_1 = \{$ inciso$_{punt}$, CM, SENT $\}$

inciso$_{punt-mod}$ $\rightarrow$ inciso$_{punt}$ $\backslash$ $*(S_1, 20)$ CMMO / ;
$S_1 = \{$ inciso$_{punt}$, CM, SENT $\}$

Medios y finales

inciso $\rightarrow$ ($*(S_1, 20)$ $\backslash$ CMII $*(S_2, 10)$ /) ;
$S_1 = \{$ (, SENT $\}$
$S_2 = \{$), inciso, CM, SENT, VERBFIN $\}$

inciso $\rightarrow$ — $*(S_1, 20)$ $\backslash$ CMII $*(S_2, 10)$ / — ;
$S_1 = \{$ —, SENT $\}$
$S_2 = \{$ —, inciso, CM, SENT, VERBFIN $\}$

inciso $\rightarrow$ ($*(S_1, 20)$ $\backslash$ CMII OP(DET) PRON $*(S_2, 30)$ /) ;
$S_1 = \{$ (, SENT $\}$
$S_2 = \{$), inciso, CM, SENT, PRON $\}$

inciso $\rightarrow$ — $*(S_1, 20)$ $\backslash$ CMII OP(DET) PRON $*(S_2, 30)$ / — ;
$S_1 = \{$ —, SENT $\}$
$S_2 = \{$ —, inciso, CM, SENT, PRON $\}$

inciso $\rightarrow$ ($*(S_1, 20)$ $\backslash$ CMII OP(DET) PRON $*(S_2, 20)$ OP(DET) PRON

$*(S_2, 20)$ /) ;
$S_1 =$ { (, SENT }
$S_2 =$ {), inciso, CM, SENT, PRON }

inciso → — $*(S_1, 20)$ \ CMII OP(DET) PRON $*(S_2, 20)$ OP(DET) PRON $*(S_2, 20)$ / — ;
$S_1 =$ { —, SENT }
$S_2 =$ { —, inciso, CM, SENT, PRON }

inciso → ($*(S_1, 20)$ \ CMII VERBFIN $*(S_2, 30)$ /) ;
$S_1 =$ { (, SENT }
$S_2 =$ {), inciso, CM, SENT, VERBFIN }

inciso → — $*(S_1, 20)$ \ CMII VERBFIN $*(S_2, 30)$ / — ;
$S_1 =$ { —, SENT }
$S_2 =$ { —, inciso, CM, SENT, VERBFIN }

C.2.3. Incisos entre paréntesis o rayas

inciso → \ ($*(S, 50)$) / ;
$S =$ { (, SENT }

inciso → \ — $*(S, 50)$ — / ;
$S =$ { —, SENT }

C.2.4. Estructuras bipolares

prop_{bip} → SENT\ CONJBIP $*(S, 50)$ VERBFIN $*(S, 50)$ / CM $*(S, 50)$ VERBFIN $*(S, 50)$ SENT;
$S =$ {CONJBIP, CM, VERBFIN, SENT}

prop_{bip} → prop_{bip}\ CM $*(S, 50)$ VERBFIN $*(S, 50)$ / SENT;
$S =$ {CONJBIP, CM, VERBFIN, SENT}

prop_{bip} → CMSP\OP(CONJ) CONJBIP $*(S, 50)$ VERBFIN $*(S, 50)$ / CM $*(S, 50)$ VERBFIN $*(S, 50)$ SENT;
$S =$ {CONJBIP, CM, VERBFIN, SENT}

C.2.5. Incisos

Iniciales

$\text{inciso}_{ini-con}$ → SENT \ $*(S, 20)$ CMCO / ;
$S =$ {CM, SENT}

$\text{inciso}_{ini-mod}$ → SENT \ $*(S, 20)$ CMMO / ;
$S =$ {CM, SENT}

$\text{inciso}_{ini-con}$ → inciso_{ini}\ $*(S, 20)$ CMCO / ;

