Criterios de etiquetación y desambiguación morfosintáctica de corpus en español

Montserrat Civit Torruella

Índice general

Índice de cuadros

Índice de figuras

Índice de siglas

AGTK	Annotation Graph ToolKit
BDS	Base de Datos Sintácticos del español (Universidad de Santiago)
BNC	British National Corpus
CLiC	Centre de Llenguatge i Computació (Universitat de Barcelona)
CORDE	Corpus Diacrónico del Español
CREA	Corpus de Referencia del Español Actual
EAGLES	Expert Advisory Group on Language Engineering Stantdards
ELRA	European language Archives Community
ISST	Italian Sintanctic Semantic Treebank
LDC	Linguistic Data Consortium
MACO	Morphological Analyzer Corpus Oriented
NERC	Network of European Reference Corpora
OLAC	Open Language Archives Community
OTA	Oxford text Archive
PDT	Prague Dependency Treebank
PTB	Penn TreeBank
PLN	Procesamiento del Lenguaje Natural
SGML	Standard Generalized Markup Language
TACAT	TAgged Corpus Text Analyzer
TALP	Tecnologies Aplicades al Llenguatge i la Parla (Universitat Politècnica de Catalunya)
TEI	Text Encoding Initiative
TUT	Turin University Treebank
XML	Extensive Markup Language

Introducción

El procesamiento de corpus presenta una problemática que muchas veces va más allá de los planteamientos y criterios que aparecen en las gramáticas. Suele ocurrir que en este ámbito se adopten criterios *ad hoc* o bien se adopten soluciones sin un análisis crítico de la base lingüística del problema. Nuestro objetivo es proponer una sistemática para el análisis de corpus que ha consistido en la definición de criterios de anotación basados en el análisis crítico de diferentes propuestas gramaticales y notaciones computacionales. Asimismo se han desarrollado nuevos recursos de análisis y se han mejorado los ya existentes. En definitiva, se aporta, con una evaluación crítica, una metodología razonada para el etiquetado de corpus del español, extensible a otras lenguas románicas.

Los objetivos fundamentales han sido la mejora de recursos existentes en el marco de trabajo CLiC-TALP; la definición de criterios de anotación basados en el análisis crítico de material existente y el desarrollo de recursos de ingeniería lingüística para el análisis y la etiquetación automática de corpus. La propuesta de criterios de anotación está fundamentada lingüísticamente y pensada para ser utilizada como base para otros trabajos. Hemos partido como marco de trabajo del entorno de tratamiento de corpus desarrollado por los grupos de investigación **CLiC** y **TALP**, respectivamente, del Departamento de Lingüística de la Universidad de Barcelona y del Departamento de Lenguajes y Sistemas Informáticos de la Universidad Politécnica de Catalunya. En este contexto, las aportaciones realizadas se concretan en los siguientes puntos:

1. Los sistemas de anotación morfosintáctica de corpus suelen utilizar un conjunto de etiquetas que suelen definirse con criterios particulares para una determinada aplicación o un determinado tipo de corpus; además, el tratamiento de las categorías cerradas (determinante, pronombre, conjunción, etc.) suele ser muy asistemático y poco fundamentado lingüísticamente. Nuestra aportación en este sentido ha consistido en la redefinición de un conjunto de etiquetas ya existente, de modo internamente consistente, basado en los criterios propuestos por el grupo Eagles y justificado desde un punto de vista lingüístico. Existía una versión anterior que se ha revisado y actualizado. Para ello, se han estudiado las diferentes clasificaciones de palabras realizadas desde la teoría gramatical y se ha justificado en cada caso la decisión tomada (cf. capítulo 2). Estas etiquetas se han implementado en el sistema de análisis morfológico **MACO**.

2. En el marco del sistema de tratamiento de corpus de los grupos **CLiC-TALP**, tras el análisis morfológico se realiza un proceso de desambiguación morfosintáctica au-

tomática que selecciona la etiqueta correcta para cada palabra (**RELAX**). Este desambiguador infiere restricciones de desambiguación a partir de corpus previamente anotados, y admite la introducción de restricciones elaboradas manualmente. Nuestra aportación a este módulo ha sido, por un lado, la elaboración de criterios para la validación manual del corpus de aprendizaje, el **Corpus CLiC-TALP** y, por otro, la introducción de restricciones para el desambiguador basadas en conocimiento lingüístico. La tasa de error del desambiguador es del 6 %, que se reduce al 4 % gracias a la incorporación de estas reglas (cf. capítulo 3).

3. La aproximación seguida en el procesamiento de corpus en nuestro marco de trabajo es modular, de manera que el resultado de cada proceso es la entrada para el proceso siguiente. Los distintos módulos de procesamiento no tienen un correlato directo con los niveles tradicionales del análisis lingüístico. La razón estriba en el hecho de que las diferentes herramientas de análisis están enfocadas a tratar determinados problemas específicos. El proceso que sigue a la desambiguación morfológica es un analizador sintáctico basado en *chunks* que trata algunas de las unidades léxicas que no se pueden resolver en el análisis anterior (perífrasis verbales, tiempos compuestos, etc.) e identifica los constituyentes básicos que preparan el texto para un procesamiento sintáctico en profundidad. Nuestra aportación en este punto ha sido el desarrollo de **GramEsp**, una gramática de *chunks* para el español, de amplia cobertura, aunque el análisis resultante es superficial (cf. capítulo 4).

   Aquí se cierra una primera fase de tratamiento de corpus basada en criterios estrictamente formales y de alta calidad en los resultados.

4. El último apartado de la tesis (cf. capítulo 5) sienta las bases para el procesamiento sintáctico en profundidad. Nuestra aportación ha consistido en la definición de los criterios de anotación sintáctica manual del corpus **Cast3LB**, para el desarrollo de un banco de árboles sintáctico que será la base para la elaboración de una gramática del español donde se traten constituyentes complejos y funciones sintácticas.

En la figura 1 se presenta en forma gráfica todo el proceso de tratamiento de corpus en el marco de trabajo **CLiC-TALP** y se destacan las aportaciones realizadas en este trabajo (aparecen a la derecha y están relacionadas con cada uno de los módulos de procesamiento de análisis).

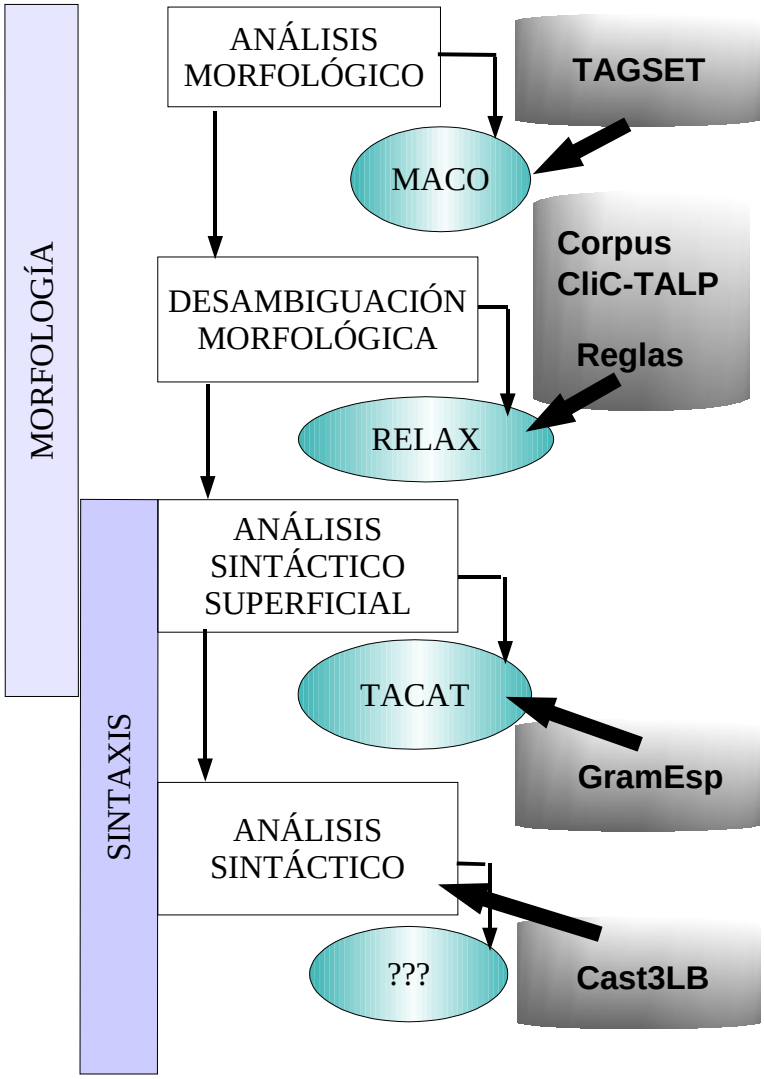

Figura 1: Cadena de procesos de análisis

Los ejemplos que citamos en este trabajo provienen de diversas fuentes. La mayoría provienen del corpus **CLiC-TALP** y están identificados al final con una letra y un número entre paréntesis que indica el archivo de origen. Otros, provienen del LexEsp o de los autores citados; en ambos casos se indica la fuente.

Los proyectos que han permitido el desarrollo de este trabajo han sido:

1. Sistema de Diálogo para Habla Espontánea en un Dominio Semántico Restringido (més conegut com a projecte Basurde): TIC98-0423-C06
2. X-TRACT Integración de recursos lingüísticos para la extracción de información de corpus textuales y diccionarios (2000-2003). Ministerio de Educación y Cultura. Programa sectorial de promoción general del conocimiento. (PB98-1226)
3. PETRA: Interficies Orales para Aplicaciones Avanzadas de Mensajería Unificada (TIC-2000-0335)
4. HERMES: Electronic Libraries with Multilingual Information Retrieval and Semantic Processing. (TIC2000-0335-C03-02)
5. X-TRACT-II: Plataforma para el desarrollo e integración de recursos de ingeniería lingüística (BFF2002-04 226-C03-03)
6. 3LB: Construcción de una base de datos de árboles sintáctico semánticos FIT-15050-2002-244 Financiado por el Ministerio de Ciencia y Tecnología (Programa PROFIT)

Capítulo 1

Marco general: la Lingüística de Corpus

En este capítulo se presenta una visión general de la Lingüística de Corpus, con especial atención a los diferentes sistemas de anotación y análisis morfosintáctico, puesto que este trabajo se centra precisamente en la construcción de una metodología coherente y sistemática de análisis de corpus en español a diferentes niveles: morfológico y sintáctico.

1.1. La Lingüística de Corpus

La Lingüística de Corpus consiste en el estudio de las lenguas a partir de ejemplos de uso de las mismas.

Los estudios lingüísticos basados en corpus no son nuevos, ni tienen que ver directamente con la aparición de la lingüística computacional. Boas, y con él, los estructuralistas, basaban sus estudios en el análisis de textos, tal como resume Harris (1993)[1]: *The approach began ... with a large collection of recorded utterances from some language, a corpus. The corpus was subjected to a clear, stepwise, bottom-up strategy of analysis.*

Esta metodología empezó a ser utilizada ya a finales del siglo XIX en el área de la adquisición del lenguaje, en pedagogía, en estudios de lingüística comparativa y también en sintaxis y semántica[2]. Sin embargo, en los años 50 del siglo XX, la lingüística de corpus sufrió un descrédito considerable, fundamentalmente debido a las críticas realizadas desde la gramática generativa, que no son más que la continuación del debate entre racionalistas y empiristas.

El racionalismo se basa en el desarrollo de una teoría de la mente y tiene como objetivo fundamental la plausibilidad cognitiva. Los datos que utiliza son fundamentalmente juicios introspectivos del lingüista. Por su parte, el empirismo fundamenta sus estudios en la observación de datos externos al lingüista, de corpus. En McEnery y Wilson (1996a) estos términos aparecen definidos del siguiente modo:

empirismo: *an approach to a subject (in our case linguistics) which is based upon the*

[1] Citado por McEnery y Wilson (1996a): p. 3.

[2] Para un breve repaso a estos estudios, véase McEnery y Wilson (1996a): pp. 2-4.

analysis of external data (such as text and corpora);
racionalismo: *an approach to a subject (in our case linguistics) which is based upon introspection rather than external data analysis.*

Este debate tiene también que ver con el objeto de estudio de la lingüística. Desde Saussure hasta Chomsky se ha primado el estudio de la *langue – competence* por encima del estudio de la *parole – performance*. Dado que los corpus son muestras de la actuación, éstos pierden todo interés para los lingüistas situados en el marco teórico racionalista.

A pesar del racionalismo dominante en el siglo XX, la lingüística de corpus no desapareció totalmente y siguió especialmente en fonética y en estudios de adquisición de lenguaje, donde la introspección no es posible.

Cabe preguntarse, pues, cuál es el interés, desde un punto de vista tanto teórico como aplicado, de la lingüística de corpus. En primer lugar hay que señalar que el uso de datos externos al hablante, resultado de la producción lingüística, tiene una ventaja importante sobre la introspección: estos datos son directamente observables por aquél que lo desee. En segundo lugar, los corpus son una fuente de información cuantitativa importante. La frecuencia de aparición de los elementos lingüísticos no puede recuperarse por introspección y, si una estructura o una palabra no aparece en un corpus representativo de una lengua, esta información también es significativa, ya que proporciona información sobre esta estructura. Finalmente, los corpus, posibilitan la verificación objetiva de resultados, especialmente cuando se trata de desarrollar herramientas para el procesamiento del lenguaje natural, pero también cuando se trata de aportar pruebas para una teoría[3].

Sampson (2001) dedica varios de los capítulos de su libro a insistir en el hecho de que para estudiar la lengua lo único necesario es tener evidencias, datos empíricos: *the evidence on which a linguistic theory is based [...] consists of people's utterances.* El capítulo 8 de este libro se titula *Objective evidence is all we need*; en él el autor considera que *what is crucial is that any hypothesis which is challenged should be* tested *against interpersonally observable, objective data.* Este autor representa la actitud más radicalmente opuesta al racionalismo.

Desde hace algunos años, sin embargo, *Corpus* es sinónimo de *Corpus en soporte electrónico* y *Lingüística de corpus* lo es de *Lingüística de corpus en soporte electrónico*[4]. La aparición y el uso del ordenador han permitido, por una parte, buscar, recuperar, ordenar y hacer cálculos sobre los datos de forma rápida y eficaz y, por otra, hacer posible el manejo de un volumen de datos inimaginable hace tan sólo unas décadas:

> *It is truer to say that the interest in the computer for the corpus linguist comes from the ability of the computer to carry out the processes of searching for, retrieving, sorting and calculating linguistic data, whether that be textual (most*

[3]La objetividad aparece como un factor importante, mientras que la subjetividad o introspección puede causar problemas, como por ejemplo, en la dificultad que hay para valorar como gramaticales o agramaticales ciertas secuencias, lo que se traduce en una gran proliferación de '?' precediendo a las oraciones 'dudosas'. En estos casos, el juicio depende exclusivamente del lingüista.

[4]Traducción del inglés *Computer Corpus Linguistics.* Término que aparece, entre otros, en Leech (1997b), McEnery y Wilson (1996a) y Ooi (1998).

common) or digitized speech (increasingly common).[5]

La importancia del ordenador queda reflejada en las palabras de McEnery y Wilson (1996a): *Whatever philosophical advantages we may eventually see in a corpus, it is the computer which allows us to exploit corpora on a large scale with speed and accuracy, and we must never forget that. Technology has allowed a pseudo-procedure to become a valuable linguistic methodology* (pp. 18–19).

La lingüística de corpus tal como se entiende en la actualidad, está relacionada no sólo con el análisis y la interpretación de la lengua sino también con las técnicas computacionales y la metodología para el análisis de estos textos (Ooi, 1998).

Desde el punto de vista de la lingüística computacional, la aparición de corpus en soporte informático ha implicado la creación de las herramientas necesarias para abordar el tratamiento de grandes cantidades de textos. Las implicaciones que ello ha tenido sobre la lingüística computacional han sido un cambio radical de enfoque, ya que si antes se trabajaba con pequeños sistemas (*toy-systems*) que demostraban que un determinado proceso podía resolverse computacionalmente, ahora se trabaja con grandes cantidades de datos, de modo que se ha sacrificado la profundidad en el tratamiento de los análisis en favor de la cantidad de datos procesados. Los primeros sistemas de análisis sintáctico, por ejemplo, llevaban a cabo análisis profundos de las oraciones correspondientes a dominios restringidos o sublenguajes. Un análisis superficial era totalmente irrelevante para este tipo de sistemas. Ahora, sin embargo, la inmensa mayoría de analizadores realizan análisis parciales o superficiales de grandes cantidades de textos. Actualmente, el interés se centra en la creación de corpus anotados a distintos niveles y en el desarrollo de herramientas capaces de tratar de modo automático esos corpus, y de llevar a cabo de modo automático esa anotación. Esto ha implicado la recuperación de los métodos estadísticos en esta disciplina.

Según Leech (1992)[6] la *lingüística de corpus en soporte electrónico* focaliza su interés en (i) la actuación lingüística más que en la competencia; (ii) la descripción lingüística más que los universales; (iii) modelos de lenguaje no sólo cualitativos sino también cuantitativos; y (iv) un punto de vista de la investigación lingüística más empirista que racionalista.

1.1.1. ¿Qué es un corpus?

In principle, any collection of more than one text can be called a corpus. But... (McEnery y Wilson, 1996a).

Aunque históricamente la definición de corpus como colección de textos podría considerarse aceptable, ya hemos mencionado el hecho de que en los últimos años dicha definición se ha visto modificada a causa fundamentalmente de la aparición de las nuevas tecnologías. Por tanto, podemos definir un *corpus* no como una simple colección de textos, sino, siguiendo a Leech (1997b), como:

[5] McEnery y Wilson (1996a): p. 17.

[6] G. Leech (1992) "Corpora and Theories of linguistic performance", en Svartvik, J. (ed.) 1992 *Directions in Corpus Linguistics: Proceedings of Nobel Symposium*, No 82 Stockholm, 4-8 August, 1991, Berlin: Mouton de Gruyter, citado por Ooi (1998).

> *a body of language material which exists in electronic form, and which may be processed by a computer for various purposes such as linguistic research and language engineering.*

Además, podemos completar esta definición con la que proporcionan McEnery y Wilson (1996a):

> *a finite collection of machine-readable text, sampled to be maximally representative of a language or variety.*

y Sampson (2001):

> *a siseable sample of real-life usage in English or another language under study, compiled and used as a source for generating hypotheses about the nature of the language. [...] Nowadays [...] language corpora are machine-readable.*

Los tres autores mencionan en sus definiciones el hecho, anteriormente destacado, de que los corpus están en *formato electrónico*. Esto presenta dos grandes ventajas: por una parte, los textos pueden leerse y manipularse (esto es, analizarse, procesarse, etc.) con gran rapidez y pueden enriquecerse con nueva información. Y, por otra, se facilita el intercambio de textos.

También en los tres casos se menciona el hecho de la *representatividad*. En la constitución de un corpus de referencia debe realizarse una selección que represente el subconjunto que se quiere analizar; de otro modo sólo se obtendría un modelo sesgado. La representatividad es, en realidad, una respuesta a las críticas que Chomsky hizo hacia la Lingüística de Corpus y que aparecen mencionadas tanto en (McEnery y Wilson, 1996a) como en (Sampson, 2001). La crítica se basaba en el hecho de que un conjunto finito de oraciones de una lengua no puede tomarse por la lengua, que es infinita. La respuesta a esta crítica ha sido la de intentar recoger, en la constitución de un corpus, una amplia muestra de la variación y la diversidad lingüística. Para ello el primer paso es definir de la forma más clara posible los límites de población que se desea estudiar para luego definir los procesos de toma de muestras. En (McEnery y Wilson (1996a): p.79) se presentan dos métodos utilizados por los creadores de dos corpus distintos. Los textos del corpus Lancaster-Oslo/Bergen (LOB corpus) fueron seleccionados a partir de la *British National Bibliography* y de la *Willings' Press Guide*. Por su parte, el Brown Corpus se constituyó a partir de textos publicados en 1961. En el caso de los corpus orales suele llevarse a cabo un muestreo demográfico, es decir, una selección de los informantes basada en su edad, sexo, región, clase social, etc. Aquí la estadística aporta sus bases metodológicas para el muestreo y la selección de datos. De este modo fue recogida la parte oral del British National Corpus; pero se complementó con otros datos de carácter oral como entrevistas y procesos judiciales. Debe señalarse, sin embargo, que la representatividad está en función del dominio del corpus: no es lo mismo un corpus que quiera ser el reflejo de una lengua (los llamados *corpus de referencia*), donde la variedad deberá ser mayor, que un corpus que se constituya con el objetivo de reflejar un dominio específico de la lengua: la representatividad aquí queda restringida al dominio concreto que se trate.

Tanto McEnery y Wilson como Sampson mencionan explícitamente el hecho de que los corpus tienen un *tamaño finito*. Ésta es una característica, en la práctica, inherente a los corpus: por lo general, el tamaño que tendrá se establece al principio y, cuando se llega a él, el corpus está constituido. Por lo tanto, además de seleccionar la población de textos, hay que determinar el tamaño de los ejemplos, tanto en lo referente a la longitud de cada uno como la cantidad óptima de texto que debe incluirse en el corpus. Si no se hiciera así, los objetivos marcados en la fase inicial de constitución del corpus podrían alterarse. Si los corpus se aumentan, es necesario tener en cuenta los mismos criterios iniciales de creación. Sin embargo, hay algunas excepciones a la *finitud* de los corpus, y ambas tienen que ver con los objetivos del corpus: son los llamados *Monitor Corpora*, entre los que destaca sobre todo el *Cobuild Bank of English* que se está construyendo en la Universidad de Birmingham por el equipo de John Sinclair. Otro ejemplo de *Monitor corpora* lo ofrece el *Russian Monitor Corpora* (Yablonsky, 2000).