$S = \{\text{inciso}_{ini}, \text{CM}, \text{SENT}\}$

$\text{inciso}_{ini-mod} \rightarrow \text{inciso}_{ini} \backslash *(S, 20)$ CMMO / ;
$S = \{\text{inciso}_{ini}, \text{CM}, \text{SENT}\}$

$\text{inciso}_{ini} \rightarrow$ SENT $\backslash$ PREP $*(S_1, 20)$ CMI / $*(S_2, 20)$ CM NO(VERBFIN);
$S_1 = \{\text{inciso}, \text{VERBFIN}, \text{PREP}, \text{CM}, \text{SENT}\}$
$S_2 = \{\text{inciso}, \text{CM}, \text{SENT}\}$

$\text{inciso}_{ini} \rightarrow$ SENT $\backslash$ PREP $*(S_1, 20)$ CMI / $*(S_2, 20)$ VERBFIN $*(S_2, 20)$ SENT;
$S_1 = \{\text{inciso}, \text{VERBFIN}, \text{PREP}, \text{CM}, \text{SENT}\}$
$S_2 = \{\text{inciso}, \text{CM}, \text{SENT}, \text{VERBFIN}\}$

Sin verbo

inciso $\rightarrow \backslash$ CMII $*(S, 10)$ CMIF/;
$S = \{\text{inciso}, \text{PUNT}, \text{CM}, \text{VERBFIN}, \text{SENT}\}$

inciso $\rightarrow \backslash$ CMII $*(S, 10)$ /SENT;
$S = \{\text{inciso}, \text{PUNT}, \text{CM}, \text{VERBFIN}, \text{SENT}\}$

Con pronombres

inciso $\rightarrow \backslash$ CMII OP(DET) OP(NOM) PRON $*(S, 20)$ CMIF /;
$S = \{$ inciso, CM, SENT, PRON, PUNT$\}$

inciso $\rightarrow \backslash$ CMI OP(PRON) OP(DET) PRON *(S,30) /SENT;
$S = \{$ inciso, CM, SENT, PRON, PUNT$\}$

inciso $\rightarrow \backslash$ CMII OP(DET) PRON $*(S, 20)$ PRON $*(S, 20)$ CMIF /;
$S = \{$ inciso, CM, SENT, PRON, PUNT$\}$

inciso $\rightarrow \backslash$ CMI OP(PRON) OP(DET) PRON $*(S, 30)$ PRON $*(S, 30)$ / SENT;
$S = \{$ inciso, CM, SENT, PRON, PUNT$\}$

Con participios

inciso $\rightarrow \backslash$ CMI VPART $*(S, 50)$ CMIF / ;
$S = \{\text{inciso}, \text{VERB}, \text{CM}, \text{SENT}, \text{PUNT}\}$

inciso $\rightarrow$ VERBFIN $*(S_1, 50) \backslash$ CMII VPART $*(S_2, 50)$ / SENT;
$S_1 = \{\text{VERB}, \text{SENT}\}$
$S_2 = \{\text{inciso}, \text{VERB}, \text{CM}, \text{SENT}, \text{PUNT}\}$

inciso $\rightarrow$ VERBFIN $*(S_1, 50)$ inciso $\backslash$ VPART $*(S_2, 50)$ / SENT;

$S_1 = \{$CMI, VERB, SENT $\}$
$S_2 = \{$inciso, VERB, CM, SENT, PUNT$\}$

Con adverbios

inciso $\rightarrow \setminus$ CMI ADV $*(S, 50)$ CMIF / ;
$S = \{$inciso, ADV, VERBFIN CM, SENT, PUNT$\}$

inciso $\rightarrow \setminus$ CMI ADV $*(S, 50)$ / SENT;
$S = \{$inciso, ADV, VERBFIN, CM, SENT, PUNT$\}$

inciso $\rightarrow$ VERBFIN $*(S_1, 50)$ inciso $\setminus$ VPART $*(S_2, 50)$ / SENT;
$S_1 = \{$CMI, VERB, SENT $\}$
$S_2 = \{$inciso, ADV, VERBFIN, CM, SENT, PUNT$\}$