Finalmente, Leech y Sampson hablan de los objetivos para los que se crean los corpus. Las aplicaciones de los corpus son muy numerosas y se dividen en dos grandes parcelas, la computacional y la lingüística. Dedicamos la sección 1.1.1.2 a comentarlas.

Además de todas estas características, hay que mencionar el hecho de que idealmente los corpus deberían estar disponibles para los investigadores. Las ventajas que esto supone son, en primer lugar, que el corpus se convierte en un criterio con que evaluar o comparar diferentes herramientas y/o metodologías de estudio; y, en segundo lugar, que las variaciones entre diferentes estudios realizados con el mismo corpus deben atribuirse a la metodología y no a los datos.

A modo de resumen, Leech (1997b) menciona que el valor de un corpus como herramienta de investigación no reside sólo en su tamaño, sino también en su diversidad (que se corresponde con la representatividad de McEnery y Wilson (1996a)), en el cuidado o atención con el que ha sido recogido (atención prestada, por ejemplo, a la corrección ortográfica si el corpus es escrito) y, especialmente, en la anotación, que es, para el autor, el valor añadido del corpus:

> *Corpus annotation is widely accepted as a crucial contribution to the benefit a corpus brings, since it enriches the corpus as a source of linguistic information for future research and development.*

Por su parte, J. Sinclair (EAGLES, 1996a) sostiene que:
(i) *The corpus should be as large as could possibly be envisaged with the technology of the time.*
(ii) *It should include samples from a broad range of material in order to attain some sort of representativeness.*
(iii) *There should be an intermediate classification into genres between the corpus in total and the individual samples.*
(iv) *The samples should be of an even size.*
(v) *The corpus as a whole should have a declared provenance.*

Además, el mismo autor señala que las características que se suponen a cualquier corpus son: cantidad, calidad, simplicidad y documentación. *Cantidad* tiene como valor por defecto *grande*. Se supone que los corpus contienen grandes cantidades de texto. En la práctica,

el tamaño puede depender de la disponibilidad de textos y de la posibilidad de hacerlos públicos. *Calidad* se define por defecto como *auténtica*, esto es, los textos deben recogerse a partir de material genuino, original, tal como las personas lo producimos. Según el autor, si en algo interviene el lingüista el corpus debe considerarse de otro modo dado que las modificaciones introducidas alteran la fuente. *Simplicidad* se define como *texto plano*, es decir, una cadena de caracteres ASCII. Si se introducen en el corpus marcas de cualquier clase, éstas deben poder separarse del texto fácilmente. Este mismo criterio de separabilidad se aplica también a cualquier tipo de anotación lingüística que se pueda añadir a la fuente. Por último, *documentación* tiene como valor por defecto *documentado*, lo que significa que todos los detalles sobre los elementos del corpus deben mantenerse, aunque separados del propio texto. El modelo que el autor propone para la documentación de los corpus es la DTD (*Document Type Definition*).

En este trabajo nos centraremos fundamentalmente en los corpus anotados con información lingüística.

1.1.1.1. Tipología de corpus

La clasificación de los corpus puede llevarse a cabo atendiendo a la realización de la lengua que recogen, a sus objetivos, a su contenido o al número de lenguas que incluyen. En todos los casos, los corpus pueden estar anotados o no (para este aspecto, cf. sección 1.3).

1. **Orales –vs– textuales.** Según aparece en el *Informe sobre recursos lingüísticos para el español* (Instituto-Cervantes, 1996) los corpus orales pueden entenderse desde diferentes puntos de vista. Bien se trata de recopilaciones de realizaciones fonéticas producidas en condiciones de grabación muy controladas y cuyo objetivo es el estudio de las manifestaciones fonéticas de una determinada lengua, o el entrenamiento de sistemas de reconocimiento y síntesis del habla; bien se trata de colecciones de grabaciones orales de procedencia muy diversa y obtenidas en situaciones reales de uso que se utilizarán para llevar a cabo estudios sobre la lengua en su manifestación oral.

   El primer tipo de corpus orales presentan la señal acústica digitalizada grabada en un entorno controlado y, en ocasiones, se acompaña esta señal de la transcripción ortográfica y/o fonética.

   Los segundos suelen consistir en transcripciones ortográficas más o menos enriquecidas y sus objetivos van desde el análisis léxico a estudios sobre el desarrollo del lenguaje, pasando por análisis conversacional, del discurso o de variación geográfica.

   Por otra parte, los corpus textuales son colecciones de textos recogidas con fines muy diversos: constituir corpus de referencia de una lengua, bases de datos para el estudio diacrónico, para el sincrónico, para la adquisición de segundas lenguas, etc.[7]

2. **Anotados –vs– no anotados.** Los corpus no anotados son los que recogen texto plano, mientras que los corpus anotados son aquellos a los que se ha añadido algún

[7] En el informe del Instituto-Cervantes (1996) aparece una relación de los corpus orales y escritos del español. Si bien esta referencia es algo antigua, es la más completa que se ha hecho.

tipo de información, especialmente lingüística (cf. sección 1.2 para una distinción entre codificación de los corpus y anotación).

3. **Fines generales –vs– fines específicos.** Según aparece en el informe del Instituto-Cervantes (1996), los corpus pueden clasificarse, atendiendo a sus objetivos, como corpus con fines generales o corpus con fines específicos. Los primeros recogen muestras representativas de la lengua, ya orales, ya escritas; los segundos se han constituido con el fin de estudiar aspectos concretos de la lengua, ya sean gramaticales, léxicos, pragmáticos, prosódicos, etc. Sin embargo, es de notar que la reutilización de los corpus es algo frecuente, de modo que un corpus creado con un fin concreto puede reutilizarse con otra finalidad si sus características y diseño lo permiten.

4. **Contenido general –vs– contenido específico.** En cuanto al contenido, los corpus pueden ser un reflejo de toda la variedad de tipología textual y/o comunicativa de una lengua, o, al contrario, reflejar sólo un tipo concreto de textos o una situación de habla determinada.

5. **Monolingües –vs– multilingües.** Por último, si bien la mayoría de corpus existentes son monolingües, cada vez es mayor el número de corpus bilingües o, incluso, multilingües. Los corpus que contienen textos de más de una lengua, pueden subclasificarse en dos grupos: el primero lo formarían aquellos corpus que simplemente recogen textos en más de una lengua, sin que haya ninguna relación entre dichos textos. El segundo lo constituyen los corpus en que los textos, en dos o más lenguas, guardan entre sí alguna relación. Según sea esta relación, se tratará de *corpus comparables* o de *corpus paparelos*. Por *corpus comparables* se entienden aquellos que *contienen textos en distintos idiomas, que, sin ser traducciones, comparten similar origen, temática, extensión y número: partes meteorológicos, ofertas laborales, artículos periodísticos, etc. Es decir, los textos no se reunen de manera arbitraria, sino que se escogen de acuerdo a unos criterios de selección comunes* (Matínez, 1999). Los *corpus paralelos* son aquellos que *contienen una misma colección de textos en más de una lengua, es decir, cuando a las versiones originales les acompañan sus traducciones*[8]. El proceso que más valor añade a los corpus bilingües o multilingües es sin duda la alineación de los textos. Por *alineación* se entiende el hecho de explicitar las relaciones de equivalencia entre las distintas unidades de una y otra lengua. El nivel de alineación puede variar entre alineación de segmentos (textos o párrafos); oraciones; sintagmas; expresiones o palabras.

1.1.1.2. Aplicaciones de los distintos tipos de corpus

Las utilidades de los corpus abarcan todos los ámbitos del estudio de la lengua, tanto desde el punto de vista teórico como también aplicado. Sobre este último, cabe señalar que la importancia de los corpus es creciente, especialmente para aquellos sistemas de ingeniería lingüística llamados *sistemas cognitivamente implausibles*[9]. Estos sistemas tienen

[8] Abaitua (2000).
[9] McEnery y Wilson (1996a): pp. 133–134.

como objetivo la construcción de modelos de análisis del lenguaje sin tener en cuenta si el sistema diseñado es similar al humano o no. Se oponen a los llamados *cognitivamente plausibles* cuyo objetivo es la obtención de modelos cognitivos pensados para ser relevantes sobre cómo los humanos realizan ciertas tareas adscritas a la 'inteligencia'. Mientras estos últimos utilizan grandes bases de conocimiento y dejan un poco de lado los corpus, los primeros los necesitan de un modo casi esencial, puesto que son las únicas fuentes masivas de datos de que se puede disponer:

> *Corpora may be used to aid in the construction of systems which are interested in claims of cognitive plausibility - but, where cognitive plausibility is sacrified to brute force mathematical modelling, corpora are the* sine qua non *of such approach. Corpora provide the necessary raw data for approaches to language engineering based upon abstract numerical modelling*[10].

Las aplicaciones de los corpus abarcan casi todos los ámbitos de la lingüística y de la ingeniería lingüística. En el ámbito del **habla** los corpus proporcionan información sobre las variables que la condicionan, como el sexo o la edad, además de mostrar la producción lingüística en su forma más genuina. Desde otro punto de vista, se utilizan para la construcción de modelos de síntesis y reconocimiento de voz. La **lexicografía** se sirve de los corpus como fuente para la obtención de datos y ejemplos con los que construir diccionarios y lexicones computacionales, como en el caso del corpus de la RAE, del del Institut d'Estudis Catalans o de los diccionarios Cobuild. Para los **estudios gramaticales**, además de la información lingüística sobre el uso de la lengua, los corpus, especialmente si están anotados, permiten inferir y probar analizadores y desambiguadores morfológicos (cf. (Padró, 1998), (Màrquez y Padró, 1997) y (Civit, Martí, y Padró, 2003) para el español, por ejemplo) y analizadores sintácticos (véanse a este respecto los siguientes trabajos: Engelson y Dagan (1996), Carroll, Briscoe, y Sanfilippo (1998), Carroll, Minnen, y Briscoe (1999), Carroll, Minnen, y Briscoe (2003), Böhmova y Hajicova (1999), Brants y Plaehn (2000)), componentes necesarios para los sistemas de recuperación y extracción de información, sistemas de pregunta–respuesta o de detección de entidades con nombre (Arévalo et al., 2002), (Carreras, Màrquez, y Padró, 2003). En el ámbito de la **semántica**, la principal contribución de los corpus ha sido y es la de proporcionar una aproximación objetiva al estudio del significado. Con frecuencia, el significado de las palabras se describe por referencia a las intuiciones del lingüista. En cambio, en los textos podemos observar la influencia que el contexto (morfológico, sintáctico o prosódico) tiene en la determinación de los significados[11]. En el marco de la desambiguación de sentidos (WSD), disponer de corpus anotados se hace imprescindible si los métodos que se utilizan son supervisados. De forma breve, los métodos supervisados son aquellos que utilizan corpus previamente anotados a partir de los cuales se infiere el conocimiento necesario y frente a los cuales se evalúan; se oponen a los métodos no supervisados, que utilizan fuentes de conocimiento externas (diccionarios, bases de datos, taxonomías, etc.) para anotar el texto. La **pragmática** y la **sociolingüística** pueden, a partir de corpus convenientemente anotados (Payrató (1996), Payrató y

[10] McEnery y Wilson (1996a): p. 134.

[11] En este sentido destacan los trabajos de J. Véronis (Véronis, 2000) y (Véronis, 2001b).

Alturo (2002), Llisterri (1997)), realizar estudios sobre el desarrollo de las interacciones comunicativas, el funcionamiento de los turnos de palabra, etc. o estudiar fenómenos como la interferencia lingüística.

Otros ámbitos en los que el uso de corpus ha significado una aportación relevante son según Abaitua (2000) la enseñanza de segundas lenguas y la traducción. Según este autor, disponer de corpus bilingües o multilingües permite el estudio de las influencias de la lengua materna sobre la segunda lengua; por otra parte, si esos corpus están alineados pueden utilizarse para la creación de memorias de traducción.

En todos estos casos, los corpus pueden utilizarse para comprobar (y/o rechazar) hipótesis lingüísticas (Sampson, 2001); para describir la lengua; para evaluar herramientas y algoritmos y para procesar automáticamente el lenguaje natural.

1.2. Codificación –vs– Anotación de corpus

El corpus puede consistir en texto plano (el texto en sí, sin ningún tipo de información añadida: corpus no anotados) o bien puede presentar información complementaria (corpus anotado).

Por lo general podemos hablar de dos tipos de enriquecimiento de los textos: (i) el marcaje del texto o codificación y (ii) la anotación o enriquecimiento del texto con información lingüística de diversa índole.

El texto en sí puede marcarse, por ejemplo, con SGML (*Standard Generalized Markup Language*) o con XML (*Extensive Markup Language*), que es lo que se propone desde TEI [12], NERC [13] e EAGLES [14]. XML es una versión reducida de SGML que facilita el marcaje de los propios documentos. La ventaja de este tipo de marcaje es que se establece una clara distinción entre los datos y la anotación, con lo que el texto original puede recuperarse con facilidad. Por ejemplo, en la propuesta TEI, todo texto o documento consta de dos partes: la cabecera y el texto propiamente dicho. La cabecera contiene información sobre autor, título y fecha; la edición empleada para la creación del texto en formato electrónico y la información sobre la práctica de codificación adoptada. En el texto se marcan el inicio y final de párrafo, los elementos destacados en negrita o cursiva, por ejemplo. La gran ventaja de este tipo de marcación es que facilita enormemente el intercambio de textos en soporte informático así como la detección de aquellos elementos textuales que aportan más información que la estrictamente lingüística.

A continuación puede observarse un ejemplo de corpus anotado a nivel morfológico y sintáctico en formato XML. Se trata del análisis de la oración *Una estrella que ayer recobró en parte un brillo semiolvidado*. La etiqueta *STX TYPE* corresponde a los nodos del árbol etiquetados sintácticamente; *WRD FORM* señala que lo que sigue es la palabra de la oración; y *POS* indica que lo siguiente es la categoría morfológica de la palabra.

```
<STX TYPE="S">
    <STX TYPE="sn">
```

[12] Text Encoding Initiative (http://www.tei-c.org).

[13] Network of European Reference Corpora.

[14] Expert Advisory Group on Language Engineering Standards (EAGLES, 1997) y (Ces, 2000).

```
        <STX TYPE="espec-fs">
            <WRD FORM="Una" POS="di0fs0" />
        </STX>
        <STX TYPE="grup-nom-fs">
            <WRD FORM="estrella" POS="ncfs000" />
        </STX>
    </STX>
    <STX TYPE="relatiu">
        <WRD FORM="que" POS="pr0cn000" />
    </STX>
    <STX TYPE="sadv">
        <WRD FORM="ayer" POS="rg" />
    </STX>
    <STX TYPE="grup-verb">
        <WRD FORM="recobró" POS="vmis3s0" />
    </STX>
    <STX TYPE="grup-sp">
        <STX TYPE="prep">
            <WRD FORM="en" POS="sps00" />
        </STX>
        <STX TYPE="sn">
            <STX TYPE="grup-nom-fs">
                <WRD FORM="parte" POS="ncfs000" />
            </STX>
        </STX>
    </STX>
    <STX TYPE="sn">
        <STX TYPE="espec-ms">
            <WRD FORM="un" POS="di0ms0" />
        </STX>
        <STX TYPE="grup-nom-ms">
            <WRD FORM="brillo" POS="ncms000" />
            <STX TYPE="s-a-ms">
                <WRD FORM="semiolvidado" POS="aq0ms0" />
            </STX>
        </STX>
    </STX>
    <WRD FORM="." POS="Fp" />
</STX>
```

Si el corpus es oral, se pueden marcar también elementos como las pausas, los elementos kinésicos, las superposiciones entre los hablantes, información contextual, etc.

En cuanto a la anotación o enriquecimiento del texto en sí, ésta significa añadir, a la base textual, una descripción de las unidades, lo que implica también una interpretación

de los datos. Dedicamos la próxima sección a comentar este proceso.

1.3. Anotación de corpus

La anotación lingüística puede definirse del siguiente modo:

> *If corpora is said to be* **unannotated** *it appears in its existing raw state of plain text, whereas* **annotated** *corpora has been enhanced with various types of linguistic information. Unsurprisingly, the utility of the corpus is increased when it has been annotated, making it no longer a body of text where linguistic information is implicitly present, but one which may be considered a repository of linguistic information. The implicit information has been made explicit through the process of concrete annotation.* (sic.) (McEnery y Wilson, 1996a).

Según Leech (1997b) la anotación de corpus

> *can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data*

La tarea de anotación consiste pues, por un lado, en la explicitación de la información contenida en el texto, hecho que implica una interpretación del mismo. La labor de anotación es interpretativa en el sentido de que es producto de la 'comprensión' humana del texto y de que no hay una fórmula objetiva y mecánica para decidir qué etiqueta o etiquetas deben aplicarse a un determinado fenómeno lingüístico (Leech, 1997b). Por otra parte, también podemos considerar 'interpretativo' el hecho de que hay que decidir qué fenómenos se anotan, cuáles no y hasta qué nivel de detalle se quiere llegar en la explicitación de la información lingüística contenida en el texto. En algunos casos, resulta difícil separar representación e interpretación de los datos: por ejemplo, la transcripción de una grabación implica un componente interpretativo importante del que es difícil sustraerse (cf. Benveniste (1998)).

El paso previo a la anotación de corpus es el establecimiento de los criterios de anotación que se desea incorporar al corpus.

G. Leech (en su "Corpus annotation schemes", publicado en la revista *Literary and Linguistics Computing* 8(4):275–81) propone 7 máximas para la anotación de corpus:

> (i) *It should be possible to remove the annotation from an annotated corpus in order to revert to the raw corpus.*
> (ii) *It should be possible to extract the annotations by themselves from the text.*
> (iii) *The annotation scheme should be based on guidelines which are available to the end user.*
> (iv) *It should be made clear how and by whom the annotation was carried out.*
> (v) *The end user should be made aware that the corpus annotation is not infallible, but simply a potentially useful tool.*
> (vi) *Annotation schemes should be based as far as possible on widely agreed and theory-neutral principles.*

(vii) *No annotation scheme has the a priori right to be considered as a standard.* (Leech citado por McEnery y Wilson (1996a).)

Comentamos a continuación cada una de estas máximas

(i) *Debería ser posible eliminar la anotación del corpus anotado con el fin de poder obtener el texto plano.* De este modo se facilita la reutilización del corpus con otros fines porque puede reconstruirse con facilidad y se posibilitan, igualmente, ulteriores anotaciones.

(ii) *Debería ser posible extraer las propias anotaciones del texto.* Se trata de poder trabajar exclusivamente con la anotación. Por ejemplo, en el caso de que se trabaje con anotación morfológica, debería ser posible obtener sólo las etiquetas, para, por ejemplo, realizar estudios estadísticos sobre frecuencias de categorías.

(iii) *El esquema de anotación debería basarse en una guía que debería estar disponible para el usuario final.* De este modo, el usuario puede conocer los criterios de anotación seguidos, el significado de las etiquetas o marcas utilizadas, etc. Esto es especialmente relevante cuando se dan casos de ambigüedad y los anotadores han optado por una solución de entre las posibles y no por otra.

(iv) *Debería indicarse claramente cómo y por quién se ha llevado a cabo la anotación.* Debe señalarse si el proceso ha sido automático, semiautomático o totalmente manual; qué herramientas se han utilizado, etc.

(v) *El usuario final debería ser plenamente consciente del hecho de que la anotación de corpus no es algo infalible, sino, simplemente, una herramienta potencialmente útil.* Y ello es así porque la anotación de corpus es también, y por definición, un acto de interpretación, tanto de la estructura del texto como de su contenido.

(vi) *Los esquemas de anotación deberían basarse, en la medida de lo posible, en principios aceptados de modo general y teóricamente neutros.* Si se sigue este principio, se facilita enormemente la reutilización de los corpus, incluso para fines distintos de aquéllos para los que fueron creados.

(vii) *Ningún esquema de anotación tiene, a priori, el derecho de considerarse como un estándar.* No se trata aquí de estándares de codificación de la información (en los que es más o menos fácil llegar a consensos), sino de la estandarización de los contenidos de la anotación, ya que algunas aplicaciones pueden necesitar una mayor granularidad en el contenido que otras. A este respecto, Leech (1997b) destaca: *the need is to encourage convergent practice without imposing a straitjacket of uniformity which would inhibit flexibility and productive innovation.*

La anotación de los textos es, pues, un valor añadido importante, ya que es lo que permite la extracción de recursos, la evaluación de las herramientas de PLN, el aprendizaje automático de modelos lingüísticos, etc. Y a pesar de ello, hay pocos corpus anotados libremente disponibles para la comunidad científica. La razón hay que buscarla en el alto coste

que ello implica, especialmente si se pretende que la anotación satisfaga ciertos criterios de calidad, como la coherencia, la consistencia interna y la buena documentación, que son los que otorgan valor a la anotación.

1.3.1. Tipos de anotación

Seguimos para esta sección la presentación de McEnery y Wilson (1996a).

1. **Ortografía**.

   Si bien aparentemente la ortografía no debería representar ningún problema, sí que lo es para lenguas con caracteres idiosincrásicos que no están contemplados en los teclados de ordenador convencionales. Para ello se ha desarrollado el estándar UNICODE que permite la representación de los caracteres románicos así como de los demás *tal como son*. Por otra parte, la trasliteración ortográfica sí es un problema cuando se trata de corpus orales, puesto que en ellos no hay puntuación explícita, y cualquier intento de reconstruir frases y sintagmas es un acto de interpretación por parte del transcriptor. Además, en este tipo de corpus, hay otras dificultades que son en realidad la característica distintiva de los corpus orales: la espontaneidad. Es muy fácil encontrar repeticiones, muletillas, cortes del acto comunicativo, superposiciones, elementos ininteligibles para el transcriptor. En la mayoría de las ocasiones, estos elementos se reproducen en la trasliteración. También deben mencionarse los sonidos fáticos, los movimientos que aportan significado lingüístico, como asentir o negar con la cabeza. Por último, el ruido, las risas, los aplausos en conferencias, etc.