De discurso

inciso $\rightarrow \setminus$ CMID $*(S, 20)$ VERBFINDISC $*(S, 20)$ / SENT ;
$S = \{$VERBFIN, CM, SENT$\}$

inciso $\rightarrow \setminus$ CMID $*(S, 20)$ VERBFINDISC $*(S, 20)$ / CM ;
$S = \{$VERBFIN, CM, SENT$\}$

inciso $\rightarrow$ inciso$_{ini} \setminus$ CMID $*(S_2, 5)$ VERBFINDISC $*(S, 5)$ CMID / ;
$S_1 = \{$inciso, VERBFIN, CM, SENT$\}$
$S_2 = \{$VERBFIN, CM, SENT$\}$

Otras

inciso $\rightarrow$ CMSP $\setminus *(S, 5)$ CMIF / ;
$S = \{$VERBFIN CM, SENT$\}$

inciso $\rightarrow \setminus$ CMII $*(S, 5)$ / CMSP ;
$S = \{$VERBFIN CM, SENT$\}$

inciso $\rightarrow$ inciso$_{ini} \setminus *(S, 10)$ CMIF / $*(S, 10)$ VERBFIN ;
$S = \{$inciso, VERBFIN, CM, SENT$\}$

inciso $\rightarrow \setminus$ CMII $*(S_1, 5)$ PRON $*(S_1, 10)$ VERBFIN $*(S_2, 10)$ CMIF / ;
$S_1 = \{$inciso, PRON, VERBFIN, CM, SENT$\}$
$S_2 = \{$inciso, VERBFIN, CM, SENT$\}$

inciso $\rightarrow \setminus$ CMI $*(S, 10)$ CMI / OP(ADV) OP(ADV) VERBFIN;
$S = \{$inciso, VERBFIN, CM, SENT$\}$

C.2.6. Series proposicionales

Separadas por comas

$\text{prop}_{conj} \rightarrow$ CMSP \ CONJ *(S_1,50) VERBFIN *(S_2,50) / SENT;
$S_1 = \{\text{prop, CONJ, CMSP, SENT}\}$
$S_2 = \{\text{prop, CMSP, VERBFIN, SENT}\}$

$\text{prop}_{sinconj} \rightarrow$ CMSP \ *(S_1,50) VERBFIN *(S_2,50)/ SENT;
$S_1 = \{\text{prop, CONJ, inciso}_{ini}\text{, CMSP, SENT}\}$
$S_2 = \{\text{prop, CMSP, VERBFIN, SENT}\}$

prop → CMSP \ CONJ *(S_1,20) VERBFIN *(S_2,20)/ SENT;
$S_1 = \{\text{prop, conj, CMSP, SENT}\}$
$S_2 = \{\text{prop, CMSP, VERBFIN, SENT}\}$

prop → SENT op(prop) op(CMSP) op(prop) op(CMSP) op(prop) CMSP \ CONJ *(S,20) / SENT;
$S = \{\text{prop, CONJ, CMSP, SENT}\}$

prop → CMSP \ CONJ *(S_1,50) VERBFIN *(S_2,50)/ CMSP;
$S_1 = \{\text{prop, CONJ, SENT}\}$
$S_2 = \{\text{prop, VERBFIN, SENT}\}$

prop → prop CMSP \ *(S_1,20) VERBFIN *(S_2,20)/ CMSP CONJ;
$S_1 = \{\text{prop, CONJ, CMSP, SENT}\}$
$S_2 = \{\text{prop, VERBFIN, CMSP, SENT}\}$

prop → SENT \ *(S_1,20) VERBFIN *(S_2,20)/ CMSP;
$S_1 = \{\text{prop, inciso}_{ini}\text{, CMSP, SENT}\}$
$S_2 = \{\text{prop, VERBFIN, CMSP, SENT}\}$