   Un ejemplo de trasliteración de corpus oral es el siguiente, tomado de *http://www.lllf.uam.es/corpus/corpus_lee.html#A* y que pertenece al *Corpus Oral de referencia del Español Contemporáneo* elaborado entre 1991 y 1992 en la Universidad Autónoma de Madrid en cooperación con IBM España. En este fragmento puede observarse cómo se tratan fenómenos típicos de la lengua oral, como la superposición de turnos de palabra, que se marcan con el código *&ltsimultáneo&*, o las pausas, marcadas con el código *&ltsilencio*.

```
&ltADEP008A.WPT>
     &ltfuente=televisión>
     &lt2-91>
     &ltlocalización=Madrid>
     &lttérminos=partido, puntos, baloncesto, defensor, tiro libre>
     &ltH1=Varón, comentarista deportivo>
     &ltH2=Varón, comentarista deportivo>
     &ltH3=Varón, locutor en la cancha>

     &lttexto>
...
&ltH1> Sibilio... Ahora anotó, los tres puntos de esta segunda
       parte los ha conseguido Sibilio.
```

```
       &ltsilencio> 51 - 36... &ltextranjero&gtTrumbo</extranjero>
       Coneti. Balón para Solozábal,
       lanzamiento triple y lo convierte.
&ltH2> No se puede... dejar tirar a un hombre como Nacho Solozábal
       con esa comodidad y a esa distancia, y más perdiendo de la...
       de la diferencia que está perdiendo el Taugrés; hay que salir a
       morder cada posesión y ataque del equipo contrario.
&ltH1> Siete triples lleva ya... el Barcelona. Sibilio... Y en la
       línea de fondo, con muchas dificultades,
       Piculín taponó y reboteó... Solozábal... Triple de Lisar...
       Ahora no.
&ltH2> Ese ha sido el más fácil que ha tirado... ¿eh?
       Los otros cuatro que había metido, cuatro si no me equivoco,
&ltH1> &ltsimultáneo&gtSí, sí.
&ltH2> ...había</simultáneo> sido mucho más compleja la posición,
       con un defensor delante...
       Ahora que estaba con el hombre más próximo que le... marcase
       a cuatro metros, lo ha fallado.
&ltH1> Sibilio... Cinco puntos ya en esta segunda parte.
       &ltsilencio> Epi asiste a &ltextranjero&gtTrumbo</extranjero>
       ... vamos a ver si es capaz de subir la bola. No, porque hubo
       falta personal de Arlaukas, segunda. &ltsilencio>
...
```

2. **Anotaciones lingüísticas**.

   a) **Categoría gramatical**. También llamada anotación gramatical o *pos-tagging*. Es la anotación lingüística más básica de las que se lleva a cabo. Se trata de asignar a cada unidad léxica una etiqueta con información sobre su categoría gramatical, aunque también suele incluir información sobre las características morfológicas de la unidad (género, número, caso, persona, etc.).

   A continuación puede observarse un fragmento de la anotación de categoría gramatical del LOB corpus (con el conjunto de etiquetas C1)[15]:

   Joanna_NP stubbed_VBD out_RP her_PP$ cigarette_NN with_IN unnecessary_JJ fierceness_NN ._.

   donde las etiquetas significan:

IN	preposición
JJ	adjetivo
NN	nombre común singular
NP	nombre propio singular
PP$	pronombre posesivo
RP	partícula adverbial
VBD	forma de pasado de un verbo léxico

[15] Tomado de McEnery y Wilson (1996a): p. 47.

Otro ejemplo de anotación morfosintáctica es el siguiente, tomado del Corpus CLiC-TALP

Me_PP1CS000 gusta_VMIP3S0 la_DA0FS0 cultura_NCFS000 del_SPCMS pelotazo_NCMS000 porque_CS sacrifica_VMIP3S0 la_DA0FS0 búsqueda_NCFS000 de_SPS00 lo_DA0NS0 útil_AQ0CS0 en_SPS00 favor_NCMS000 del_SPCMS cultivo_NCMS000 de_SPS00 lo_DA0NS0 admirable_AQ0CS0 ._Fp

En este ejemplo el valor de las etiquetas es el siguiente[16]:

aq0cs0	adjetivo calificativo, de género invariable y número singular
cs	conjunción subordinante
da0fs0	artículo femenino singular
da0ns0	artículo neutro singular
fp	signo de puntuación: punto
ncfs000	nombre común femenino singular
ncms000	nombre común masculino singular
pp1cs000	pronombre personal de primera persona singular y género común
sps00	preposición simple
spcms	preposición contraída masculino singular
vmip3s0	verbo principal, indicativo presente tercera persona del singular

b) **Lematización**. La lematización consiste en la asignación a cada ítem léxico de su lema, esto es, de la palabra que utilizaríamos si quisiéramos buscarla en el diccionario. Suele realizarse junto con la anotación morfológica.

La asignación de los lemas a las palabras no es algo inmediato, sino que hay que establecer los criterios de lematización a priori y tomar decisiones sobre las formas que se utilizarán como tales. Por ejemplo, en el caso de los nombres con variación de género, hay que decidir si la forma masculina y la femenina comparten el lema o si tienen lema distinto. En el caso de los pronombres personales, hay que tomar en consideración aspectos como la persona, el género o el número para determinar qué formas van a tomarse como lema.

A continuación puede observarse un ejemplo de lematización tomado del corpus Susanne (Sampson, 1995). La lematización aparece en la tercera columna[17]:

```
YB      <minbrk>        -       [Oh.Oh]
AT      The             the     [O[S[Nns:s.
NP1s    Fulton          Fulton  [Nns.
NNL1cb  County          county  .Nns]
JJ      Grand           grand   .
```

[16] En el capítulo 2 aparece el etiquetario completo así como la explicación de todas estas etiquetas.

[17] La presentación de este ejemplo se ha simplificado. La primera columna corresponde a la etiqueta morfológica de las palabras; la segunda a las palabras del texto y la última al análisis sintáctico.

```
NN1c     Jury            jury     .Nns:s]
VVDv     said            say      [Vd.Vd]
NPD1     Friday          Friday   [Nns:t.Nns:t]
AT1      an              an       [Fn:o[Ns:s.
NN1n     investigation   investigation .
IO       of              of       [Po.
NP1t     Atlanta         Atlanta  [Ns[G[Nns.Nns]
GG       +<apos>s        -        .G]
JJ       recent          recent   .
JJ       primary         primary  .
NN1n     election        election.Ns]Po]Ns:s]
VVDv     produced        produce  [Vd.Vd]
YIL      <ldquo>         -        .
ATn      +no             no       [Ns:o.
NN1u     evidence        evidence.
YIR      +<rdquo>        -        .
CST      that            that     [Fn.
DDy      any             any      [Np:s.
NN2      irregularities  irregularity    .Np:s]
VVDv     took            take     [Vd.Vd]
NNL1c    place           place    [Ns:o.Ns:o]Fn]Ns:o]Fn:o]S]
YF       +.              -        .O]
```

Otro ejemplo de lematización es el siguiente, tomado del corpus CLiC-TALP[18]:

```
Medardo_Fraile  medardo_fraile NP00000
juega           jugar          VMIP3S0
a               a              SPS00
un              uno            DI0MS0
cinismo         cinismo        NCMS000
fácil           fácil          AQ0CS0
y               y              CC
divertido       divertido      AQ0MSP
.               .              Fp
No              no             RN
quiero          querer         VMIP1S0
decir           decir          VMN0000
que             que            CS
lo              él             PP3CNA00
sea             ser            VSM03S0
,               ,              Fc
cínico          cínico         AQ0MS0
```

[18]En este caso, la primera columna corresponde a las palabras del texto; la segunda a la lematización y la tercera a la categoría morfosintáctica de cada elemento.

```
o               o               CC
divertido       divertido       AQ0MSP
,               ,               Fc
```

c) **Análisis sintáctico**. En este nivel de anotación se marcan las relaciones que establecen entre sí los diferentes ítems léxicos. Por lo general suele tomarse como punto de partida la anotación morfológica. La anotación sintáctica puede presentar diferentes niveles de profundidad: desde el llamado *skeleton parsing* o análisis superficial en el que sólo se marcan los grandes constituyentes de la oración sin tener en cuenta su estructura interna, hasta el *full parsing* en que se representa del modo más detallado posible toda la estructura de la oración.

A continuación aparece un ejemplo de análisis sintáctico tomado del PennTreeBank (Bies et al., 1995):

```
(S (SBAR-ADV (SINV had
                   (NP-SBJ Casey)
                   (VP thrown
                       (NP the ball)
                       (ADV-MNR harder))))
   ,
   (NP-SBJ it)
   (VP would
       (VP have
           (VP reached
               (NP home plate)
               (PP-TMP in
                       (NP time))))))
```

El ejemplo siguiente corresponde al análisis sintáctico parcial (*chunking*) de una oración del corpus CLiC-TALP:

```
( S
  ( sn
    ( espec.ms
      ( da0ms0 El )
    )
    ( grup.nom.ms
      ( ncms000 libro )
    )
  )
  ( S.F.R
    ( relatiu
      ( pr0cn000 que )
    )
  )
  ( gv
```

```
    ( vmip1p0 leemos )
  )
  ( gv
    ( vmip3s0 intenta )
  )
  ( S.NF.C
    ( infinitiu
      ( vmn0000 explicarlo )
    )
  )
  ...
```

La misma frase con análisis sintáctico total tiene esta forma:

```
(
  (S
    (sn
      (espec.ms
        (da0ms0 El))
      (grup.nom.ms
        (ncms000 libro)
        (S.F.R
          (relatiu
            (pr0cn000 que))
          (sn.e *0*)
          (gv
            (vmip1p0 leemos)))))
    (gv
      (vmip3s0 intenta))
    (S.NF.C
      (infinitiu
        (vmn0000 explicarlo))
...
```

d) **Semántica**. La anotación semántica puede referirse a la semántica de la oración o semántica de la palabra. En el primer caso se marcan las relaciones semánticas entre las palabras del texto (por ejemplo, el agente y el paciente de una determinada acción). En el segundo, se anotan los sentidos de cada una de las palabras en relación a una fuente externa de conocimiento (una base de datos, un diccionario, etc.) donde se detallan los diferentes sentidos.

A continuación presentamos dos ejemplos de textos etiquetados semánticamente: en primer lugar, un fragmento de SEMCOR, un corpus en que nombres, adjetivos verbos y adverbios reciben una etiqueta semántica; en segundo lugar, un fragmento del corpus del español que participó en Senseval-2, y donde sólo una palabra de cada oración está etiquetada semánticamente.

En el caso de SEMCOR (http://www.cise.ufl.edu/depot/www/wordnet/semcor.htm) las etiquetas asociadas a cada palabra tienen un formato atributo–valor y se marcan con SGML. A los nombres, verbos, adjetivos y adverbios se les asocia un sentido de WordNet. Estas palabras tienen información sobre la categoría morfológica (*pos*), el lema (*lemma*) y el número de sentido de WordNet (*wnsn*), tal como puede apreciarse en el siguiente ejemplo:

```
<contextfile concordance=brown>
<context filename=br-a01 paras=yes>
<p pnum=1>
<s snum=1>
<wf cmd=ignore pos=DT>The</wf>
<wf cmd=done rdf=group pos=NNP lemma=group wnsn=1 lexsn=1:03:00::
            pn=group>Fulton_County_Grand_Jury</wf>
<wf cmd=done pos=VB lemma=say wnsn=1 lexsn=2:32:00::>said</wf>
<wf cmd=done pos=NN lemma=friday wnsn=1 lexsn=1:28:00::>Friday</wf>
<wf cmd=ignore pos=DT>an</wf>
<wf cmd=done pos=NN lemma=investigation wnsn=1 lexsn=1:09:00::>
            investigation</wf>
<wf cmd=ignore pos=IN>of</wf>
<wf cmd=done pos=NN lemma=atlanta wnsn=1 lexsn=1:15:00::>Atlanta</wf>
<wf cmd=ignore pos=POS>'s</wf>
<wf cmd=done pos=JJ lemma=recent wnsn=2 lexsn=5:00:00:past:00>recent</wf>
<wf cmd=done pos=NN lemma=primary_election wnsn=1 lexsn=1:04:00::>
            primary_election</wf>
<wf cmd=done pos=VB lemma=produce wnsn=4 lexsn=2:39:01::>produced</wf>
<punc>``</punc>
<wf cmd=ignore pos=DT>no</wf>
<wf cmd=done pos=NN lemma=evidence wnsn=1 lexsn=1:09:00::>evidence</wf>
<punc>''</punc>
<wf cmd=ignore pos=IN>that</wf>
<wf cmd=ignore pos=DT>any</wf>
<wf cmd=done pos=NN lemma=irregularity wnsn=1 lexsn=1:04:00::>
            irregularities</wf>
<wf cmd=done pos=VB lemma=take_place wnsn=1 lexsn=2:30:00::>
            took_place</wf>
<punc>.</punc>
</s>
</p>
```

En el caso del corpus creado para la participación española en Senseval-2 (http://www.sle.sharp.co.uk/senseval2/), se etiqueta una sola palabra de cada frase del corpus. Las clases de palabras que se etiquetaron fueron nombres, adjetivos y verbos. La etiquetación se hace de acuerdo con un diccionario establecido a partir de EuroWordNet (por lo general, para cada palabra hay menos sentidos en el diccionario de los que aparecen en EuroWordNet, aunque el *synset* de EurowordNet aparece en la definición). El siguiente ejemplo, muestra una parte del corpus del nombre *bomba* al que correspondían los siguientes significados:

Diccionario:

```
bomba#NCFS#1#Artefacto explosivo#SIN:?#02172888n#
bomba#NCFS#2#Mecanismo que bombea fluidos#SIN:?#02946143#
```

Corpus:

```
000044 ( Se ha calculado que un terremoto muy violento
libera una cantidad de energía equivalente a la de cien
<tag "1">bombas</> termonucleares, o 100.000 bombas
atómicas corrientes ) .

000045 Se trataba de una especie de <tag "1">bomba</>
de relojería que ha provocado la ruina de la citada compañía .

000046 La energía en reserva equivale a la de doscientas <tag
"1">bombas</> atómicas del tipo Hiroshima .

000047 Cuando parecía que los frutos estaban el alcance
de la mano, la realidad fue muy distinta : guerras más
cruentas , si cabe , que las anteriores , genocidios de
todos los colores , la <tag "1">bomba</> atómica con sus
tétricas consecuencias , exacerbación de la exploración
del hombre por el hombre .

000048 La diferencia más importante entre un motor normal
y otro de alta presión estriba en el hecho de que en los
primeros el combustible es conducido , mediante
<tag "2">bombas</> de alta potencia , hasta la cámara de
combustión , y es impulsado por un generador autónomo cuyos
gases de escape salen al exterior sin ser aprovechados ;
en el caso del motor de alta presión , en cambio , existe
una precámara - situada delante la cámara de combustible
- en la que se queman el hidrógeno y el oxígeno.

000049 Los gases de escape accionan una turbina que genera
la energía para las <tag "2">bombas</> .
```

e) **Anotación discursiva**. En este aspecto destaca especialmente la anotación de la correferencia anafórica, en la que se marca cuál es el referente de cada uno de los pronombres que aparece en el texto.
El siguiente fragmento, tomado de Navarro et al. (2003) muestra la anotación de la correferencia:

```
<REF ID:R1 MIN: Medardo Fraile> Medardo Fraile </REF> juega a
un cinismo fácil y divertido. No quiero decir que lo sea,
```

```
cínico o divertido, sino que ante un mazo de hojas grabadas
<COREF ID:R2 REF:R1 TYPE:SUBJ-ELLIP STATUS:CIERTO> (SN) </COREF>
coloca <REF ID:3 MIN:un cristal> un cristal bien tallado </REF>
y <COREF ID:R4 REF:R3 TYPE:CLIT STATUS:CIERTO> lo </COREF>
hace girar para que el sol rompa contra  <COREF ID:R5 REF:R3
TYPE:PRON STATUS:CIERTO> él </COREF> sus rayos.
```

f) **Transcripción fonética**. Este tipo de anotación es, evidentemente, exclusivo de los corpus orales. El principal problema para este tipo de anotación es la falta de un estándar ampliamente utilizado. Los símbolos del alfabeto fonético internacional deben ser representados por caracteres accesibles a través del teclado. Un método ampliamente utilizado es el sistema SAMPA (*Speech Assessment Methods Phonetic Alphabet*) desarrollado bajo el patrocinio de la Unión Europea.

g) **Prosodia**. Este tipo de anotación pretende captar en una forma escrita los rasgos suprasegmentales del habla: acento primario, entonación y pausas.

h) ***Problem-oriented tagging***. En este caso ya no se trata de niveles de anotación ampliamente utilizados sino de anotación orientada a una tarea particular: por ejemplo, anotar sólo las aposiciones que aparecen en un texto; anotar sólo los adjuntos verbales, etc. Las dos diferencias fundamentales de este tipo de anotación respecto de todos los anteriormente comentados son, por una parte, que no es exhaustivo; por otra, que el esquema de anotación seguido no se determina por su amplia cobertura o por su neutralidad respecto de la teoría, sino por su relevancia respecto del tratamiento del problema que se desea estudiar.

1.3.2. Métodos de anotación de corpus

Tres son las formas básicas de llevar a cabo la anotación de corpus: manual, automática o semiautomática. En el primer caso, especialistas en la materia, por lo general lingüistas, anotan manualmente el texto. En el segundo, la anotación la llevan a cabo herramientas informáticas sin que haya ninguna revisión del proceso por parte del especialista. Por último, la anotación semiautomática consiste en que el experto valida (acepta) o corrige la anotación propuesta por la herramienta informática, o en que el experto lingüista interactúa con el programa para anotar el texto.

El cuadro 1.1 presenta diferentes corpus anotados sintácticamente así como el modo en que esta anotación se ha llevado a cabo. **M** significa *manual*; **SA**, semiautomática y **A**, automática.

La anotación manual presenta el grave inconveniente de tener un alto coste en términos de tiempo. Además puede presentar problemas de consistencia, sobre todo si la realizan diversas personas. La gran ventaja es que el nivel de anotación puede llegar a ser muy específico y detallado. La anotación automática es muchísimo más rápida, pero el nivel de detalle no suele ser tan fino y además hay que tener en cuenta los errores que pueden aparecer (hasta el momento ningún sistema automático ni manual presenta un nivel de error igual de cero). Por todo ello, la mayoría de sistemas de anotación son semiautomáticos:

corpus | ref.	M	SA	A
ISST(Montemagni et al., 2003)		+	
PDT(Hajic, 1998)		+	
UAM(Moreno y López, 1999)		+	
Le Monde(Abeillé, Clément, y Kinyon, 2000)		+	
NEGRA(Brants y Plaehn, 2000)		+	
TIGER(Brants et al., 2002)		+	
PennTB(Marcus, Santorini, y Marcinkiewicz, 1993)		+	
TurcoTB(Oflazer et al., 2003)		+	
TUT(Bosco et al., 2000)		+	
Floresta(Afonso et al., 2002)		+	
Port. medieval(Rocio et al., 2003)		+	
Japonés(Kurohashi y M.Nagao, 1998)		+	
BulTreeBank(Simov et al., 2002)		+	
Hebreo Moderno(Sima'an et al., 2001)		+	
Ruso(Boguslavsky et al., 2002)		+	
Sueco(Santamarta, Lindberg, y Gambäck, 1995)		+	
Susanne(Sampson, 1995)	+		
Polaco(Marciniak et al., 2003)	+		
BNC http://www.hcu.ox.ac.uk/BNC/			+
Cast3LB(Civit y Martí, 2002)		+	

Cuadro 1.1: Procesos de anotación sintáctica

ello garantiza la consistencia y reduce el tiempo necesario así como la tasa de error. Por lo general, puede afirmarse que la anotación automática es eficaz para el análisis morfológico: es rápida y la calidad es buena, ya que, como se comentará en el capítulo 3 (página 136) la tasa de error se halla en torno al del 5 % y los errores son sistemáticos y, por tanto, detectables y fácilmente corregibles. Lo mismo ocurre con la anotación sintáctica a nivel de *chunking*. Sin embargo, para el análisis sintáctico profundo y para la anotación semántica lo mejor es la anotación manual o la automática supervisada, dado que el conocimiento requerido va mucho más allá del contexto local: el contexto puede ser de frases, incluso de párrafos y debe utilizarse conocimiento externo al propio texto.

La anotación semiautomática suele llevarse a cabo mediante editores que permiten la manipulación de los datos. Estas herramientas van desde los simples editores de texto a creaciones más sofisticadas, como el **AGTKToolkit** (figura 1.1), **@nnotate** (figura 1.2) o el interfaz utilizado para la anotación del PDT (figura 1.3). El gran reto actual de los sistemas de anotación es llegar a la anotación automática de calidad.

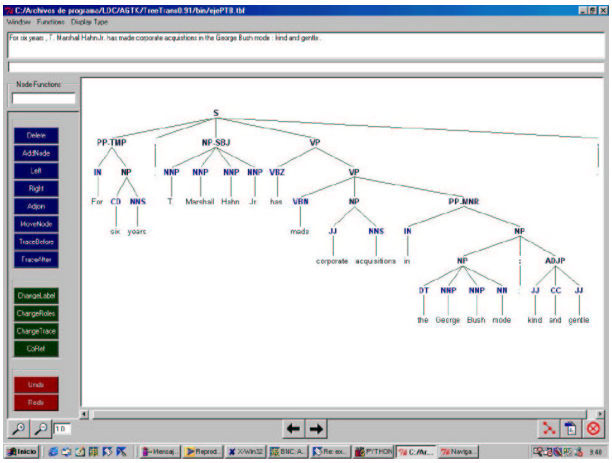

Figura 1.1: Interfaz AGTK para la anotación sintáctica

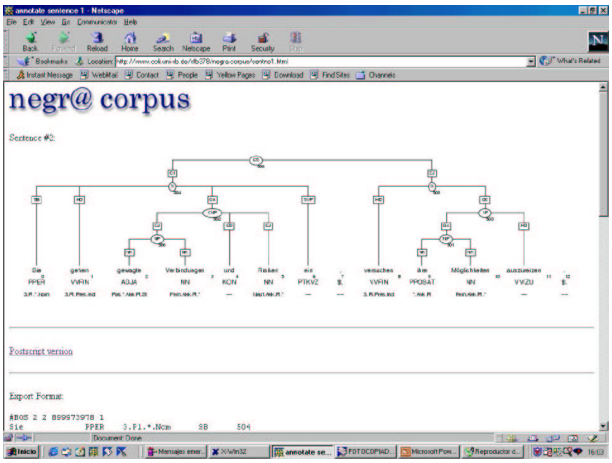

Figura 1.2: Interfaz @nnotate para la anotación sintáctica

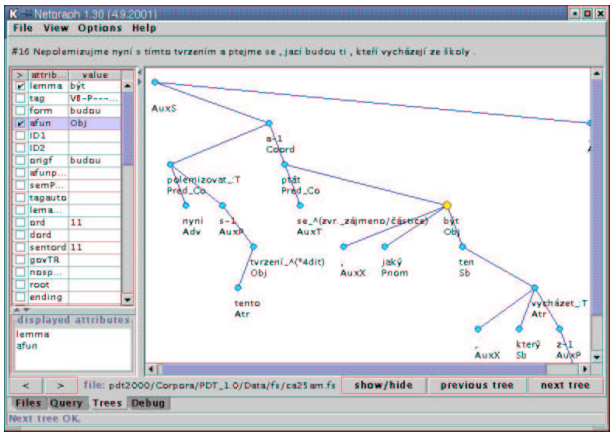

Figura 1.3: Interfaz PDT para la anotación sintáctica

En Véronis (2001a) se presenta el estado actual de las técnicas de anotación automática de corpus. Su esquema es el que reproducimos en la figura 1.4, donde *operativo* significa que existe software comercial o gratuito de gran difusión; *prototipo* denota la existencia de software en los laboratorios; e *investigación* indica que se están llevando a cabo trabajos prospectivos que todavía no se utilizan en sistemas de anotación reales.

La evaluación de los sistemas de etiquetado es compleja ya que el número de etiquetas que cada uno utiliza es distinto, como también lo es la información que contiene cada etiqueta[19], dado que ello refleja las características morfológicas de la lengua. En la anotación morfológica, los sistemas presentan, como término medio, un 95 % de acierto. Sin embargo, estos resultados deben relativizarse. Según Véronis (2001a), en general, el 60 % de las palabras no son ambiguas y, por otra parte, si se selecciona la etiqueta más frecuente para las palabras ambiguas ya se obtiene un 90 % de acierto[20]. Así, los sistemas de análisis morfológico se centran en ese 10 % de palabras ambiguas de difícil resolución. A todo esto cabe una última reflexión: ¿qué se entiende por *etiquetación correcta*? En muchos casos, lo que es *correcto* o útil para la ingeniería lingüística es bastante discutible desde el punto de vista lingüístico. Sin embargo, los sistemas automáticos son operativos.

En el caso de la sintaxis, hay dos limitaciones fundamentales a la anotación automática total: en primer lugar, nadie ha sido capaz de escribir una gramática formal lo suficientemente amplia para una lengua humana cualquiera. En segundo lugar, la desambiguación de las estructuras posibles para una frase dada implica tener acceso a información de tipo semántico y pragmático que, en la actualidad, no está al alcance de los sistemas. Por ello se lleva a cabo una anotación parcial (*shallow parsing*) que, si bien no proporciona la precisión deseada, sí presenta una amplia cobertura y robustez.

[19] Un sistema que asigne etiquetas del tipo N, V, A (para nombre, verbo, adjetivo) no es comparable a otro que asigne etiquetas como NCMS, NCFS, NCMP para nombre común masculino-singular, femenino-singular, masculino-plural, porque la extensión de cada etiqueta es distinta.

[20] Datos tomados de (Véronis, 2001a): p. 9.

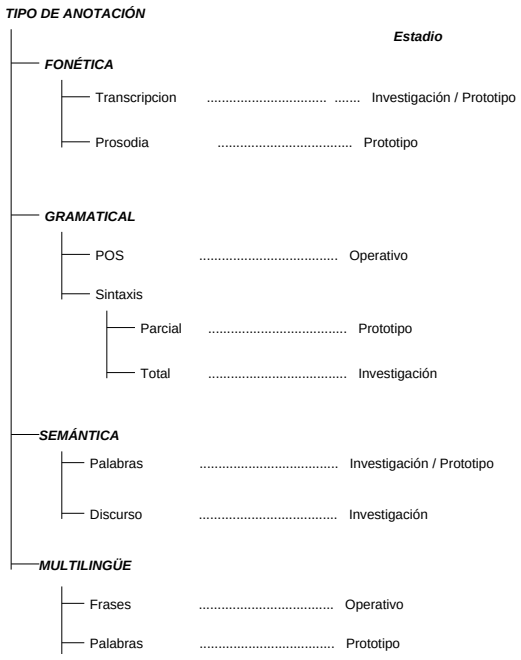

Figura 1.4: Estado de la anotación automática según J. Véronis (2001a)

1.4. La creación de Corpus

Los primeros corpus, entendidos como recopilaciones de textos en soporte informático creados con unos objetivos concretos, fueron para la lengua inglesa. Como pioneros destacan el *Brown Corpus*[21] y el *Lancaster-Oslo/Bergen Corpus (LOB)*[22]. El *Brown Corpus* fue recogido por W.N. Francis y H. Kucera, en la Brown University, en Providence, RI. Contiene un millón de palabras del inglés americano escrito procedentes de textos publicados en el año 1961. Contiene 500 textos de unas 2000 palabras cada uno y pertenecientes a 15 categorías textuales distintas. Además, hay una versión etiquetada a nivel morfológico. El *LOB* corpus fue creado en cooperación por las universidades de Lancaster, la de Oslo y el Centro Noruego para las Humanidades en Bergen. Se concibió como un equivalente del Brown Corpus pero para el inglés Británico. Los textos que lo componen también fueron publicados en el año 1961 y también consta de un millón de palabras.

Aproximadamente en la misma época se constituyó el *London-Lund Corpus of Spoken English*[23]. Este es el primer corpus de lengua oral que se recogió. Los hablantes eran

[21] Información detallada sobre este corpus puede encontrarse en la URL siguiente: http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

[22] Información detallada sobre este corpus puede encontrarse en la URL: http://www.hd.uib.no/icame/lob/lob-dir.htm

[23] Información sobre este corpus puede encontrarse en http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM y en (Svartvik y Quirk, 1980).

mayoritariamente británicos. En total, contiene unas 500.000 palabras.

En los años siguientes, se siguieron constituyendo corpus, cada vez de mayor tamaño. Destacan, para el inglés, el *British National Corpus*[24], el *International Corpus of English* y, por último, el *Bank of English*. El primero contiene cien millones de palabras del inglés moderno, tanto escrito como hablado. Fue creado por un consorcio de editores británicos, la Universidad de Oxford y la de Lancaster así como la Biblioteca Británica. El corpus está codificado siguiendo las propuestas de la TEI y utilizando el ISO standard 8879 (SGML: Standard Generalized Markup Language) para representar tanto la estructura de los textos como la salida del analizador CLAWS[25]. El *Bank of English* fue iniciado en 1991 por la editorial COBUILD y la Universidad de Birmingham. El objetivo principal de este corpus es el de proporcionar datos para la creación de diccionarios. Tiene la particularidad de ser un *monitor* corpus, esto es, un corpus al que constantemente se está añadiendo nuevo material. En la actualidad, consta de más de 320 millones de palabras. Por último, el *International Corpus of English* recoge un millón de palabras procedentes de cada país o región donde el inglés es la primera lengua. Incluye tanto textos escritos como orales. Todos los textos son posteriores a 1989. El responsable del proyecto es el *Survey of English Usage* en el *University College* de Londres.

Para el inglés, el ICAME (*International Computer Archive of Modern and Medieval English*[26]) recoge y distribuye información sobre el material disponible en formato electrónico para el procesamiento automático y para investigación lingüística. Unos objetivos parecidos tiene el *English Language Corpora and Corpus resources*[27]. Otros centros que recogen y distribuyen recursos lingüísticos son el LDC (Linguistic Data Consortium) en la Universidad de Pennsylvania[28]; el OLAC (Open Language Archives Community)[29]; OTA (The Oxford Text Archive)[30] y ELRA (European Language resources Association)[31].

[24] http://www.hcu.ox.ac.uk/BNC/index.html

[25] Garside (1987).

[26] http://www.hd.uib.no/icame.html

[27] http://www.hcu.ox.ac.uk/BNC/corpora.html

[28] *The Linguistic Data Consortium is an open consortium of universities, companies and government research laboratories. It creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes. The University of Pennsylvania is the LDC's host institution. The LDC was founded in 1992 with a grant from the Advanced Research Projects Agency (ARPA), and is partly supported by grant IRI-9528587 from the Information and Intelligent Systems division of the National Science Foundation.* Información detallada sobre este consorcio puede encontrarse en http://www.ldc.upenn.edu/

[29] *OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources.* La URL es http://www.language-archives.org.

[30] *The OTA works closely with members of the Arts and Humanities academic community to collect, catalogue, and preserve high-quality electronic texts for research and teaching. The OTA currently distributes more than 2500 resources in over 25 different languages, and is actively working to extend its catalogue of holdings.* La URL de este centro es http://ota.ahds.ac.uk/

[31] *ELRA is the driving force to make available the language resources for language engineering and to evaluate language engineering technologies. In order to achieve this goal, ELRA is active in identification, distribution, collection, validation, standardisation, improvement, in promoting the production of language resources, in supporting the infrastructure to perform evaluation campaigns and in de-*

Los corpus en otras lenguas no empezaron a desarrollarse hasta más tarde. En la sección siguiente mencionados los principales corpus del español.

1.4.1. Corpus en español

Mencionamos y describimos brevemente aquí los principales corpus existentes para el español o que contienen textos en esta lengua. Esta información es complementaria con la del informe del Instituto Cervantes del año 1996[32]. Establecemos una división entre corpus anotados lingüísticamente (a cualquier nivel) y corpus no anotados. En el primer grupo incluimos también aquellos corpus con algún tipo de marcaje (html u otros)[33].

1. **Corpus no anotados**

   a) **BEC (Biblioteca Electrónica Cristiana)**,
   disponible en *http://www.multimedios.org/*.
   Es una colección de textos de la religión cristiana.

   b) **ECI/MCI Corpus (European Corpus Initiative Multilingual Corpus)**
   disponible en *http://www.elsnet.org/resources/eciCorpus.html*
   Es un corpus de 98 millones de palabras que incluye muchas lenguas europeas, entre ellas el español.

   c) **Elaleph Com Biblioteca**
   disponible en *http://www.elaleph.com/cgi-bin/biblioteca.cgi*
   Contiene 413 textos literarios españoles.

   d) **Parnaseo LEMIR Texts (Textos de Literatura Española Medieval y del renacimiento)**
   disponible en *http://parnaseo.uv.es/Lemir.htm*
   Aquí pueden encontrarse textos medievales y renacentistas, algunos de los cuales son ediciones críticas realizadas directamente a partir de los manuscritos originales.

   e) **VISL online English Corpus**
   disponible en *http://visl.hum.ou.dk*
   VISL permite hacer preguntas on-line a textos no anotados en danés, alemán, inglés y español.

2. **Corpus anotados**

   a) **Antología del Ensayo Iberoamericano**,
   disponible en *http://ensayo.rom.uga.edu/antologia/*

veloping a scientific field of language resources and evaluation. These activities are achieved through ELRA's operational body ELDA (Evaluation & Language resources Distribution Agency). La URL es http://www.icp.grenet.fr/ELRA/home.html

[32] Instituto-Cervantes (1996).

[33] En la sección 1.4.1.1 aparecen mencionados otros corpus orales del español.

Se trata de una parte del "Repertorio Ibero e Iberoamericano de Ensayistas y Filósofos" que contiene textos filosóficos y de crítica. Todos los textos están en formato html.

b) **Bible of University of Maryland Parallel Corpus Project**,
disponible en *http://benjamin.umd.edu/parallel/bible.html*

En esta URL se proporcionan versiones de la Biblia en diferentes lenguas, entre ellas el español. Los textos están anotados siguiendo las recomendaciones del CES.

c) **CALLHOME Spanish Transcripts**,
disponible en *http://morph.ldc.upenn.edu/Catalog/LDC96T17.html*

Son transcripciones de fragmentos de conversaciones telefónicas. La señal acústica está alineada con la transcripción ortográfica. Sólo está disponible para los miembros del LDC.

d) **CHILDES Database**
disponible en *http://childes.psy.cmu.edu/*

CHILDES es un sistema que proporciona datos y herramientas para el estudio del lenguaje infantil. Uno de sus componentes es el español.

e) **Gonzalo de Berceo - Obras completas**
disponible en *http://geocities.com/urunuela1/berceo/bercoe1.htm*

Además de la obra completa de este autor se incluye bibliografía, documentación crítica y vocabulario sobre su obra. Los textos están marcados en html.

f) **Hub-4Ne - Spanish Broadcast News Transcripts 1997**
disponible en *http://morph.ldc.upenn.edu/Catalog/LDC98T29.html*

Contiene 30 horas de habla con su transcripción de noticias de radio. Las transcripciones están en formato SGML. Sólo está disponible para los miembros del LDC.

g) **Hub-5 Spanish Transcripts**
disponible en *http://morph.ldc.upenn.edu/Catalog/LDC98T27.html*

Se trata de grabaciones de conversaciones telefónicas no superiores a los 30 minutos.

h) **IntraText Library**
disponible en *http://www.eulogos.it/default.htm*

Se trata de una biblioteca con volúmenes, especialmente de temática religiosa, en nueve lenguas europeas.

i) **CRATER Parallel Corpus**
disponible en *http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html*

Se trata de una página web que permite consultas a un corpus multilingüe con los textos anotados a nivel morfológico.

j) **JOC-CES Multilingual**
disponible en *http://www.lpl.iniv-aix-fr/projects/multext/MUL4.html*

En principio se trata de corpus paralelos y anotados además con información morfológica del inglés, alemán, francés, italiano, y español procedentes del *JOC (Journal of the European Community)* y del *CES (Corpus Encoding Standard)*

k) **Lieder and Songs Texts Page**
disponible en *http://www.recmusic.org/lieder/*

En esta web pueden encontrarse textos de diferentes poetas y compositores; del total de textos, 124 están en español.

l) **Lopez Ornat Corpus**
disponible en *http://www.sis.ucm.es/Spanish*

Este corpus contiene 662 grabaciones de una niña, María, desde los 19 meses hasta los 4 años de edad. El texto está marcado a nivel lingüístico y psicolingüístico.

m) **Mark Davies' collection**
disponible en *http://www.ilstu.edu/˜ mdavies/texts.htm*

Un enlace más actualizado a este corpus es *http://corpusdelespanol.org*. Es un corpus de cien millones de palabras del español histórico y moderno. Se trata de una base de datos relacional que permite gran cantidad y diversidad de consultas. Para más información, puede consultarse Davies (2002) y también la sección 1.4.1.3 del presente trabajo.

n) **UN Parallel Text Corpus (United Nations Parallel Texts)**
disponible en *http://morph.ldc.upenn.edu/Catalog/LDC94T4A.html*

Este corpus contiene textos en inglés, francés y español de la Oficina del Servicio de Conferencias (Office of Conference Services) de la ONU entre 1988 y 1993.

ñ) **UAM Treebank**
Banco de árboles sintácticos que contiene 1500 oraciones anotadas de modo básicamente manual, elaborado por la Universidad Autónoma de Madrid (cf. sección 1.4.1.4).

o) **LexEsp**
Léxico informatizado del español (Sebastián et al., 2000). Contiene cinco millones y medio de palabras del español anotadas a nivel morfosintáctico de modo automático (cf. sección 1.4.1.5).

p) **CREA y CORDE**
Corpus sincrónico y diacrónico del español, elaborados por la Real Academia Española de la Lengua (cf. sección 1.4.1.1).

q) **CLiC-TALP**
Corpus de 100.000 palabras analizado automáticamente a nivel morfosintáctico y validado manualmente. Constituye un *gold-standard* para sistemas de desambiguación automática basados en ejemplos (cf. sección 1.4.1.5). Ha sido desarrollado por la Universitat de Barcelona y la Universitat Politècnica de Catalunya

r) **BDS**
Base de Datos sintácticos del español, desarrollada en la Universidad de Santiago, bajo la dirección del Dr. Guillermo Rojo. Contiene 160000 cláusulas anotadas sintácticamente de modo manual (cf. sección 1.4.1.2).

s) **ARTHUS**
El *Archivo de textos hispánicos de la Universidad de Santiago de Compostela* contiene en la actualidad textos pertenecientes a diferentes etapas de la historia del español. Todos ellos han sido introducidos en ordenador mediante escáner y programas de reconocimiento óptico de caracteres, están en formato ASCII y tienen una codificación mínima en formato COCOA que permite, con los programas de recuperación adecuados, conocer texto, página y línea en que se encuentran los ejemplos buscados. Información sobre este corpus puede encontrarse en http://www.bds.usc.es/corpus.html.

t) **Cast3LB**
Banco de árboles sintáctico-semánticos del español, desarrollado conjuntamente por las universidades de Barcelona, Politécnica de Catalunya, Politécnica de Valencia y la Universidad de Alicante. Contendrá, al final de su elaboración, 100000 palabras anotadas a nivel sintáctico, semántico y pragmático (cf. sección 1.4.1.6).

Consideramos que determinados corpus del español merecen una atención especial por sus características: la cantidad de datos que contienen, su representatividad y la calidad y el tipo de anotación que incluyen.

Por un lado, presentamos los corpus de la Real Academia Española de la Lengua, CREA y CORDE; en segundo lugar, la *Base de datos sintácticos del español actual (BDS)* creado en la Universidad de Santiago de Compostela; en tercer lugar, el *Corpus del Español* de Mark Davies; un Treebank del español creado en la Universidad Autónoma de Madrid; LexEsp, un corpus de 5,5 millones de palabras anotado a nivel morfológico; el corpus CLiC-TALP y el corpus CAST3LB. Dedicamos las siguientes secciones a comentar estos recursos de forma más detallada.

1.4.1.1. Los corpus CREA y CORDE

El trabajo que se está llevando a cabo en la Real Academia Española consiste en la creación de un Banco de datos del español que se presenta en dos secciones, una diacrónica, el corpus CORDE, y otra sincrónica, el corpus CREA. Juntos contienen unos 270 millones de palabras. Sobre su contenido, se indica lo siguiente: *El Banco de datos recoge el español que se emplea o se empleó en todos los territorios de habla hispana y en todas las épocas de su historia. Es importante el peso concedido al español de América, que se divide en seis grandes zonas lingüísticas, y supone un 50 % del número de registros. Además, los dos corpus se estructuran según una serie de hipercampos o géneros, que incluyen tanto la lengua literaria como la no literaria.*

Ambos corpus se han codificado con marcas SGML siguiendo las recomendaciones de la TEI, de modo que pueda facilitarse la recuperación de información así como el intercambio

de textos. Además, según consta en Sánchez et al. (1999) los corpus se han anotado a nivel morfológico. Presentamos a continuación una breve descripción de estos corpus, con comentarios sobre el origen de los textos, las áreas generales del saber que incluyen y otras características propias de cada uno.

El **Corpus CREA** es un corpus de referencia de la lengua moderna, es decir, un corpus lo suficientemente representativo como para proporcionar información relevante sobre la situación actual de la lengua española. El objetivo último es conseguir 160 millones de registros, tanto de la lengua escrita como de la oral. El período de tiempo que cubre va desde 1975 hasta la actualidad y la procedencia de los textos es muy diversa, tanto en lo referente al origen geográfico, como al medio del que provienen: radio, televisión, narrativa, periódicos y revistas. La utilidad de este corpus no sólo es la realización de los diccionarios de la Academia, donde proporcionar ejemplos de uso de las palabras o de sus combinaciones resulta esencial, sino que también es el punto de partida para la investigación lingüística o la elaboración de distintas aplicaciones y recursos lingüísticos.

Según aparece en la página web de la RAE (*http://www.rae.es*) la selección de textos para la parte escrita se ha realizado *de acuerdo con cuatro grandes criterios de clasificación, independientes entre sí: medio, origen, fecha e hipercampo*:

MEDIO

Prensa	49 %
Libros	49 %
Efímeros (material no publicado)	2 %

ORIGEN

España	50 %
Hispanoamérica	50 %

En cuanto a las **fechas**, el período 1975 – 2004 se ha dividido en lustros, y se otorga más representatividad a los textos más modernos.

Por último, por **hipercampos** se entienden áreas generales del saber, que son:

Hipercampo 1 Ciencias y tecnología
Hipercampo 2 Ciencias sociales, creencias y pensamiento
Hipercampo 3 Política, economía, comercio y finanzas
Hipercampo 4 Artes
Hipercampo 5 Ocio y vida cotidiana
Hipercampo 6 Salud
Hipercampo 7 Ficción

La parte oral del CREA se ha obtenido mediante convenios con diversas instituciones. Las grabaciones se transcriben ortográficamente y se codifican.

En la actualidad, esta parte está dividida en dos grandes grupos de textos: unos proceden de grabaciones de programas de radio y televisión, mientras que otros proceden de corpus orales previamente existentes que se han adaptado a la transcripción y codificación de la RAE. Estos corpus son:

ACUAH: Análisis de la Conversación de la Universidad de Alcalá de Henares.
ALFAL: Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico, de la Asociación de Lingüística y Filología de América Latina.
Caracas-77: Estudio sociolingüístico de Caracas, 1977.