prop → SENT \ *(S_1,50) VERBFIN *(S_2,50)/ : *(S_1,50) ;
$S_1 = \{\text{prop, :, inciso}_{ini}\text{, CM, SENT}\}$
$S_2 = \{\text{prop, VERBFIN, CM, SENT}\}$

prop → SENT \ *(S_1,50) VERBFIN *(S_2,50)/ SENT ;
$S_1 = \{\text{prop, inciso}_{ini}\text{, CM, SENT}\}$
$S_2 = \{\text{prop, VERBFIN, CM, SENT}\}$

prop → SENT \ *(S_1,50) VERBFIN *(S_2,50)/ SENT;
$S_1 = \{\text{prop, inciso}_{ini}\text{, CM, SENT}\}$
$S_2 = \{\text{prop, VERBFIN, CM, SENT}\}$

Limitando con inciso

$\text{prop} \rightarrow \text{inciso}_{ini}$ \ *(S_1,20) VERBFIN *(S_2,20)/ CMSP;
$S_1 = \{\text{prop, inciso}_{ini}\text{, CMSP, SENT}\}$
$S_2 = \{\text{prop, VERBFIN, CMSP, SENT}\}$

$\text{prop} \rightarrow \text{inciso}_{ini}$ \ *(S_1,50) VERBFIN *(S_2,50)/ SENT;

$S_1 = \{\text{prop, inciso}_{ini}\text{, CM, SENT}\}$
$S_2 = \{\text{prop, VERBFIN, CM, SENT}\}$

Otra puntuación

prop → **:** \ *(S_1,50) VERBFIN *(S_2,50)/ SENT;
$S_1 = \{\text{prop, :, CM, SENT}\}$
$S_2 = \{\text{prop, CM, VERBFIN, SENT}\}$

prop → CMSP \ CONJ *(S_1,50) VERBFIN *(S_2,50)/ **:** ;
$S_1 = \{\text{prop, :, inciso}_{ini}\text{, CONJ, CM, SENT}\}$
$S_2 = \{\text{prop, VERBFIN, CM, SENT}\}$

prop → inciso_{ini} \ *(S_1,50) VERBFIN *(S_2,50)/ **:** ;
$S_1 = \{\text{prop, :, inciso}_{ini}\text{, CM, SENT}\}$
$S_2 = \{\text{prop, VERBFIN, CM, SENT}\}$

prop → **¿**\ *(S_1,50) VERBFIN *(S_2,50)/ **?**;
$S_1 = \{\text{prop, ¿, inciso}_{ini}\text{, CM, SENT}\}$
$S_2 = \{\text{prop, VERBFIN, CM, SENT}\}$

C.2.7. Análisis final

Inciso/Proposición

inciso_{est} → SENT\ *(S_1,50) CM / *(S_1,50) VERBFIN *(S_2,50) SENT;
$S_1 = \{\text{prop, CM, VERBFIN, SENT}\}$
$S_2 = \{\text{CM, SENT}\}$

inciso_{est} → SENT OP(inciso) OP(inciso) \ PREP *(S_1,50) CM / *(S_2,50) VERBFIN *(S_3,50) SENT;
$S_1 = \{\text{PREP, CM, prop, SENT}\}$
$S_2 = \{\text{VERBFIN, CM, prop, SENT}\}$
$S_3 = \{\text{CM, SENT}\}$

inciso_{est} → SENT OP(inciso) OP(inciso) \ PREP *(S_1,50) PREP *(S_1,50) CM / *(S_2,50) VERBFIN *(S_3,50) SENT;
$S_1 = \{\text{PREP, CM, prop, SENT}\}$
$S_2 = \{\text{VERBFIN, CM, prop, SENT}\}$
$S_3 = \{\text{CM, SENT}\}$

inciso_{est} → SENT \ PREP *(S_1,50) CM / *(S_2,50) VERBFIN *(S_3,50) SENT;
$S_1 = \{\text{PREP, CM, prop, SENT}\}$
$S_2 = \{\text{VERBFIN, CM, prop, SENT}\}$
$S_3 = \{\text{CM, SENT}\}$

prop_{est} → inciso_{est}\ *(S_1,50) VERBFIN *(S_2,50) / SENT;
$S_1 = \{\text{inciso}_{est}\text{, CM, prop, SENT}\}$
$S_2 = \{\text{CM, SENT, VERBFIN, prop }\}$