Caracas-87: Estudio sociolingüístico de Caracas, 1987.
CEAP: Corpus de Encuestas en Asunción de Paraguay.
COVJA: Corpus oral de la variedad juvenil universitaria del español hablado en Alicante.
CSC: Corpus para el estudio del español hablado en Santiago de Compostela.
CSMV: Corpus Sociolingüístico de Mérida-Venezuela.
UAM: Corpus Oral de Referencia del Español Contemporáneo.
Y, por último, material público procedente de Internet.

El corpus oral constituye aproximadamente un 10 % del total del CREA. A fecha de 31 de octubre de 2001, es posible acceder a casi 9 millones de registros procedentes de transcripciones de la lengua hablada, con más de 1.600 documentos. Los materiales se clasifican de acuerdo con los siguientes criterios:

Medio: que se determina por valores del canal comunicativo (radio, televisión, grabación en directo, telefónica, etc.) o por valores de procedencia: si es o no una grabación, si es texto previamente transcrito o a la vez transcrito y codificado.

Origen: al igual que la parte escrita, el 50 % de textos provienen de España y el 50 % restante de Hispanoamérica.

También lo referente a la **época** se corresponde con los criterios adoptados para la parte escrita del corpus.

Por último, sobre el **género**, puede hablarse de dos subgéneros: los textos procedentes de grabaciones de radio y televisión, que constituyen la parte más importante del corpus oral y las transcripciones de discursos políticos, conversaciones telefónicas, etc.

El CREA ha sido concebido como un corpus de estructura abierta, es decir, sin final, lo que implica la actualización continua de sus datos. A medida que se van incorporando los textos más recientes al corpus, los más antiguos pasan a formar parte del CORDE, de modo que el CREA siempre abarcará únicamente los últimos veinticinco años. Los textos se mantienen en constante revisión y se introducen nuevas muestras para conservar el equilibrio del material, de modo que sea representativo de las diversas tendencias del español de hoy en día. Los textos que se están incorporando actualmente corresponden al período 2000-2004.

Según consta en la página web de la RAE, *el Corpus diacrónico del español (***CORDE***) es un corpus textual de todas las épocas y lugares en que se habló español, desde los inicios del idioma hasta el año 1975, en que limita con el Corpus de Referencia del Español Actual. El CORDE está diseñado para extraer información con la que estudiar las palabras, sus significados, la gramática y su uso a través del tiempo.*

A fecha de Octubre de 2001 contaba con un total de 136 millones de registros, procedentes de muy diversos géneros: prosa, verso y, en cada modalidad, textos líricos, dramáticos, históricos, jurídicos, etc.

Por otra parte, se pretende que el corpus recoja todas las variedades geográficas, históricas y de género, de modo que el conjunto sea lo más representativo posible.

La aplicación fundamental de este corpus será la creación de un Diccionario Histórico de la lengua española.

Los criterios de selección de textos han sido los siguientes:

MEDIO
Libro 97 %
Prensa 3 %
ORIGEN
España 74 %
Hispanoamérica 25 %
Español sefardí y otros 1 %
MODO
Prosa 85 %
Verso 15 %
ÉPOCA
Orígenes hasta 1491
1492-1712
1713-1974

1.4.1.2. BDS

La *Base de Datos del español Actual (BDS)* contiene el análisis sintáctico de unas 160.000 cláusulas (algo más de 1.450.000 palabras) que provienen de la parte contemporánea del Archivo de Textos Hispánicos de la Universidad de Santiago (ARTHUS). La construcción de esta base de datos se ha llevado a cabo de modo totalmente manual y no contiene anotación morfológica ni está lematizada. Se han anotado la estructura y los constituyentes funcionales de la cláusula, entendiéndola como *unidad gramatical que se organiza entorno a un elemento que desempeña la función de predicado* (Rojo, 2001). La BDS consiste en la codificación del análisis realizado manualmente y la conexión con el texto para permitir la visualización de los ejemplos. En la BDS cada registro contiene cinco tipos de información:

1. datos sobre el verbo que actúa como predicado así como información sobre su localización en el texto escrito;

2. datos sobre la cláusula: tipo (independiente, coordinada, de gerundio, etc.), función de la cláusula, voz, modalidad, polaridad, forma verbal utilizada, forma verbal de verbo que la domina (si es pertinente), persona y número, número de argumentos y orden de los elementos;

3. datos sobre cada función sintáctica de la cláusula[34]: sujeto, complemento directo, indirecto y complementos argumentales como suplemento, complemento adverbial, modal y otros complementos preposicionales, complemento agente, complemento predicativo (del sujeto, del objeto o de otro elemento). Cuando se ha considerado necesario también se ha marcado el tipo de elemento que realiza de función (frase nominal, demostrativo, relativo *que*, cláusula de *que* en indicativo, etc.);

4. observaciones sobre predicativos no argumentales o fenómenos como el dequeísmo;

[34]Sólo se han anotado los argumentos verbales, pero no los circunstanciales ni las estructuras de los sustantivos deverbales, por ejemplo.

5. informaciones que se han obtenido de forma automática a partir de los campos anteriores.

Dada la falta de unanimidad en torno al análisis sintáctico de ciertas construcciones o al carácter de ciertas funciones se estableció un *Manual de fichado* que recoge las decisiones tomadas durante la construcción de la BDS. En este manual se detalla, por ejemplo, el inventario de las funciones de carácter argumental; el estatus de argumento o no de ciertos elementos respecto de ciertos verbos, etc.

La figura 1.5 muestra el resultado de la consulta de verbos en un esquema con sujeto, suplemento y predicativo.

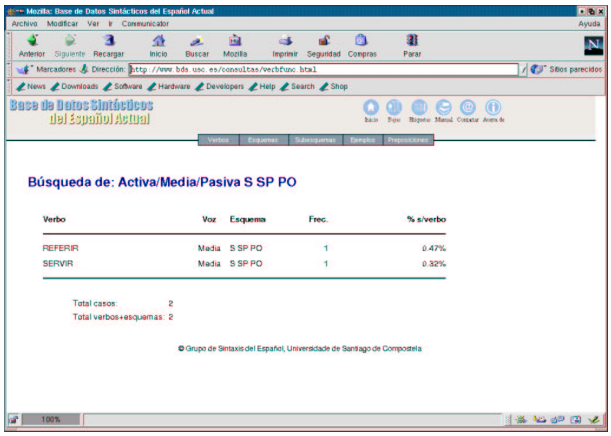

Figura 1.5: Interfaz de consulta de la BDS

1.4.1.3. Corpus del Español (de Mark Davies)

El *Corpus del Español* de Mark Davies consta de 100 millones de palabras recogidas de más de diez mil textos y transcripciones del español que abarcan un período que va desde el siglo XIII al XX. Las palabras están en una base de datos relacional. La base de datos contiene cada unigrama, bigrama y trigrama de todo el corpus. Para cada uno de estos n-gramas hay información sobre su frecuencia en cada siglo y en los tres registros en que se ha dividido el corpus (literatura, oral y miscelánea). Esta base de datos central está relacionada con muchas otras que contienen los lemas, las categorías gramaticales, sinónimos, etimologías, etc. (puede consultarse (Davies, 2002) para más detalles sobre la organización de la base de datos. Asimismo, el corpus puede consultarse en la web: http://www.corpusdelespanol.org/).

Algunas de las búsquedas que permite esta base de datos relacional son las siguientes:
(i) simple *pattern-matching* de formas:

búsqueda		**resultado**
grit*	$\Longrightarrow$	gritos, gritándoles, ...

(ii) colocaciones:

búsqueda		**resultado**
lo * posible	$\Longrightarrow$	lo {antes, mayor, máximo} posible ...
música *.adj	$\Longrightarrow$	música {clásica, folclórica, electrónica}...

(iii) lemas:

búsqueda		**resultado**
decir.*	$\Longrightarrow$	(todas las formas de este lema)

(iv) categoría gramatical:

búsqueda		**resultado**
*.v_inf	$\Longrightarrow$	(infinitivos)

(v) sinónimos y antónimos:

búsqueda		**resultado**
!inteligente	$\Longrightarrow$	vivo, capaz, agudo, ...

(vi) combinaciones de las anteriores:

búsqueda		**resultado**
!mandar.* que *.v_subj_ra	$\Longrightarrow$	hicieron que dijera
		mandó que volvieran

La figura 1.6 muestra el resultado de la busqueda del lema *resistir* en la parte contemporánea del corpus.

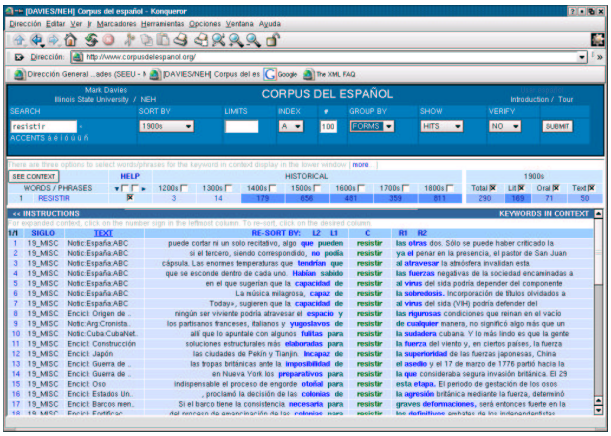

Figura 1.6: Interfaz de consulta del Corpus del Español

1.4.1.4. UAM-Treebank

La Universidad Autónoma de Madrid inició hace unos años la construcción de un TreeBank (corpus anotado con información de la estructura sintagmática). Las principales referencias de este trabajo son: Moreno y López (1999), Moreno et al. (2000) y Moreno et al. (2003). A continuación presentamos brevemente el trabajo realizado por este grupo de investigación.

El treebank constituido en la Universidad Autónoma de Madrid consta de 1000 oraciones, con un promedio de 16,1 palabras por oración, tomadas de dos fuentes distintas: por una parte, 900 oraciones provienen de *El País Digital*, mientras que, por otra, las 100

restantes provienen de revistas de la OCU. Las primeras se tomaron aisladamente, fuera de todo contexto; las segundas tienen un contexto de varios párrafos.

Las estrategias básicas de codificación de la información lingüística son las siguientes:

- en la anotación se combinan etiquetas y rasgos que especifican la información sintáctica de cada elemento;
- los niveles de anotación son:
  1. categorías sintácticas (nombre, adjetivo, etc.)
  2. funciones sintácticas (sujeto, complemento directo, indirecto, ...)
  3. rasgos sintácticos (género, número, ...)
  4. (algunos) rasgos semánticos (humano, tiempo, ...)
- la información se especifica una sola vez, para evitar la redundancia;
- en caso de duda del anotador al asignar un determinado rasgo, se opta por no asignar ninguno;
- en principio, se anotan sólo los elementos sintácticos superficiales, aunque se marcan los sujetos elípticos y la elipsis de las coordinaciones.

Sobre la ambigüedad en los análisis, la opción tomada es confiar en la intuición del anotador y no dudar en recurrir a cualquier tipo de información extratextual para desambiguar. Si ello no es posible, se marca la ambigüedad en el análisis.

El tratamiento que se da a los elementos elípticos es similar al propuesto en el PennTreeBank (Marcus, Santorini, y Marcinkiewicz, 1993). Se marca el sujeto vacío de las oraciones finitas y de las no finitas, además de marcarse la elipsis que se produce en las coordinaciones. En ambos casos se añade un índice de referencia al elemento vacío y un identificador al elemento correferente; pero en ningún caso se marcan huellas (como sí se hace en el PennTreeBank).

Las herramientas utilizadas para la construcción del treebank fueron: un *tagger* estadístico que proporcionaba la categoría y los rasgos flexivos más frecuentes para cada palabra; un *chunker* para reconocer los sintagmas nominales, adjetivos, adverbiales y preposicionales; un interfaz gráfico para anotar las oraciones; un verificador de la asignación correcta de los rasgos a cada categoría; y, por último, un generador de reglas de estructura sintagmática para detectar posibles anotaciones incorrectas.

A continuación presentamos una frase de este treebank (tomada de Moreno, López, y Sánchez (1999)):
Muerto el perro, se acabó la rabia

```
(S IMPERSONAL
   (CL ABS-PART TIME
      (VP UNTENSED PART MASC SG
         (V "<muerto>" "morir" UNTENSED PART MASC SG)
         (NP SUBJ MASC P3 SG
```

```
         (ART "<el>" "el" DEF MASC SG)
         (N "<perro>" "perro" MASC SG))))
(PUNCT "," COMMA)
(VP TENSED PAST IND SG IMPERSONAL
   (SE-MARK "<se>" "se" IMPERSONAL)
   (V "<acabó>" "acabar" TENSED PAST IND P3 SG)
   (NP OBJ1
      (ART "<la>" "la" DEF FEM SG)
      (N "<rabia>" "rabia" FEM SG))))
```

1.4.1.5. LexEsp y CLiC-TALP

El corpus LexEsp: *Léxico Informatizado del español* (Sebastián et al., 2000) consta de cinco millones y medio de palabras, recogidas entre los años 1978 y 1995. Es un corpus representativo del español estándar escrito porque presenta varios estilos narrativos, procedentes de distintas fuentes (prensa, literatura, etc.) e incluye también muestras tanto del español peninsular como del de América. Recoge un número reducido de palabras por obra y no más de tres obras por autor. Las fuentes son las que aparecen en el cuadro 1.2.

narrativa	40 %
divulgación científica	10 %
ensayo	10 %
prensa diaria	25 %
semanarios	10 %
prensa deportiva	5 %

Cuadro 1.2: Fuentes de LexEsp

Se recogen muestras de 329 novelas con unas 6000 palabras por obra aproximadamente. Las revistas de divulgación científica utilizadas han sido *Muy interesante, Mundo científico* e *Investigación y Ciencia*, así como algunos artículos de divulgación publicados en suplementos de periódicos como *El País* y *ABC*. Los fragmentos de ensayo provienen de unas 88 obras, a razón de unas 5700 palabras por obra. Las parte procedente de prensa se ha obtenido de *El País, ABC, El Mundo, El Periódico, Diario 16, El Independiente* y *La Vanguardia*. Hay que reseñar que esta parte se compone de otras tres: editoriales (15 %), articulistas (50 %) y noticias (35 %). Por último, la parte de prensa deportiva proviene de las publicaciones *As, Marca* y *Mundo deportivo*.

Este corpus se anotó a nivel morfológico de forma totalmente automática, con MACO (Carmona et al., 1998) y RELAX (Padró, 1998).

El corpus **CLiC-TALP** es un subconjunto reducido de LexEsp. En concreto contiene 100000 palabras, analizadas automáticamente con las mismas herramientas pero validado posteriormente de modo manual. Por ello constituye un corpus de referencia para estudios lingüísticos y es además un *gold-standard* utilizado como base para el aprendizaje automático para la desambiguación morfosintáctica (Civit, Castellón, y Martí, 2001a) y (Civit, Castellón, y Martí, 2001b).

1.4.1.6. Cast3LB

La elaboración del corpus **Cast3LB** forma parte de un proyecto más amplio, **3LB**, en el que participan las siguientes universidades: Universitat de Barcelona, Universitat Politècnica de Catalunya, Universidad de Alicante, Universidad Politécnica de Valencia y la Euskal Herriko Unibertsitatea. Este proyecto tiene tiene como objetivo la creación de tres bancos de árboles con anotación sintáctica, semántica y pragmática para las tres lenguas implicadas: español (Civit y Martí (2002), Civit et al. (2003), Navarro et al. (2003)), catalán y euskera (Aduriz et al. (2002), Aduriz et al. (2003)). Información detallada sobre este proyecto puede encontrarse en la siguiente página web: *http://www.dlis.ua.es/projectes/3lb*.

El principio general del proyecto es construir un sistema flexible que pueda aplicarse a diferentes lenguas, y que a la vez sea consistente en todos los niveles de anotación y con respecto a los datos.

En el nivel sintáctico, se sigue la anotación por constituyentes, con parentización y etiquetado de los mismos y marcaje adicional de las funciones sintácticas. Los principios básicos son: no tratar la elipsis (excepto en el caso de los sujetos de los verbos finitos), no alterar el orden superficial de los elementos; y realizar una anotación neutra desde el punto de vista lingüístico[35].

La figura 1.7 muestra el interfaz utilizado para la anotación sintáctica de este corpus.

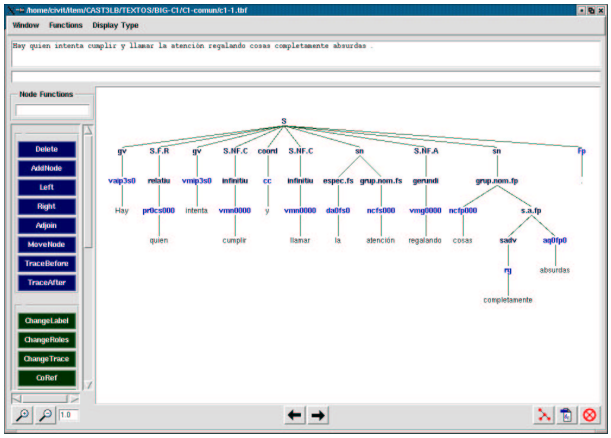

Figura 1.7: Interfaz AGTK para la anotación sintáctica

En el nivel semántico, se lleva a cabo la anotación de nombres, verbos y adjetivos con los *synsets* de EuroWordNet (Alonge et al., 1998). El proceso de anotación es *transversal*, esto es, no se anotan las palabras siguiendo el orden lineal de aparición en los textos, sino que, por orden de frecuencia, se anotarán todas las apariciones de una misma palabra en el corpus, luego las apariciones de la siguiente, y así sucesivamente.

Por último, en el nivel pragmático, se lleva a cabo la anotación de las relaciones de anáfora y correferencia. En este último aspecto, se tienen en cuenta los pronombres personales (átonos o tónicos) y los sujetos elípticos. Sin embargo, no se marcan las expresiones

[35] Véase, para más detalles sobre la anotación, el capítulo 5.

definidas. Los tipos de antecedentes que se consideran son los sintagmas nominales y otras expresiones correferentes.

En este capítulo hemos presentado el marco general en el que se inserta nuestra investigación, con una especial mención a los corpus desarrollados para el español. Algunos de ellos, **CLiC-TALP** y **Cast3LB** constituyen una parte importante del trabajo desarrollado y se tratan con mayor detalle en los capítulos 3 y 5 respectivamente.

En lo que sigue presentamos las aportaciones realizadas en este trabajo de investigación. En el capítulo 2, la propuesta de etiquetado; en el capítulo 3 las mejoras en el sistema de desambiguación así como los criterios para la desambiguación manual del corpus **CLiC-TALP**; en el capítulo 4, **GramEsp**, una gramática para el análisis superficial del español; y, por último, en el capítulo 5, los criterios para la anotación sintáctica de corpus en español.

Capítulo 2

Anotación morfológica

En este capítulo tratamos la fundamentación lingüística el sistema de anotación morfosintáctica utilizado. Tras una breve introducción a la anotación morfosintáctica de corpus, se presenta el sistema de etiquetas adoptado teniendo en cuenta tanto las propuestas estándar de Eagles como la clasificación de las palabras realizada desde la teoría lingüística.

Queremos destacar el hecho de que nuestra propuesta no parte de cero, sino de un sistema de análisis morfosintáctico previamente existente, y que muestro trabajo ha consistido en redefinir el conjunto de etiquetas utilizado, en reclasificar las palabras según este nuevo etiquetario, en eliminar inconsistencias detectadas, en simplificar y reducir el número de categorías y en fundamentar desde un punto de vista lingüístico el sistema de análisis morfosintáctico del español[1].

La figura 2.1 sitúa este trabajo en el marco de los procesos de análisis del lenguaje de CLiC-TALP.

2.1. Introducción

La anotación morfológica consiste en asignar a cada ítem léxico en el texto un código que indique su categoría gramatical o clase de palabra. Además, la mayoría de sistemas de anotación incluyen en estos códigos o etiquetas la información morfológica relevante de la palabra según la aplicación para la que se requiera dicha anotación.

La anotación morfológica se considera una base fundamental para incrementar la especificidad de la recuperación de datos y también es un fundamento esencial para posteriores formas de análisis como el análisis sintáctico y la anotación semántica (McEnery y Wilson, 1996a).

Antes de iniciar la anotación morfológica en sí hay que considerar diversos aspectos. En primer lugar cómo dividir el texto en *individual word tokens*, es decir, en lo que podemos llamar *unidades léxicas*. En segundo lugar, hay que diseñar el conjunto de etiquetas que se aplicará a esas unidades léxicas. Por último, hay que establecer los criterios que decidirán qué etiqueta debe aplicarse a cada una de las unidades léxicas. Todos estos aspectos, de

[1] Esta redefinición general se ha aplicado también al catalán.

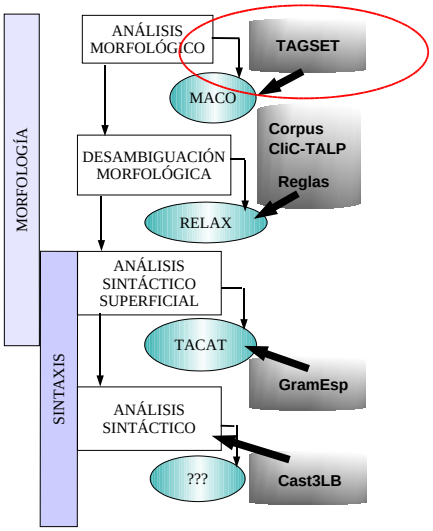

Figura 2.1: Procesos de análisis (1): establecimiento del tagset

índole lingüística, se ven no obstante influidos por otras cuestiones de índole práctica: (i) cuál va a ser el método de anotación de corpus: manual, automático o semiautomático; (ii) si la anotación va a ser automática, qué técnicas van a utilizarse; y (iii) qué niveles de rapidez, precisión y consistencia se requieren. Detallamos a continuación estos aspectos.

2.1.1. Establecimiento de las unidades léxicas

Por lo general, las palabras están delimitadas por espacios en blanco. Pero este criterio ortográfico no se corresponde siempre con los criterios morfosintácticos. Por ello se habla de *unidades léxicas* en vez de hablar de *palabras*. Revisamos a continuación los casos, en español, en que no hay una correspondencia exacta entre unidades léxicas y palabras ortográficas.

1. **Multipalabras**: *más de una palabra ortográfica se corresponde con una unidad léxica.*

   Este caso se produce, por ejemplo, en las locuciones. Sobre las multipalabras hay que tener también en cuenta si se va a permitir la presencia de elementos discontinuos. Por ejemplo: *a_ causa_ de* puede considerarse una unidad léxica formada por tres palabras ortográficas; y el conjunto puede etiquetarse como preposición. Pero, si aparece algún elemento incrustado en la locución, como en *a causa fundamentalmente de* hay que determinar el tratamiento que se dará a la unidad multipalabra.

   (2.1) *Quedarían un tanto enmascarados* **a_causa_de** *ese fiasco del rival* (d2).
   ... presentan serias dificultades **a causa fundamentalmente de** *su baja estabilidad térmica* (dc3).

En algunos esquemas de anotación se opta por los llamados *ditto tag*, que tienen la forma *tag-n-m*, donde *n* indica el número de elementos del término multipalabra y *m* la posición de la palabra actual en la secuencia. Este sistema aplicado al caso anterior daría como resultado, siendo *p* la etiqueta para la preposición: *a_p31 causa_p32 fundamentalmente_adv de_p33*.

2. **Contracciones**: *una palabra ortográfica se corresponde con más de una unidad léxica.*

   En este caso se hallan, en español, las palabras *al* y *del* y los infinitivos y gerundios seguidos de clíticos

   (2.2) *La vida sexual* **del** *titi [...] es sorprendente* (dc1).
   ... con Antonio Peñalver encadenado **al** *cabecero de la cama* (c1).
   ... cómo nos ayudaría a tomar posesión de ella sin **vulgarizarla**, **degradarla**, **envilecerla** - *según los casos* (a10).
   Nada nos repugnará tanto como **encontrárnoslas** *en mitad de la calle a plena luz del día* (a12).

2.1.2. El conjunto de etiquetas

El conjunto de etiquetas es la lista de etiquetas utilizada para una determinada tarea de anotación gramatical (Leech, 1997a).

La situación ideal es la de fundamentar lingüísticamente el conjunto de etiquetas. Pero en la práctica es posible que una etiqueta bien fundamentada desde el punto de vista lingüístico sea inaplicable de modo automático. Si una distinción gramatical es difícil de establecer utilizando sólo contexto local, ésta será probablemente descartada por el sistema automático de análisis. Por ello debe haber una solución de compromiso entre lo que es deseable lingüísticamente y lo que es realizable computacionalmente. Un ejemplo de esta situación aparece, por ejemplo, si se desea en español establecer la distinción entre las formas "LE" dativas, como las que aparecen en los ejemplos 2.3 (a-b) y acusativas (leísmo), como las de 2.3 (c-d), como en las siguientes oraciones:

(2.3) (a) **le** *llamaban Oso Hormiguero porque...* (t2).
(b) **le** *había abandonado para siempre...* (t6).
(c) *jamás* **le** *faltaban los clientes porque no se negaba a nada* (t2).
(d) *... contemplando la imagen que* **le** *devolvía el espejo* (t6).

Leech (1997a) define *etiqueta* como *a word-class embodied in an annotative device associated with a word in the text.* Cuando se definen las etiquetas hay que tener en cuenta tres criterios básicos: concisión, claridad y analizabilidad. Por una parte, es deseable que las etiquetas sean lo más breves posible. Así, por ejemplo, *NC* es preferible a *Nombre_común*. Por otra parte, cuanto más claras sean las etiquetas más fácil es para las personas recordarlas y trabajar con ellas. Por ejemplo, para los nombre comunes *NC* es preferible a *XY*. Dado que la codificación de la información es arbitraria, nada impide utilizar *XY*, pero las etiquetas mnemotécnicas (como *NC*) facilitan el trabajo. Por último, la analizabilidad de la etiqueta significa que es más fácil el trabajo si la etiqueta puede descomponerse para trabajar sólo con alguno de sus componentes. Por ejemplo, comparemos dos sistemas (imaginarios):

Glosa	Sistema 1	Sistema 2
nombre_común_masculino_singular	NCMS	N
nombre_común_femenino_singular	NCFS	O
nombre_común_masculino_plural	NCMP	P
nombre_común_femenino_plural	NCFP	Q

Si lo que queremos es extraer del corpus sólo los nombres comunes, en el primer caso (sistema 1) bastará con extraer todas las palabras cuya etiqueta empiece por NC; en el segundo caso habrá que hacer cuatro consultas (N, O, P, Q). Si queremos extraer sólo los nombres femeninos, en el primer caso seleccionaremos NCF; en cambio en el segundo, habrá que seleccionar O y Q.

Quizá el ideal es establecer un conjunto lógico de etiquetas, tal como propone Leech (1997a): p. 27:

> *the idea of a logical tagset is that the relations between the word categories symbolized by tags should be representable as a hierarchical tree with attributes being inherited from one level of the tree to another.*

Un ejemplo de esta propuesta es el conjunto de etiquetas C7 utilizado para el BNC por el equipo de Lancaster, del que reproducimos un fragmento (tomado de Leech (1997a): p. 28):

```
(C7) ---J---J--------------------JJ     general adjective, unmarked
       |   |    |                        (good, old)
       |   |    |-R---------------JJR    general adjective, comparative
       |   |    |                        (better, older)
       |   |    |-T---------------JJT    general adjective, superlative
       |   |                             (best, oldest)
       |   |-K--------------------JK     catenative adjective
       |                                 ([be] able, willing [to])
       |
       |
       |-A---T--------------------AT     article, neutral for number
           |    |                        (the, no)
           |    |-1---------------AT1    article, singular
           |                             (a, an)
           |-PPGE-----------------APPGE  possessive determiner
                                         (my, their)
```

Por su parte, el sistema de codificación de la información morfosintáctica que se propone desde Eagles no sigue este esquema: no es un sistema lógico como el que propone Leech (1997a), sino que sigue otros principios, como por ejemplo el que todas las etiquetas para una categoría gramatical tengan el mismo número de dígitos y que cada posición de esa etiqueta represente siempre al mismo atributo, tal como puede observarse en el siguiente ejemplo extraído de EAGLES (1996c) que también sigue EAGLES (1996b), uno de cuyos autores es el mismo G. Leech. Por ejemplo, para la categoría Pronombre-Determinante **PD** se propone la siguiente etiquetación:

(i)	Persona:	1.Primera	2.Segunda	3.Tercera	
(ii)	Género:	1.Masculino	2.Femenino	3.Neutro	
(iii)	Número:	1.Singular	2.Plural		
(iv)	Posesivo:	1.Singular	2.Plural		
(v)	Caso:	1.Nominativo	2.Genitivo	3.Dativo	4.Acusativo
		5.No-genitivo	6.Oblicuo		
(vi)	Categoría:	1.Pronombre	2.Determinante	3.Ambos	
(vii)	Tipo-pron.:	1.Demostrativo	2.Indefinido	3.Posesivo	4.Interrog./Rel.
		5.Pers./Refl.			
(viii)	Tipo-Det.:	1.Demostrativo	2.Indefinido	3.Posesivo	4.Interrog./Rel.
		5.Partitivo			

Si la aplicamos a algunas palabras del español el resultado podría ser:

esto PD-31–11-

nuestro PD1112-333

Cada dígito se utiliza para un atributo determinado y todas las palabras que pertenezcan a esa categoría deben presentar una etiqueta con el mismo número de dígitos. Si algún atributo no es aplicable para una palabra concreta puede subespecificarse con **-** o con **0** o de la forma que se desee, aunque esa posición no puede desaparecer. Para las otras categorías es de aplicación el mismo principio: todas las etiquetas de una determinada categoría deben tener el mismo número de dígitos, aunque este número puede diferir entre categorías.

Un caso extremo de este proceder es el adoptado por el equipo que desarrolló el *Prague Dependency Treebank* ((Bemova et al., 1999) y (Hajic y Hladká, 1998)). Todas las etiquetas de todas las categorías presentan 15 dígitos y cada uno de ellos se utiliza para un rasgo morfológico concreto. La primera posición es siempre para la categoría morfosintáctica:

N	nombre	D	adverbio	T	partícula
V	verbo	C	numeral	Z	puntuación
A	adjetivo	R	preposición	X	indefinido
P	pronombre	J	conjunción	I	interjección

La segunda posición es para la subcategoría; la tercera para el género; la cuarta para el número; la quinta para el caso; etc. Si un atributo no es pertinente para una determinada palabra, su posición la ocupa un guión **-**. Por ejemplo, las palabras de la frase *Sance je presto minimalní.* (literalmente 'una posibilidad es incluso tan mínima') reciben entre muchas otras las siguientes etiquetas (la correcta es la que figura en primera lugar)[2]:

```
sance      NNFS1-----A----
           NNFP1-----A----
           NNFP4-----A----
           NNFP5-----A----
je         VB-S---3P-AA---
           PPNS4--3-------
presto     Dg-------1A----
minimalní  AAFS1----1A----
```

[2]Ejemplo tomado de Bemova et al. (1999).

```
          AAFP1----1A----
          AAFP4----1A----
          AAIP1----1A----
          AANP4----1A----
.         F:-------------
```

Otro elemento a tener en cuenta es el tamaño del conjunto de etiquetas. Ello depende del nivel de especificidad que se desee denotar. Por ejemplo, es posible anotar sólo las categorías morfológicas de las palabras (nombre, verbo, adjetivo, etc.), añadir subcategorías (nombre_común, nombre_propio, adjetivo_calificativo, adjetivo_determinativo, etc.), o bien, por último, añadir rasgos morfológicos (nombre_común_masculino_singular_nominativo; adjetivo_determinativo_masculino_singular_nominativo, etc.).

A continuación presentamos un cuadro comparativo con el tamaño del conjunto de etiquetas utilizadas en distintos corpus[3].

Corpus	**Número de etiquetas**		**Referencia**
	Cat.	**Total**	
TOSCA		32	(Leech, 1997a)
PennTB		36	(id.)
BNC (C5)		61	(id.)
Brown		77	(id.)
LOB		132	(id.)
LondonLund		197	(id.)
TOSCA–ICE		270	(id.)
TurcoTB	12		(Oflazer et al., 2003)
Port. medieval	24		(Rocio et al., 2003)
Japonés	43		(Kurohashi y M.Nagao, 1998)
Hebreo Moderno	31		(Sima'an et al., 2001)
ISST		236	(Montemagni et al., 2003)
PDT		3030	(Hajic, 1998)
Le Monde		250-212	(Abeillé, Clément, y Kinyon, 2000)
NEGRA		54	(Brants y Plaehn, 2000)
PennTB		48	(Marcus, Santorini, y Marcinkiewicz, 1993)
Susanne		355	(Sampson, 1995)

Por lo general, el tamaño del conjunto de etiquetas crece proporcionalmente a la riqueza morfológica inflectiva de la lengua. Esto puede observarse claramente si se compara el número de etiquetas morfológicas utilizadas para el Prague Dependency Treebank (3030) con lo que ocurre en la mayoría de corpus del inglés, donde si exceptuamos el conjunto de etiquetas del Susanne Corpus, raramente se sobrepasa el centenar de etiquetas.

En la determinación del conjunto de etiquetas intervienen dos factores que (Leech, 1997a) clasifica en internos y externos. Los factores externos son de índole lingüística y tienen que ver con los desiderata del usuario, mientras que los internos o computacionales están relacionados con el hecho de conocer hasta qué punto una etiqueta es útil para el

[3]Este cuadro no es exhaustivo; las siete primeras referencias son citas de Leech (1997a)

proceso de desambiguación y para aumentar la precisión del desambiguador[4].

2.2. Sistema de análisis adoptado y definición del tagset

Los procesos automáticos de anotación de corpus responden generalmente al esquema que aparece en la figura 2.2, adaptado de McEnery y Wilson (1996a).

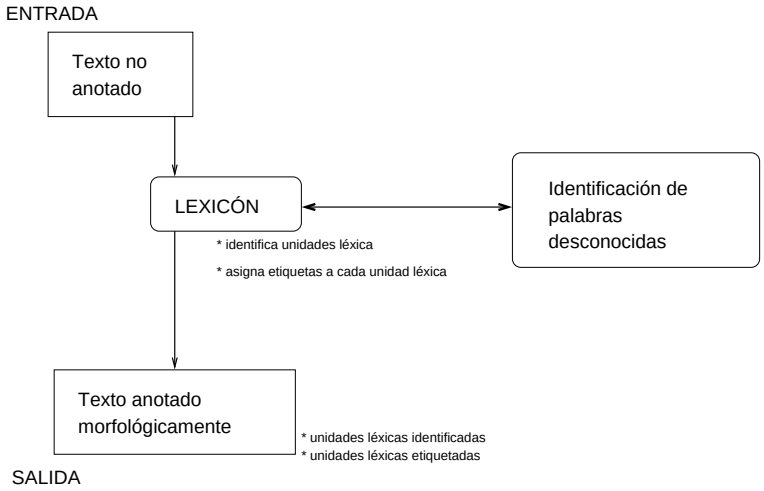

Figura 2.2: Esquema básico de los sistemas de anotación morfológica

Tales sistemas toman como input el texto escrito. Tras la segmentación (*tokenization*) del texto en palabras, el sistema intenta ver si cada palabra está presente en el lexicón de formas o *formario* en soporte electrónico. Si es así, el sistema asigna la lista de etiquetas asociadas a esa palabra en el lexicón a la palabra en el texto. Si no es así, el abanico de posibilidades se abre. Una de ellas es el análisis morfológico propiamente dicho, aunque en general más que análisis es *guesswork*, basado, por ejemplo, en los finales de palabra. Un caso concreto podría ser la palabra *tokenización*. Si no se halla en el lexicón pero el sistema puede segmentar la palabra y leer su final en *-ción*, entonces puede asignarle la etiqueta de nombre_común_femenino_singular a partir de un archivo donde se especifique que estos finales pertenecen a este tipo de nombres. En otros sistemas, la identificación de palabras desconocidas se lleva a cabo en fases posteriores cuando, por ejemplo, ya se conoce la etiqueta de la palabra anterior y la de la posterior y puede calcularse probabilísticamente la etiqueta del elemento desconocido.

Por lo general estos sistemas son modulares: pequeños módulos con tareas muy concretas se encadenan para constituir el analizador. Por ejemplo, en el esquema de análisis CLiC-TALP el analizador morfológico **MACO** consta de varios módulos, cada uno de ellos especializado en el tratamiento de distintos fenómenos. La figura 2.3 muestra esta concatenación de módulos, que presentamos a continuación de modo breve.

[4]Sobre la desambiguación morfológica, cf. capítulo 3.

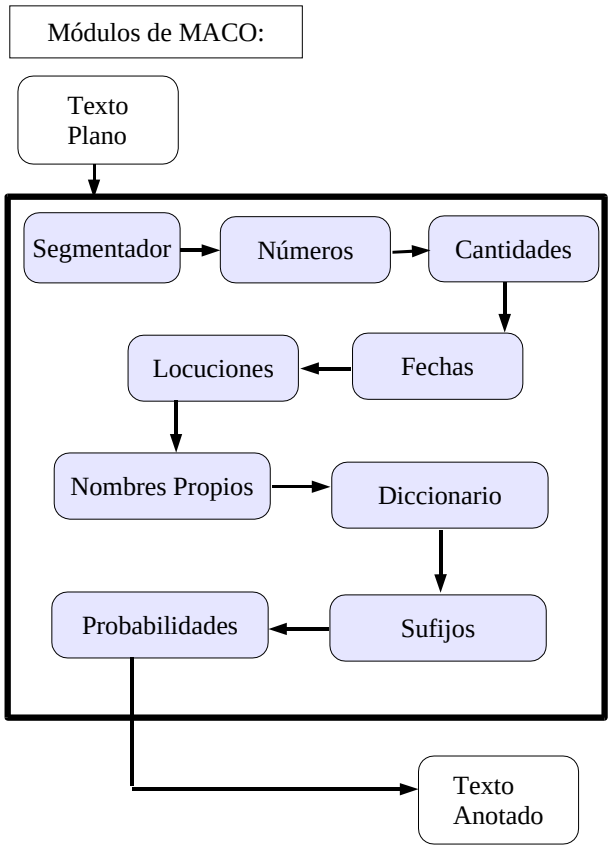

Figura 2.3: Módulos del analizador morfológico MACO

En primer lugar aparece el segmentador. Este módulo es el encargado de verticalizar el texto y tratar los signos de puntuación y las abreviaturas (especificadas en un archivo aparte). Algunas de las abreviaturas reconocidas son las siguientes:

s.a.	sr.	sra.	mr.	mrs.	ilm.
ilmo.	ilma.	ex.	excm.	excmo.	excma.
jr.	ma.	dr.	dra.	ed.	pd.
ps.	ss.	mm.	cm.		

El módulo de números, además de reconocer expresiones numéricas, convierte números expresados en palabras en su valor en cifras. El módulo de cantidades reconoce cantidades monetarias y porcentajes. El módulo de fechas reconoce estas expresiones: desde días de la semana, a años, incluso siglos, como puede apreciarse en los siguientes ejemplos:

```
el día 4 de mayo
          el el DA0MS0
          día_4_de_mayo [??:4/5/??:??.??] W
```

```
el siglo XX
          el el DAOMS0
          siglo_XX [s:xx] W

el martes
          el el DAOMS0
          martes [martes:??/??/??:??.??] W
```

El módulo de locuciones, trabaja con un archivo en el que están declaradas expresiones que pueden formar locuciones. Respecto de lo comentado en la sección 2.1.1 sobre el tratamiento de las unidades léxicas, es preciso señalar, por una parte, que hemos dado a las expresiones multipalabra una única etiqueta correspondiente a su categoría morfosintáctica y que no hemos utilizado los *ditto-tags*; por otra, las contracciones se han tratado como una única unidad a la que se asigna la etiqueta de *preposición*[5]. Estas expresiones multipalabra están clasificadas en dos tipos: locuciones ambiguas y locuciones inambiguas. Las primeras son expresiones que en un determinado contexto pueden actuar como locuciones, y en otros no, como *esto es*; las segundas son expresiones que siempre actúan como locuciones, como *pese a que*. En este último caso, el módulo une las palabras y les otorga la categoría de adverbio, preposición o conjunción, según les corresponda[6]. En el primer caso las marca como posibles inicio o final de locución para que el desambiguador determine cuál es la etiqueta adecuada en función del contexto. Aparecen como locuciones inambiguas los pronombres posesivos, puesto que están formados por artículo más posesivo[7].

El módulo de nombres propios reconoce como tales cualquier secuencia de palabras que empiece por mayúsculas y que no esté ya marcada (dado que este módulo es posterior al de abreviaturas y fechas, es posible que algunas unidades del texto hayan ya recibido una etiqueta). Estas secuencias pueden incluir palabras funcionales, como en *Banco de la Pequeña Empresa* que se trata como una sola unidad léxica. Sin embargo, si la palabra funcional está en mayúscula y está al principio de frase no se considera parte del nombre propio.

Tras la detección de los nombres propios interviene el diccionario del analizador, que es el que reconoce el resto de palabras. En la actualidad este diccionario contiene 921051 formas que representan un total de 1134441 interpretaciones[8]

El penúltimo módulo es el de sufijos; en este módulo se trata de encontrar una interpretación para una palabra mediante un tratamiento de los sufijos: se eliminan los caracteres al final de la palabra y se comprueba si el segmento resultante es una forma del diccionario del analizador. De este modo se tratan determinados procesos como la formación de los adverbios en *-mente*, las formas verbales con clíticos, la derivación apreciativa, etc, dado que estas formas no se hallan declaradas en el diccionario. Este módulo establece, en el primer campo el *sufijo* -, entendiendo por *sufijo* el final de palabra que hay que eliminar; en el segundo el *sufijo* + a añadir para reconstruir el lema (si aparece un * ello significa

[5] Véase la sección 2.2.9.2 para más detalles sobre este tema.

[6] Aunque hay pocas, también hay expresiones multipalabra con categoría nominal y adjetiva.

[7] Cf. sección 2.2.5.6 para más detalles sobre el tratamiento de estas formas.

[8] Este diccionario se obtiene a partir de un generador de formas que recibe como input las raíces y los sufijos junto con las reglas combinatorias de los mismos (Arévalo et al., 2001).

que no hay que añadir ningún sufijo); en el tercero, se impone una condición sobre la etiqueta morfosintáctica de la palabra candidata (*pos lema*); en el cuarto se señala la etiqueta morfológica que debe asignarse a la palabra de salida (si aparece un * se mantiene la misma etiqueta que tiene el lema); en el quinto (*+ acento*) se indica si debe probarse lo mismo acentuando la última letra (1) o no (0); en el sexto (*- acento*), eliminando los acentos; por último, el séptimo campo (*pos salida*) indica cuál es el lema que debe asignarse a la palabra de salida (si el valor es 0, ello indica que el lema de la palabra de salida no es la propia forma sino el lema hallada tras la supresión del sufijo). Una muestra del archivo de sufijos es la siguiente:

sufijo -	**sufijo +**	**pos lema**	**pos salida**	**+ acento**	**- acento**	**lema salida**
lo	*	^ V	*	1	1	0
la	*	^ V	*	1	1	0
los	*	^ V	*	1	1	0
las	*	^ V	*	1	1	0
le	*	^ V	*	1	1	0
les	*	^ V	*	1	1	0
me	*	^ V	*	1	1	0
te	*	^ V	*	1	1	0
nos	*	^ V	*	1	1	0
os	*	^ V	*	1	1	0
se	*	^ V	*	1	1	0

El último módulo, el de probabilidades, asigna las probabilidades léxicas a cada etiqueta propuesta que luego utilizará el desambiguador automático (véase capítulo 3) como punto de partida para realizar sus cálculos. Por otra parte, este módulo es el que asigna las posibles etiquetas a las palabras *desconocidas*, es decir, a aquellas que no han sido tratadas por ninguno de los módulos anteriores.