$\text{prop}_{est} \rightarrow \text{inciso}_{est} \backslash$ *(S$_1$,50) VERBFIN *(S$_2$,50) / prop;
$S_1 = \{$ inciso_{est}, CM, prop, SENT$\}$
$S_2 = \{$ CM, prop, VERBFIN, SENT$\}$

$\text{prop}_{est} \rightarrow$ SENT$\backslash$ *(S$_1$,50) VERBFIN *(S$_2$,50) /$\text{inciso}_{est-fin}$;
$S_1 = \{$prop, CM, SENT$\}$
$S_2 = \{$VERBFIN, prop, CM, SENT$\}$

$\text{inciso}_{est-fin} \rightarrow$ SENT *(S$_1$,50) VERBFIN *(S$_2$,50) $\backslash$ CM *(S$_2$,50) / SENT;
$S_1 = \{$ CM, SENT, $\text{inciso}_{est}\}$
$S_2 = \{$ prop, inciso_{est}, CM, VERBFIN, SENT$\}$

Proposición/Proposición

$\text{prop}_{est} \rightarrow$ SENT $\backslash$ *(S$_1$,50) VERBFIN *(S$_2$,50) / CM *(S$_1$,50) VERBFIN *(S$_2$,50) SENT;
$S_1 = \{$CM, SENT, prop $\}$
$S_2 = \{$prop, CM, VERBFIN, SENT$\}$

$\text{prop}_{est} \rightarrow \text{prop}_{est}$ CM $\backslash$ *(S$_1$,50) VERBFIN *(S$_2$,50) / SENT;
$S_1 = \{$CM, SENT, prop $\}$
$S_2 = \{$prop, CM, VERBFIN, SENT$\}$

$\text{prop}_{est} \rightarrow$ SENT $\backslash$ *(S$_1$,50) VERBFIN *(S$_2$,50) / CM *(S$_1$,50) VERBFIN *(S$_2$,50) CM prop;
$S_1 = \{$CM, SENT, prop $\}$
$S_2 = \{$prop, CM, VERBFIN, SENT$\}$

$\text{prop}_{est} \rightarrow \text{prop}_{est}$ CM $\backslash$ *(S$_1$,50) VERBFIN *(S$_2$,50) / CM prop;
$S_1 = \{$CM, SENT, prop $\}$
$S_2 = \{$prop, CM, VERBFIN, SENT$\}$

Bibliografía

[1] S. P. Abney. Parsing by chunks. En R. C. Berwick, S. P. Abney y C. Tenny, editores, "Principle-Based Parsing: Computation and Psycholinguistics", páginas 257–278. Kluwer Academic Publishers, Dordrecht, Netherlands (1991).

[2] Steven P. Abney, Robert E. Schapire y Yoram Singer. Boosting applied to tagging and PP attachment. En "Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)", páginas 38–45, College Park, MD (1999). ACL.

[3] James Allen. "Natural Language Understanding". The Benjamin/Cummings Publishing Company, California, US, 2 edición (1996).

[4] Murat Bayraktar, Bilge Say y Varol Akman. An analysis of english punctuation: The special case of comma. *International Journal of Corpus Linguistics* **3**(1), 33–57 (enero 01 1998).

[5] William J. Black y Argyrios Vasilakopoulos. Language-independent named entity classification by modified transformation-based learning and by decision tree induction. En "Proceedings of CoNLL-2002", páginas 159–162. Taipei, Taiwan (2002).