Además de todo esto, MACO realiza también la lematización del texto (véase la sección 2.3).

Los elementos que figuran en el analizador morfológico MACO se han obtenido de diversas fuentes. Por un lado se han recogido datos procedentes de diccionarios, y, por otro, se han utilizado corpus para obtener nuevas palabras que incorporar al analizador. El proceso de mantenimiento de este recurso es constante.

En este capítulo nos ocupamos fundamentalmente del diccionario, puesto que es el módulo que contiene la mayoría de palabras. En adelante, comentamos los criterios de anotación y clasificación de las palabras.

2.2.1. La propuesta de codificación del grupo Eagles como punto de partida

Para la etiquetación de palabras en el analizador hemos tomado la propuesta de Eagles que hemos adaptado al español. El objetivo era utilizar un sistema estándar para hacerlo compatible con otros sistemas existentes para otras lenguas.

Presentamos en este apartado el resultado del trabajo del grupo Eagles dedicado a

la anotación de Lexicones y Corpora tal como aparece en *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages*[9]. En el apartado siguiente detallamos la adaptación al español de esta propuesta.

Eagles propone un estándar de codificación basado en el estudio comparativo de lexicones ya construidos y de corpus ya anotados hasta el momento, por lo que el resultado del trabajo es más una síntesis de lo ya existente que una propuesta original, en la que se aporten criterios de discriminación entre categorías o criterios de adscripción de las palabras a las categorías. Esta propuesta admite diferentes soluciones para un mismo problema. Así, por ejemplo, se ofrecen dos formas de categorizar los demostrativos modificadores del nombre, bien como adjetivos (de tipo determinativo demostrativo), bien como determinantes (demostrativos)[10]. El objetivo último de Eagles es proponer una codificación consensuada que permita la comparación o correspondencia (*mappability*) entre distintos recursos existentes a nivel europeo con vistas a la reutilización y al uso compartido de los ya existentes y de los que puedan crearse.

En efecto, el procedimiento seguido para la elaboración de estas pautas consistió, por un lado, en examinar las principales prácticas de anotación de lexicones y corpora con el fin de llegar a una propuesta consensuada; por otro lado, se llevó a cabo un test de los resultados sobre distintas lenguas europeas para comprobar la viabilidad del procedimiento. De todo ello, resultó un conjunto común de posibles distinciones morfosintácticas que debían o podían codificarse y que son las que se ofrecen como estándar de anotación.

En este documento (EAGLES, 1996c) se propone un conjunto básico de rasgos nucleares para el etiquetado de lexicones. Estos rasgos se distribuyen en tres grados de obligatoriedad o niveles de profundidad:

- **obligatorio (L0)**, que sólo incluye la categoría o *pos* de la palabra;
- **recomendado (L1)**, que es considerado como el núcleo mínimo de rasgos que deben codificarse, y
- **opcional (L2)**, que afecta a rasgos que no suelen codificarse de un modo generalizado o a rasgos pertinentes sólo para una o muy pocas lenguas.

Las categorías que se contemplan (nivel L0) son 12: nombre, verbo, adjetivo, pronombre, determinante, artículo, adverbio, adposición, conjunción, numeral, interjección, residual.

[9](EAGLES, 1996c).

[10]En esta misma situación se hallan los restantes determinantes. Un caso similar de doble categorización se da en palabras como *dónde, cómo*, puesto que podrían aparecer bien como adverbios bien como pronombres.

2.2.1.1. Sistema de codificación

En este apartado presentamos exhaustivamente los dos primeros niveles (L0, L1)[11], mientras que para el nivel L2 reproduciremos sólo los rasgos que, según este grupo de trabajo, pueden aplicarse al español.

Para los **nombres** se proponen los siguientes descriptores:

L0	Nombre			
L1	Tipo:	Común	Propio	
	Género:	Masculino	Femenino	Neutro
	Número:	Singular	Plural	
L2	Género:	Común		
	Número:	Invariable		

El *género* común y el *número* invariable no son propiamente un género y un número más que añadir a los que figuran en el nivel L1. Se trata de valores indeterminados que indican que el género y el número no pueden determinarse únicamente a partir de la forma de la palabra, sino que se necesita del contexto para determinarlos. Según figura en el propio documento (p. 38):

> *The values 'common' -c- and 'invariant' -n-, [...], for the attributes Number and Gender respectively, raise an important issue: they are not a gender or a number, buy they are a signal that gender or number cannot be determined from the form of the word alone.*

Para el **verbo** la propuesta es la siguiente:

L0	Verbo				
L1	Tipo:	Principal	Auxiliar	Modal	
	Finitud:	Finito	No-finito		
	Verb-form/Modo:	Indicativo	Subjuntivo	Imperativo	Condicional
		Infinitivo	Participio	Gerundio	Supino
	Tiempo:	Presente	Imperfecto	Futuro	Pasado
	Persona:	1	2	3	
	Número:	Singular	Plural		
	Género:	Masculino	Femenino		

La *finitud* y el *modo* son rasgos que se superponen. Su presencia responde al hecho de querer reflejar las distintas tradiciones de análisis de las formas verbales: la inglesa y la románica. Los valores que presentan ambos atributos son compartidos, hecho que puede representarse tal como aparece en la figura 2.4. Las formas finitas equivalen a los modos indicativo, imperativo, subjuntivo y condicional; mientras que las no-finitas, equivalen a los modos no personales del verbo.

En el caso del **adjetivo** los rasgos que se proponen son:

[11] Nos limitamos a reproducir literalmente la propuesta, ya que como se podrá observar, hay rasgos de los que se proponen que no pueden aplicarse al español porque no existen en la lengua: género neutro en el nombre, caso genitivo en el pronombre personal, etc., y que sin embargo se reflejan en el documento porque aparecen en los lexicones y corpus de otras lenguas utilizados como base para la propuesta.

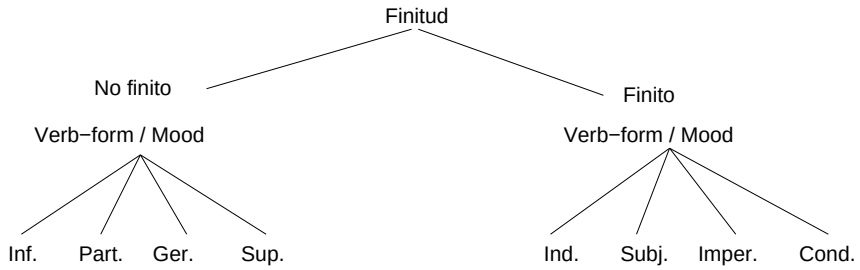

Figura 2.4: Relación *Finitud – Modo*

L0	Adjetivo					
L1	Tipo:	Calificativo	Posesivo	Indefinido	Cardinal	Ordinal
	Grado:	Positivo	Comparativo	Superlativo		
	Género:	Masculino	Femenino	Neutro		
	Número:	Singular	Plural			
	Poseedor:	Singular	Plural			
	Persona:	1	2	3		
L2	Género:	Común				
	Número:	Invariable				

En la propuesta Eagles se parte de una subclasificación de los adjetivos entre calificativos e indicativos (o determinativos). Entre estos últimos se incluyen los posesivos, demostrativos, relativos, indefinidos, numerales, interrogativos y exclamativos, aunque sólo aparecen en el nivel L1, además de los calificativos, los posesivos, indefinidos y numerales (cardinales y ordinales). Los comentarios sobre el atributo *Tipo* de la clase adjetivo finalizan con la siguiente recomendación: *Each language-specific application should specify clearly in which category the indicative adjectives are treated in order for cross-linguistic comparisons to be made possible* (p. 120). Una vez más, como ya se ha comentado anteriormente, se hace patente la redundancia del sistema de anotación propuesto, ya que una misma subclase de palabras puede pertenecer a diferentes categorías.

Por último, es de notar que más adelante se propone la categoría **Determinante** en la que aparecen también los aquí llamados adjetivos determinativos.

Los **pronombres** deberían contener los siguientes atributos:

L0	Pronombre					
L1	Tipo:	Demostrativo	Indefinido	Posesivo	Interrogativo	Exclamativo
		Relativo	Personal	Reflexivo	Recíproco	
	Persona:	1	2	3		
	Género:	Masculino	Femenino	Neutro		
	Número:	Singular	Plural			
	Caso:	Nominativo	Genitivo	Dativo	Acusativo	Oblicuo
		Objetivo				
	Poseedor:	Singular	Plural			
L2	*Politeness*	Pol	Fam			

Desde Eagles se propone una distinción entre pronombres y determinantes. Los atributos de Persona, Caso y Poseedor no son aplicables a todos los tipos de pronombres. En el nivel opcional se propone un atributo *Politeness* para marcar diferentes usos de algunas formas concretas de pronombres.

En cuanto a los **determinantes** la propuesta es la siguiente:

L0	Determinante					
L1	Tipo:	Demostrativo	Indefinido	Posesivo	Interrogativo	Relativo
	Persona:	1	2	3		
	Género:	Masculino	Femenino	Neutro		
	Número:	Singular	Plural			
	Poseedor:	Singular	Plural			

Este cuadro demuestra, junto con lo anteriormente mencionado sobre los pronombres y los adjetivos, que el trabajo del grupo Eagles consistió más en una recopilación de datos y propuestas ya existentes que en la elaboración de un marco unitario de trabajo. Según todo esto, la palabra *suyas* podría etiquetarse como adjetivo o como determinante, puesto que los posesivos aparecen como tipos de ambas categorías.

Los **artículos** se proponen aquí como una categoría separada, aunque se indica en los comentarios que en algunos lexicones y corpus se incluyen en la categoría de los determinantes:

L0	Artículo			
L1	Tipo:	Definido	Indefinido	
	Género:	Masculino	Femenino	Neutro
	Número:	Singular	Plural	

Los **adverbios** presentan los siguientes atributos:

L0	Adverbio			
L1	Tipo:	General	Partícula	
	Grado:	Positivo	Comparativo	Superlativo

Las **adposiciones** reciben las etiquetas siguientes:

L0	Adposición			
L1	Tipo:	Preposición	Posposición	Circumposición
	Formación:	Simple	Contracción	
	Género:	Masculino	Femenino	Neutro
	Número:	Singular	Plural	

Para las **conjunciones** se proponen sólo dos niveles:

L0	Conjunción		
L1	Tipo:	Coordinante	Subordinante

Los **numerales** se presentan también como una categoría aparte y con los siguientes atributos:

L0	Numeral			
L1	Tipo:	Cardinal	Ordinal	
	Género:	Masculino	Femenino	Neutro
	Número:	Singular	Plural	

En los comentarios que siguen a la propuesta se señala que se deja abierta la opción de tratar los numerales como una categoría separada del resto o de incluirlos en las clases pronombre, determinante o adjetivo.

Las **interjecciones** aparecen como categoría pero sin que se especifique ningún rasgo para ellas.

L0	Interjección

Por último se propone una clase **residual** que debe contener, en la medida de lo posible, el menor número de elementos. El ideal es que los fenómenos aquí incluidos queden asimilados a las categorías anteriores.

L0	Residual			
L1	Tipo:	Palabras extranjeras	Abreviaturas	Contracciones
		Símbolos alfabéticos	Inclasificados	Formas truncadas
		Fórmulas	Afijos	Formas compuestas
		Acrónimos	*Shortcuts*	
	Género:	Masculino	Femenino	Neutro
	Número	Singular	Plural	

2.2.1.2. Comentarios a la propuesta

Un análisis pormenorizado de la propuesta de codificación de la información morfosintáctica de Eagles, nos lleva a formular los siguientes comentarios:

1. La necesidad de disponer de estándares de codificación de la información lingüística está fuera de toda duda, ya que ello implica la posibilidad de comparación objetiva de sistemas de anotación así como la reutilización de los recursos existentes.

2. Disponer de un sistema de codificación no implica necesaria ni automáticamente disponer de un sistema de clasificación de las palabras o, dicho de otra forma, los sistemas de codificación no son en sí mismos sistemas de clasificación de las palabras porque no proporcionan más que un método para codificar la información y no un método para asignar categorías a las palabras de la lengua.

3. La propuesta Eagles no proporciona criterios lingüísticos de discriminación categorial ni de adscripción de las palabras a las distintas categorías, simplemente propone un sistema para codificar información.

4. Adaptar esta propuesta a una lengua concreta no puede hacerse de un modo directo, puesto que, si bien se trata de una propuesta de sistema de anotación orientado a estandarizar y compatibilizar los recursos existentes, deja abiertos y sin resolver

aspectos como la doble categorización o la redundancia en el tratamiento de categorías y subcategorías, lo que, además, puede dificultar la tarea de traducción y/o comparación entre aplicaciones.

2.2.2. Adaptación al español

En lo que sigue presentamos nuestra propuesta de adaptación del etiquetado Eagles al español[12].

El primer paso para adaptar la propuesta Eagles al español ha sido establecer las categorías o las clases de palabras (o las partes de la oración o las categorías sintácticas) que se van a utilizar; el segundo, adscribir las palabras de la lengua a las categorías establecidas, con independencia de que una palabra reciba más de una etiqueta.

2.2.2.1. Descripción general

La adaptación del sistema Eagles a nuestro entorno ha implicado en primer lugar, definir el conjunto de categorías de entre las diferentes opciones; en segundo lugar, seleccionar los atributos pertinentes para cada categoría; y finalmente construir el diccionario basándonos en nuestra propuesta.

Los principales objetivos de nuestro sistema de codificación son los siguientes:

1. Utilizar etiquetas que codifiquen la información morfológica explícita de las palabras.

2. Establecer como principio general que los dos primeros atributos para cada clase de palabra sean **categoría** y **tipo**, de modo que todas reciban el mismo tratamiento[13].

3. Establecer que el orden general de los atributos para las distintas clases de palabras sea siempre el mismo: categoría, tipo(s), accidentes gramaticales. No ocurre lo mismo a nivel específico (i.e. el atributo de género, por ejemplo, no siempre ocupa el mismo lugar en todas las etiquetas). El orden que hemos seguido se ajusta de la propuesta Eagles.

4. Utilizar dígitos alfanuméricos con una posición fija. Cada categoría morfológica establecida tiene un número concreto de atributos, lo que implica un número igual de dígitos en un orden determinado. El número de atributos no tiene porqué coincidir entre las distintas categorías.

5. Considerar la posibilidad de inespecificar algún atributo, es decir, de no dar su información relativa, y utilizar para ello el valor '0', de modo que se no altere el esquema general.

   Esta situación se producirá fundamentalmente en dos casos:
   (i) cuando se trate de un atributo que dependa directamente del valor que tome un

[12] Esta propuesta se ha aplicado también al catalán (Arévalo, Taulé, y Martí, 2001) y al inglés con las adaptaciones correspondientes.

[13] Este principio se ha respetado para todas las categorías mayores, y sólo las cifras, interjecciones, fechas y abreviaturas, tienen un único dígito.

atributo anterior, como el atributo de *caso* en el pronombre, que sólo es aplicable al pronombre de tipo personal, y que, por tanto, quedará inespecificado en los casos restantes.
(ii) si los atributos establecidos todavía no se utilizan en esta versión del analizador, pero está previsto introducirlos más adelante. De este modo, se deja abierta la posibilidad de refinar el sistema de anotación propuesto[14].

De este punto se desprende que no todos los atributos tienen el mismo dominio de aplicación: mientras unos se aplican a toda la categoría con independencia del tipo, otros se aplican a un tipo concreto[15] y, por tanto, aparecerán con valor '0' en los demás casos.

6. Utilizar dos tipos de valores: las listas y los booleanos[16].

7. Para aquellos casos en que no es posible establecer diferencias a nivel formal entre los valores que debe tomar un atributo, establecer un supervalor que incluya a los demás. Por ejemplo, en los valores del atributo de género en el caso del adjetivo y del nombre seguimos la solución adoptada por Eagles, que puede esquematizarse tal como aparece en la figura 2.5

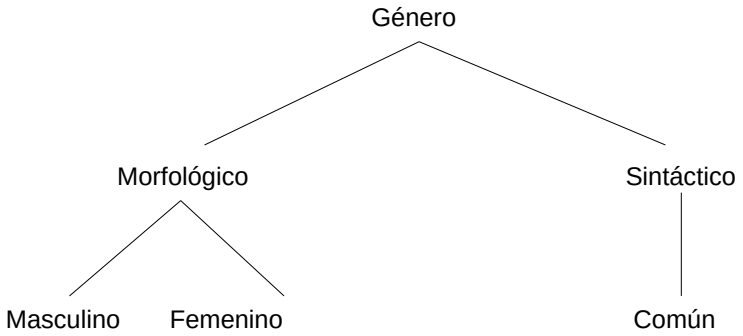

Figura 2.5: Los valores del atributo de género

Esta división no significa que haya un tercer género, en distribución con el masculino y el femenino, sino que indica que el género no puede determinarse a partir de la forma de la palabra, pues se requiere información contextual. Dicho de otro modo, en el caso concreto de los adjetivos, se trata de aquellas formas invariables que pueden acompañar a sustantivos tanto masculinos como femeninos (como por ejemplo, to-

[14]Esto ocurre por ejemplo en el caso de los nombres. Ya se ha realizado una clasificación de los nombres propios y se van a utilizar los dígitos quinto y sexto para especificar esta información (Arévalo, 2001); por otra parte, el último dígito de las etiquetas para los nombres también se utilizará en el futuro para marcar los apreciativos.

[15]El atributo de *género* en el verbo sólo se aplica a las formas participias; el de *politeness* en el pronombres, sólo a los pronombres personales; etc.

[16]Las listas contienen dos o más elementos; los booleanos contienen un solo valor que aparece o no aparece en un caso concreto.

dos los acabados en -e)[17]. Otra forma de tratar estos fenómenos sería la disyunción de valores. Por ejemplo, señalar el género de *comunista* con dos valores *M/F*. Sin embargo, por un criterio de simplicidad en la anotación hemos optado por codificar estos fenómenos adoptando la propuesta Eagles.

8. Tal como se deduce del punto anterior, no utilizar la disyunción de valores en el seno de un atributo.

2.2.2.2. Categorías gramaticales establecidas

No es siempre fácil saber si un determinado comportamiento gramatical corresponde a una clase de palabras o a una subclase de otra categoría. En gran parte depende de nuestra decisión [...] elegir entre postular que dos unidades con distinto funcionamiento pertenecen a la misma clase pero a distinta subclase, o bien entender, por el contrario, que pertenecen a clases distintas. (Bosque (1991): p. 25.)

No conviene dar a las clasificaciones más alcance que el de criterios con que tratamos de analizar e interpretar la realidad viva del idioma. (RAE (1973): p. 463.)

Ignacio Bosque (Bosque, 1991) habla de cuatro clasificaciones (binarias) que se han utilizado tradicionalmente para categorizar las palabras:

(i) categorías variables –vs– categorías invariables
(ii) clases abiertas –vs– clases cerradas
(iii) categorías llenas –vs– categorías vacías
(iv) categorías mayores –vs– categorías menores.

La primera clasificación se basa en el hecho de que ciertos grupos de palabras comparten determinadas marcas morfológicas.

La segunda, se basa en el número de elementos que puede tener cada clase. La lista de las palabras que pueden ser, por ejemplo, preposiciones es finita y no hay libertad para crear nuevas palabras que pertenezcan a esta clase. Las palabras que pertenecen a clases cerradas forman parte del conocimiento lingüístico de todos los hablantes de una lengua[18], mientras que las que pertenecen a clases abiertas, no[19].

La distinción entre categorías llenas y vacías es de naturaleza semántica y consiste en afirmar que mientras las palabras que pertenecen a la primera clase poseen significado léxico, las que pertenecen a la segunda no, por lo que no pueden definirse. Sin embargo, conviene no confundir aquí el no expresar significado léxico con el no expresar significado. En

[17] Sobre estos valores, se comenta en (EAGLES, 1996c): *Ontologically, they have a quite different status from 'm' or 'f'. [It] can be seen as a metalinguistic device, or as a third value: it can be seen as a sort of a multilabel whose semantics is a 'disjunction' operation.*

[18] Según el autor, todas excepto, quizá, *cuyo* y *sendos*.

[19] Sobre los numerales, hay que decir que casi por definición son una clase abierta, o infinita, puesto que la serie de los números naturales es infinita. Sin embargo, creemos, a diferencia de lo que postula el autor, que, a efectos de esta clasificación pueden considerarse como pertenecientes a la clase cerrada, puesto que son perfectamente identificables y también forman parte, pese a algunas vacilaciones, del conocimiento de todos los hablantes. Además, los numerales no son productivos, no pueden crearse nuevas palabras pertenecientes a esta categoría.

efecto, las palabras pertenecientes a las categorías vacías expresan relaciones gramaticales que, si bien no tienen en sí significado léxico, sí lo tienen gramatical.

Por último, la cuarta clasificación establece como pertenecientes a las clases mayores aquellas palabras capaces de tener complementos, es decir, de ser núcleo, lo cual plantea contradicciones con algunas teorías que consideran que las preposiciones no son categorías mayores o núcleos.

Entendemos por categorías gramaticales lo que J.C. Moreno Cabrera llama *partes del discurso* ((Moreno, 2000): cap. 19, pp. 403 y ss.) y que define del siguiente modo:

> *Denominamos* partes del discurso *a aquellas clases formales de unidades de significado que comparten una serie de propiedades morfológicas, sintácticas y semánticas en que pueden agruparse las palabras de una lengua natural.*

La lengua se adapta mal a las clasificaciones, por lo que en algunos casos hay vacilaciones en la categorización de las palabras. Sin embargo, puesto que lo que estamos estableciendo es un sistema de etiquetación (es decir, de clasificación), hemos de tomar decisiones sobre la correspondencia que establecemos entre las palabras de la lengua y las distintas categorías gramaticales. Aunque se intenta establecer claramente cuáles son los criterios de adscripción de las palabras a las categorías, esta adscripción no siempre es clara. Hay casos dudosos o difusos, como se verá en el resto del capítulo y en el siguiente.