[6] John Breese, David Heckerman y Kadie Carl. Empirical analysis of predictive algorithms for collaborative filtering. En "Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI-98)", páginas 43–52, San Francisco,US (1998). Morgan Kaufmann Publishers.

[7] Eric Brill y Jun Wu. Classifier combination for improved lexical disambiguation. En "Proceedings of the 17th international conference on Computational linguistics", páginas 191–195. Association for Computational Linguistics (1998).

[8] Ted Briscoe. The syntax and semantics of punctuation and its use in interpretation. En "Proceedings of the Association for Computational Linguistics Workshop on Punctuation", páginas 1–7, Santa Cruz, US (junio 1996).

[9] Ted Briscoe y John Carroll. Developing and evaluating a probabilistic lr parser of part-of-speech and punctuation labels. En "Proceedings of the ACL/SIGPARSE 4th International Workshop on Parsing Technologies", páginas 48–58, Prague/Kqrlovy Vary, Czech Republic (1995).

[10] X. Carreras, I. Chao, L. Padró y M. Padró. Freeling: An open-source suite of language analyzers. En "Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)", Lisbon, Portugal (junio 2004).

[11] Xavier Carreras, Lluís Màrquez y Lluís Padró. A simple named entity extractor using adaboost. En Walter Daelemans y Miles Osborne, editores, "Proceedings of CoNLL-2003", páginas 152–155. Edmonton, Canada (2003).

[12] Serrana Caviglia, Javier Couto, Aiala Rosá y Dina Wonsever. Reconocimiento de indicadores de inicio de proposición en español. En "VIII Simposio Internacional de Comunicación Social", Cuba (enero 2003).

[13] Serrana Caviglia, Javier Couto, Aiala Rosá y Dina Wonsever. Un sistema para la segmentación en proposiciones de textos en español. En "VI Congreso de Lingüística General", Santiago de Compostela, Spain (mayo 2004).

[14] Serrana Caviglia, Mariela Malcuori y Mariela Grassi. Corpus informatizado: textos del español del uruguay. En "IV Congreso de Lingüística General", Cádiz, España (2000).

[15] Eugene Charniak. "Statistical Language Learning". MIT Press, Cambridge, US (1993).

[16] William F. Clocksin y Christopher S. Mellish. "Programming in Prolog". Springer-Verlag, Berlin, Deutschland, 4 edición (1994).

[17] Javier Couto. Los sistemas de exploración contextual de cara al usuario. Tesis de Maestría, Universidad de la República, Montevideo, Uruguay (2002).

[18] Gustavo Crispino. "Une plate-forme informatique de l'Exploration Contextuelle : modélisation, architecture et réalisation (ContextO)". Tesis Doctoral, Université Paris-Sorbonne, Paris, France (diciembre 2003).

[19] James Cussens. Part-of-speech tagging using Progol. En S. Džeroski y N. Lavrač, editores, "Proceedings of the 7th International Workshop on Inductive Logic Programming", tomo 1297, páginas 93–108. Springer-Verlag (1997).

[20] Jean-Pierre Desclés. "Systèmes d'exploration contextuelle. Co-texte et calcul du sens", páginas 215–232. Presses Universitaires de Caen, Caen, Francia (1996).

[21] Christine Doran. Punctuation in a lexicalized grammar. En "Proceedings of the Workshop TAG+5", Paris, France (mayo 2000).

[22] Diego Garat y Dina Wonsever. A constraint parser for contextual rules. En "SCCC '02: Proceedings of the XII International Conference of the Chilean Computer Science Society (SCCC'02)", páginas 234–242. IEEE Computer Society (noviembre 2002).

[23] Claudia García y Yamandú González. Reconocimiento de entidades con nombre. Proyecto de grado, Universidad de la República (abril 2005).

[24] Ralph Grishman. The nyu system for muc-6 or where's the syntax? En "MUC6 '95: Proceedings of the 6th conference on Message understanding", páginas 167–175. Association for Computational Linguistics (1995).

[25] John C. Henderson y Eric Brill. Bagging and boosting a treebank parser. En "Proceedings of the first conference on North American chapter of the Association for Computational Linguistics", páginas 34–41. Morgan Kaufmann Publishers Inc. (2000).

[26] Derrick Higgins y Jerrold M. Sadock. A machine learning approach to modeling scope preferences. *Computational Linguistics* **29**(1), 73–96 (2003).

[27] J. Hopcroft y J. Ullman. "Introduction to Automata Theory. Languages and Computation." Addison-Wesley, USA, 2 edición (1979).

[28] Peter Jackson y Isabelle Moulinier. "Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization." Jhon Benjamins Publishing Company, Amsterdam/Philadelphia (2002).

[29] Bernard Jones. Can punctuation help parsing? Informe técnico 29, Cambridge University, Cambridge, UK (1994).

[30] Bernard Jones. Exploring the role of punctuation in parsing natural text. En "Proceeding of the 15th International Conference on Computational Linguistics (COLING)", páginas 421–425. Association for Computational Linguistics (1994).

[31] Xin Li y Dan Roth. Exploring evidence for shallow parsing. En Walter Daelemans y Rémi Zajac, editores, "Proceedings of CoNLL-2001", páginas 38–44. Toulouse, France (2001).

[32] J. W Lloyd. "Foundation of Logic Programming". Springer-Verlag, 2 edición (1988).

[33] Christopher D. Manning y Hinrich Schütze. "Foundations of Statistical Natural Language Processing". The MIT Press, Cambridge, US (1999).

[34] Tom M. Mitchell. "Machine learning". McGraw Hill, New York, US (1997).

[35] Guillermo Moncecchi. Reglas contextuales y modelos de estado finito. Tesis de Maestría, Universidad de la República, Montevideo, Uruguay (diciembre 2004).

[36] Ghassan Mourad. La segmentation de textes par l'étude de la ponctuation. En "Document électronique dynamique: actes du colloque international sur le document électronique (CIDE'99)", páginas 155–171, Damasco,Siria (1999).

[37] L. Màrquez. Machine learning and natural language processing. Informe técnico LSI-00-45-R, Universitat Politècnica de Catalunya, Catalunya, España (2000).

[38] Lluís Màrquez, Lluís Padró y Horacio Rodríguez. A machine learning approach to pos tagging. *Machine Learning* **39**(1), 59–91 (2000).

[39] Tetsuji Nakagawa, Taku Kudo y Yuji Matsumoto. Revision learning and its application to part-of-speech tagging. En "Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)", páginas 497–504 (2002).

[40] Geoffrey Nunberg. "The Linguistics of Punctuation". Número 18 en CSLI Lecture Notes. Stanford University Press, Stanford, US (1990).

[41] Geoffrey Nunberg, Ted Briscoe y Rodney Huddleston. "The Cambridge Grammar of English", capítulo 20. Cambridge University Press, Cambridge, UK (2002).

[42] Richard O'Keefe. "The Craft of Prolog". The MIT Press, Cambridge, US (1990).

[43] G. Orphanos, D. Kalles, A. Papagelis y D. Christodoulakis. Decision trees and nlp: A case study in pos tagging. En "Proceedings of ACAI'99" (1999).

[44] Miles Osborne. Can punctuation help learning? En "Learning for Natural Language Processing", páginas 399–412 (1995).

[45] Manuel Penalver Castillo. Problemas de puntuación en el español peninsular. *Estudios Filológicos* **37**, 104–114 (2002).

[46] Juan José Prada. Marcadores del discurso en español - análisis y representación. Tesis de Maestría, Universidad de la República, Montevideo, Uruguay (octubre 2001).

[47] J. Ross Quinlan. "C4.5: Programs for Machine Learning". Morgan Kaufmann, San Mateo, US (1993).

[48] Real Academia Española. Avance del diccionario panhispánico de dudas. Disponible en: http://www.rae.es/. (Último acceso: diciembre de 2004).

[49] Real Academia Española. Ortografía de la lengua española (1999).

[50] Paul Resnick y Hal R. Varian. Recommender systems. *Communications of the ACM* **40**, 56–58 (March 1997).

[51] Stuart Russell y Peter Norvig. "Artificial Intelligence: A Modern Approach". Prentice-Hall, New Jersey, US, 2 edición (2003).

[52] Tjong Kim Sang y Erik F. Introduction to the CoNLL–2002 shared task: Language-independent named entity recognition. En "Proceedings of CoNLL–2002", páginas 155–158. Taipei, Taiwan (2002).

[53] Tjong Kim Sang, Erik F. y Fien De Meulder. Introduction to the CoNLL–2003 shared task: Language-independent named entity recognition. En Walter Daelemans y Miles Osborne, editores, "Proceedings of CoNLL–2003", páginas 142–147. Edmonton, Canada (2003).

[54] B. Say y V. Akman. Current approaches to punctuation in computational linguistics. *Linguistics. Computers and the Humanities* **30**(6), 457–469 (1997).

[55] R. Schapire. The boosting approach to machine learning: An overview (marzo 2001).

[56] Robert E. Schapire y Yoram Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning* **39**(2/3), 135–168 (2000).

[57] Helmut Schmid. Part-of-speech tagging with neural networks. En "Proceedings of the 15th conference on Computational Linguistics (COLING)", páginas 172–176. Association for Computational Linguistics (1994).

[58] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. En "International Conference on New Methods in Language Processing", Manchester, UK (1994).

[59] Stuart C. Shapiro, editor. "Encyclopedia of artificial intelligence". A Wiley-Interscience Publication, US (1992).

[60] SWI-Prolog. Sitio oficial en internet. http://www.swi-prolog.org/. (Último acceso: enero de 2005).

[61] Sebastian van Delden y Fernándo Gómez. Combining finite state automata and a greedy learning algorithm to determine the syntactic roles of commas. En "ICTAI '02: Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'02)", páginas 293–300, Washington, DC, USA (2002). IEEE Computer Society.

[62] Sebastian van Delden y Fernándo Gómez. A finite state comma tagger. *International Journal on Artificial Intelligence Tools* **13**(3), 449–468 (2004).

[63] Michael White. Presenting punctuation. En "Proceedings of the Fifth European Workshop on Natural Language Generation", páginas 107–125, Leiden, the Netherlands (mayo 1995). Faculty of Social and Behavioural Sciences, University of Leiden.

[64] Jan Wielemaker. An overview of the SWI-Prolog programming environment. En Fred Mesnard y Alexander Serebenik, editores, "Proceedings of the 13th International Workshop on Logic Programming Environments", páginas 1–16, Heverlee, Belgium (diciembre 2003). Katholieke Universiteit Leuven.

[65] Jan Wielemaker. "SWI-Prolog 5.4.1 Reference Manual". Department of Social Science Informatics (SWI), Universiteit Amsterdam (2004).

[66] Jan Wielemaker y Anjo Anjewierden. "Programming in XPCE/Prolog 5.4.1". Department of Social Science Informatics (SWI), Universiteit Amsterdam (2001).

[67] Dina Wonsever. "Repérage automatique des propositions par exploration contextuelle". Tesis Doctoral, Université Paris-Sorbonne, Paris, France (marzo 2004).

[68] Dina Wonsever y Jean-Luc Minel. Contextual rules for text analysis. *Lecture Notes in Computer Science* **2004**, 509–521 (2001).

[69] Dekai Wu, Grace Ngai, Marine Carpuat, Jeppe Larsen y Yongsheng Yang. Boosting for named entity recognition. En "Proceedings of CoNLL-2002", páginas 195–198. Taipei, Taiwan (2002).

[70] Xelda. Sitio oficial del Grupo Temis en internet. http://www.temis-group.com/index. (Último acceso: enero de 2005).