> *Hay [...] una regla irrecusable, como dictada por la razón, y es que los varios miembros de la clasificación no se comprendan unos a otros.* (Bello (1847): Nota I.)

La clasificación que proponemos se basa en criterios tanto morfológicos como sintácticos. Los primeros, utilizados aisladamente, no son suficientes; sin embargo, combinados con los sintácticos sí proporcionan una clasificación adecuada de las palabras en español.

Las objeciones que presenta Ignacio Bosque a los criterios de clasificación basados exclusivamente en la morfología son las siguientes ((Bosque, 1991): pp. 31–32):

1. *no distingue específicamente entre las propiedades flexivas que se asocian sistemáticamente con una categoría [...] y aquellas otras que se caracterizan porque sólo algunos de sus miembros poseen la marca en cuestión*, como por ejemplo, *cada* que es invariable;

2. hay categorías que tienen flexión asignada léxicamente (el nombre) y otras que la reciben por concordancia (adjetivo y verbo);

3. por último, es posible que *determinado contenido pued[a] estar presente morfológicamente sin que se trate de una marca flexiva*, como el contenido temporal del prefijo *ex-* en *exembajador* .

Como respuesta general a las críticas comentadas en el párrafo anterior, cabe decir que pueden hallar solución con la sintaxis. A nivel más detallado, se pueden dar las siguientes respuestas:

1. Cuando el criterio basado estrictamente en la forma de la palabra es insuficiente para asignar una palabra a una clase, es posible utilizar criterios paradigmáticos, y *cada*, pese a ser efectivamente invariable, está en distribución con otras palabras variables que pertenecen a la categoría de determinante.

2. Sobre la segunda, cabe señalar que el hecho de que en algunos casos la flexión sea asignada léxicamente y en otros no, nos ha servido para subclasificar las palabras: nombre y adjetivo poseen el morfema de género, pero mientras en el primer caso éste es asignado léxicamente, en el segundo se manifiesta por la concordancia con el nombre, igual que ocurre con los determinantes. Lo que hemos considerado importante es que tuvieran flexión, fuera ésta asignada léxicamente o no.

3. Por último, y como respuesta a la tercera objeción, podemos señalar que sólo hemos utilizado, en la definición de variación morfológica, los que tradicionalmente se han considerado morfemas flexivos de las distintas clases de palabras que los poseen, a saber, género, número y tiempo; es cierto que hay palabras o morfemas que pueden marcar tiempo, pero no es cierto que el tiempo sea un morfema flexivo de todas las clases de palabras (sólo lo es en el verbo).

A pesar de todo ello, el principal argumento a favor de la clasificación morfológica de las palabras lo ofrece el propio I. Bosque: ((Bosque, 1991): p. 29), *las marcas morfológicas casi nunca son opcionales y constituyen rasgos formales siempre relevantes.*

En nuestro caso, el primer criterio utilizado es el morfológico: las clases de palabras poseen determinadas marcas flexivas o, por definición, carecen de ellas. Esto ya permite establecer dos grandes grupos de palabras, tal como muestra la figura 2.6:

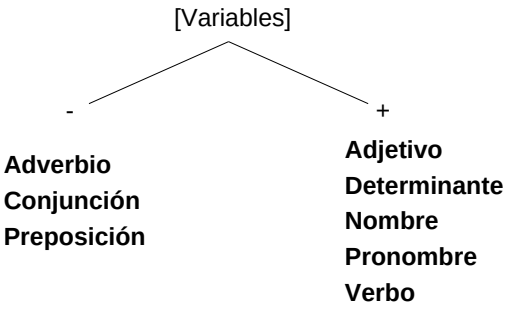

Figura 2.6: Clasificación de las palabras (1a)

Dentro de las palabras variables, todavía podemos realizar una segunda subclasificación atendiendo a los morfemas que presentan: el verbo presenta el morfema de tiempo, mientras que las restantes no. Entre éstas, el nombre tiene género y número inherente y las demás concordante. Esta división es la que aparece en, la figura 2.7:

La distinción entre adjetivo, determinante y pronombre, la establecemos en dos pasos. En primer lugar *adjetivo –vs– determinante/pronombre*. Los adjetivos pertenecen a una clase abierta de palabras, mientras que las otras dos no; y los determinantes (como

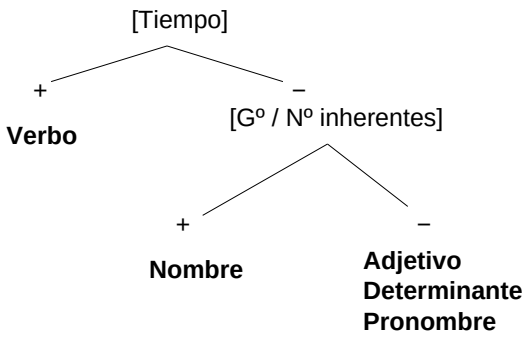

Figura 2.7: Clasificación de las palabras (1b)

los adjetivos) son adyacentes del nombre (aunque no pueden ser predicados (Demonte, 1999): pp. 136–137), mientras que los pronombres no.

Por su parte, preposiciones, conjunciones y adverbios pueden distinguirse utilizando criterios sintácticos[20]. Proponemos por tanto, utilizar en primer lugar el criterio de [± categoría mayor] que separa las conjunciones del resto; la diferencia entre la preposición y el adverbio viene dada por la transitividad, rasgo que poseen las preposiciones pero no los adverbios.

Todo lo comentado hasta aquí queda reflejado en la figura 2.8.

Establecidas ya ocho categorías de palabras, recogidas por la mayoría de las gramáticas y contempladas por la mayoría de teorías lingüísticas existentes, debemos hacer notar que al etiquetar corpus aparecen *elementos* que no pueden incluirse en ninguna de las categorías anteriores, tal como puede observarse en los ejemplos de 2.4, de 2.5 y de 2.6. En los ejemplos de 2.4 pueden observarse interjecciones, propias del lenguaje coloquial.

(2.4) (a) *... hoy que Epi,* ¡Ay!*, sólo puede ser un recurso táctico* (d1).
(b) *y todo vuelve a empezar, en ese ciclo ininterrumpido de la "era Gil".* Adiós*, Copa de la UEFA.* Adiós*, Copa del Rey. Hasta el año que viene* (d1).

En los siguientes (2.5), aparecen cifras con distintos valores. En (a) se sobreentiende

[20]La distinción entre estas tres clases de palabras también puede hacerse por criterios más funcionales, tal como aparece en Pavón (1999): pp. 567–568. que luego retomaremos nosotros:
la **preposición** *es una clase de palabras encargada de establecer una relación de modificación o subordinación entre dos constituyentes*;
el **adverbio** *se suele definir como la clase de palabras que modifica al verbo (o a la oración), al adjetivo o a otros adverbios [...] si bien existen ciertos adverbios (los de la clase de* incluso, casi*, etc. que pueden modificar prácticamente a cualquier tipo de categoría gramatical)*;
las **conjunciones** *constituyen una clase de palabras cuya misión es relacionar oraciones o elementos de una oración. El tipo de relación que establecen puede ser de coordinación ('conjunciones coordinantes') o de subordinación ('conjunciones subordinantes'). En el primer caso, las conjunciones pueden enlazar diferentes tipos de elementos [...]; en el segundo caso, el segundo término de la relación ha de ser obligatoriamente una oración.*

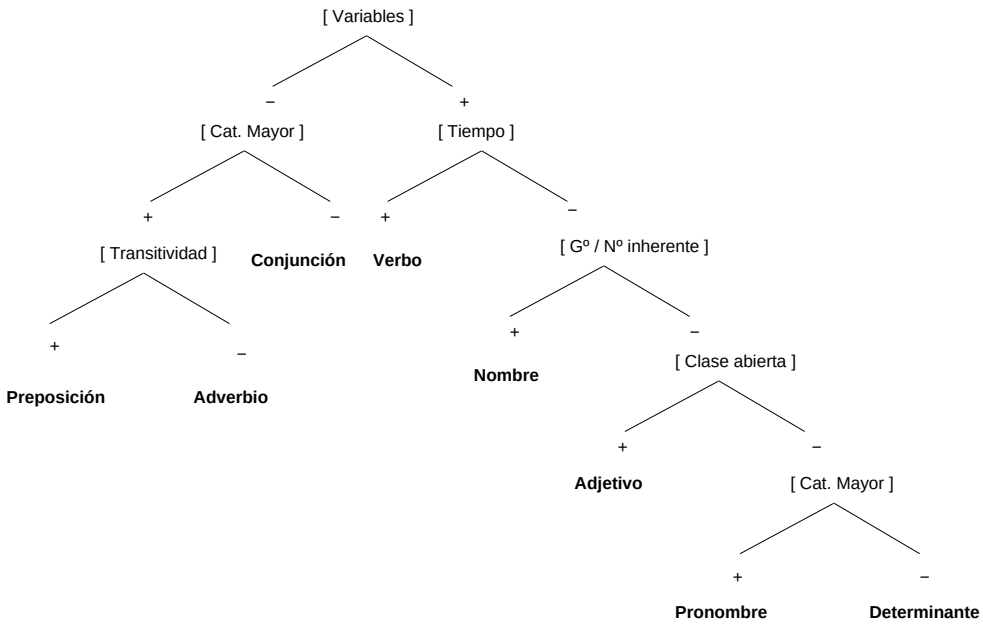

Figura 2.8: Clasificación de las palabras (1)

el jugador que llevaba el número once // el jugador que llevaba el número siete; en (b) tenemos una cantidad expresada en cifras; en (c) se referencia una fecha; en (d) aparece un porcentaje y en (e) la altura en centímetros de una persona.

(2.5) (a) *Un pase del* 11 *provocó el gol del* 7 (d1).
(b) 206 *centímetros de potencia y agilidad* (d1).
(c) *Saltaron juntos a la fama en* 1992 (d1).
(d) *Pero sigue sin meter más del* 40 % *de sus tiros* (d1).
(e) *... y pese a su talla modesta de* 2,05 *(pero con una envergadura de casi* 2,20*), está resolviendo...* (d1).

En los ejemplos de 2.6 aparecen diversos signos de puntuación que no desempeñan su función habitual: en (a) y (b) aparecen puntos que forman parte de nombres propios; en (c) y (d) el punto aparece entre caracteres numéricos, en el primer caso para señalar una hora, y en el segundo para marcar los millares; en (e) aparecen tres puntos que deben interpretarse como un único elemento con una función lingüística determinada; por último, en (f) el punto forma parte de una abreviatura.

(2.6) (a) *por eso un día* Will D. Howe *le invitó a colaborar en un libro de filología* (a1).
(b) *En* 1976 *apareció en Filadelfia* (EE. UU.) *una bacteria especial* (dc2).
(c) *Esa joven lleva a estas horas, las* 20.13*, una cesta de la compra* (a11).
(d) *Sólo en Villa El Salvador, el suburbio de la Madre Coraje, hay* 300.000 *habitantes* (a14).

(e) *"Creérselo todo requiere una aplicación constante, y una especie de genio, y creérselo todo... a medida que pase el tiempo probablemente será cada vez más difícil"* (a15).
(f) *Con espesores controlables desde* 1 *a varios* mm, *permiten lograr fácilmente el "dopado" del polímero* (fig. 1) (dc3).

Como resultado del análisis, hemos establecido pues, y siguiendo la propuesta de trabajos precedentes (como por ejemplo, EAGLES (1996c), Sampson (1995) y Santorini (1990)), otras categorías: Interjección, Abreviatura, Fecha, Moneda y Números. Finalmente, hay que señalar que los signos de puntuación se tratan también como una categoría más de palabras, porque son importantes a la hora de delimitar estructuras.

Hay por tanto catorce categorías, a saber: **adjetivo, adverbio, determinante, nombre, verbo, pronombre, conjunción, preposición, interjección, abreviatura, cifra, fecha, moneda** y **puntuación**.

En las secciones siguientes se definirán de modo más preciso todas y cada una de las clases o categorías establecidas así como sus marcas formales y/o su función. También se señalarán los criterios seguidos para adscribir las palabras a las clases.

Lo que ha prevalecido en nuestro análisis ha sido que los criterios de clasificación sean lo más claros posible y que se reduzca al máximo la ambigüedad en la asignación de categorías a las palabras.

Se dan muchos casos de lo que Leech y Wilson llaman *homonimia gramatical*[21], consistente en que una misma forma de palabra reciba más de una etiqueta. Sin embargo, este tipo de ambigüedad es inherente a una clasificación morfológica de las palabras. En algunos casos estamos ante un fenómeno de homonimia (forma verbal o pronominal de *una*); en otros estamos ante diferentes usos de las palabras, lo que Ignacio Bosque llama *duplicación de categorías* ((Bosque, 1991): pp.48–52), fenómeno que se produce especialmente entre los cuantificadores y las palabras con capacidad anafórica.

A continuación se presenta un análisis detallado de cada una de las categorías establecidas según los tratados gramaticales más representativos[22] para concluir con nuestra propuesta de anotación léxica para el español.

[21] EAGLES (1996b): p.17.

[22] Fundamentalmente, *Gramática de la Lengua Castellana* de Andrés Bello (Bello, 1847); *Esbozo de una nueva gramática de la lengua española*, de la RAE (RAE, 1973); *Gramática española*, de Alcina-Blecua (Alcina y Blecua, 1989); *Las categorías gramaticales*, de Ignacio Bosque (Bosque, 1991); *Gramática de la lengua española*, de Emilio Alarcos (Alarcos, 1994), y *Gramática Descriptiva de la Lengua Española*, de I. Bosque y V. Demonte (Eds.) (Bosque y Demonte, 1999). De modo más ocasional, también hemos utilizado *Curso Universitario de Lingüística General*, de J.C. Moreno Cabrera (Moreno, 2000) y *Sintaxis y Cognición. Introducción al conocimiento, el procesamiento y los déficits sintácticos* de M. Fernández Lagunilla y A. Anula Rebollo (Fernández y Anula, 1995).

2.2.3. Adjetivo.

2.2.3.1. Definición, clasificación y propiedades

Los rasgos morfológicos del adjetivo no se interpretan semánticamente, sólo son rasgos concordantes (Bosque, 1991).

El adjetivo no siempre ha sido tratado, como se verá a continuación, como una categoría independiente. Para algunos autores forma parte, junto con el sustantivo, de una superclase: el nombre. En la lingüística moderna, por lo general se separa del nombre aunque se reconocen sus puntos de contacto.

Para Bello (1847), adjetivo y sustantivo aparecen como clases distintas de palabras aunque admita la denominación común de *nombre* para ambas:

> *me ha parecido conveniente dar la denominación común de* Nombres *al sustantivo y al adjetivo, por la semejanza de sus accidentes y la frecuente transformación de uno en otro, sin que por esto, cuando enumero las más altas categorías en que se dividen las palabras, considere al Nombre como una de ellas, puesto que el Sustantivo y el Adjetivo ofrecen caracteres especiales, exclusivos e importantísimos, que diferencian al uno del otro y de todas las clases de palabras*[23].

Para este autor, el adjetivo es una de las dos clases de palabras (la otra es el verbo) que modifica al sustantivo. Los rasgos morfológicos que presenta el adjetivo son los de género y los de número.

Por su parte, tanto el *Esbozo*, como la *Gramática* de Alcina–Blecua tratan adjetivo y sustantivo como dos subclases de la categoría gramatical *nombre*. Los motivos de esta agrupación son que comparten los mismos morfemas, a saber, género y número, además de poder combinarse los dos con bastantes sufijos derivativos y de intercambiar sus funciones con facilidad. La diferencia importante que se señala entre ellos es que sólo los nombres sustantivos pueden clasificarse en apelativos y propios[24]. Desde un punto de vista funcional, la indistinción o sincretismo entre ambas categorías es frecuente. Por lo general sólo la construcción en que aparecen decide la categoría y en frases como *son españoles* es imposible, se dice[25], decidir si *españoles* es nombre sustantivo o nombre adjetivo. Por otra parte, muchos adjetivos aparecen tratados como sustantivos, en especial en plural: *los mejores*. Y los sustantivos pueden desempeñar la función de atributo: *vida padre, ciudad satélite* (a veces con inmovilización del morfema de número, a veces con acomodación de género: *día perro, vida perra*)[26].

Lo que separa ambas clases de nombres es que sólo los adjetivos pueden combinarse con la forma neutra del artículo *lo*, que sólo los adjetivos (aunque no todos) pueden combinarse con adverbios de grado y de modo[27], y que mientras que los sustantivos tienen una función

[23] Bello (1847): nota I.
[24] RAE (1973): p. 171-172.
[25] RAE (1973): p. 172.
[26] RAE (1973): p. 172 y 190.
[27] RAE (1973): p. 190-191.

primaria, la de los adjetivos es secundaria[28].

En cambio, Emilio Alarcos, en su *Gramática de la lengua española*, distingue claramente los sustantivos de los adjetivos y define adjetivo como la palabra que funciona como adyacente del sustantivo y que, aislado, puede cumplir por sí solo la función de atributo[29]. Según este autor,

> *hay dos tipos de adjetivos, separados funcionalmente por su diversa posibilidad de ordenación entre sí y respecto del sustantivo al que acompañan:*
> *1. los que admiten cualquier posición respecto del núcleo sustantivo del grupo unitario y respecto del otro adyacente (con el cual pueden coordinarse o yuxtaponerse). Se corresponden con los calificativos [...].*
> *2. Los que en presencia de otro adjetivo en el mismo grupo unitario exigen estar antepuestos y nunca inmediatamente pospuestos a aquel. Vienen a coincidir con los denominados determinativos.*[30]

En general, todos los autores mencionan también como característica del adjetivo la gradación, aunque los tipos propuestos difieren entre los autores. Sin embargo, como señala (Bosque (1991): pp. 122 y ss.), no todos los adjetivos admiten la gradación (adjetivos relacionales) y, además, ésta no es exclusiva del adjetivo. Señala este autor que *La gradación [...] no es en sí misma un criterio de identificación categorial*[31].

En la línea de separar claramente ambas categorías (nombre y adjetivo), se sitúa la lingüística moderna (Bosque (1991), Bosque (1999) y Demonte (1999)). No obstante, en las tres obras citadas se reconoce la proximidad entre ambas clases. En Bosque (1991) se señala como diferencia clara que sólo los sustantivos pueden aparecer *en sintagmas a los que corresponden funciones oracionales básicas, como sujeto o complemento directo*[32]. Sin embargo, para este autor, la diferenciación más clara entre nombre y adjetivo la marca la semántica de ambas clases de palabras:

> *los sustantivos 'categorizan', esto es, determinan* clases de objetos*, mientras que los adjetivos 'describen' propiedades que no constituyen clases* (Bosque (1991): p. 107.).

En (Demonte (1999): p. 133.) también se recoge esta interpretación:

> *El adjetivo es una categoría gramatical: una clase de palabras cuyos miembros tienen unas características formales muy precisas; y es también una categoría semántica: hay un tipo de significado que se expresa preferentemente por medio de adjetivos. Como categoría gramatical puede ser un atributo o modificador del nombre sustantivo; unido a él, y a sus determinantes y cuantificadores, forma una frase nominal en la cual ha de concordar en género y número con el nombre modificado.*

[28] Alcina y Blecua (1989): pp. 498-499.
[29] Alarcos (1994): p. 78.
[30] Alarcos (1994): p. 84.
[31] Bosque (1991): p. 123.
[32] Y en esto no se diferencia mucho de lo que propone E. Alarcos.

Siguiendo a I. Bosque (Bosque (1999): pp. 60–61) podemos definir la categoría adjetivo atendiendo a las características semánticas (2) y morfosintácticas (1, 3, 4) siguientes:

1. Los rasgos morfológicos del adjetivo no se interpretan semánticamente, sino que sólo son rasgos concordantes;

2. La interpretación de los sintagmas nominales formados por adjetivos es anafórica (en contraposición con lo que ocurre si están formados por nombres: interpretación no anafórica);

3. Sólo los adjetivos pueden recibir el artículo **lo** (mientras que sólo los sustantivos pueden recibir el indefinido **un**);

4. A nivel sintáctico, sólo los adjetivos pueden ser predicados de construcciones absolutas y sólo los adjetivos pueden aparecer como complemento predicativo no seleccionado (en realidad, la mayoría de nombres no puede aparecer en este tipo de construcciones).

Merece especial atención el tratamiento que se hace en todas estas obras de los numerales. Para Bello son nombres[33]; para el *Esbozo* comparten propiedades con las clases del nombre y del pronombre[34]; Alcina-Blecua los incluyen en la categoría de pronombres determinativos[35]. Por su parte, Emilio Alarcos sitúa los cardinales entre los adjetivos de tipo segundo (es decir, los determinativos) y los restantes entre los calificativos[36] debido a su comportamiento sintáctico.

En (Marcos (1999): pp. 1202 y ss.), los cardinales se asimilan a los determinantes por el orden de colocación respecto del nombre (el orden no marcado en español es nombre–adjetivo, y los cardinales suelen preceder al sustantivo). Los ordinales se consideran más próximos a los calificativos por los siguientes motivos[37]:
(i) pueden anteponerse o posponerse al nombre (*el primer libro – el libro primero*);
(ii) refieren anafóricamente al nombre: *se probó un montón de trajes pero al final se quedó con el primero*; y,
(iii) admiten anteposición de la forma *lo*, como los calificativos.

Una observación interesante que hace este autor (p. 1205) y que también se retoma en Martínez (1999) es que en la lengua actual se está dando un uso de los cardinales como ordinales. Los cardinales utilizados como tales acompañan a nombres en plural (*las diez páginas*), mientras que cuando se utilizan como ordinales el nombre aparece en singular (*la página diez*)[38].

[33] Bello (1847): § 188.
[34] RAE (1973): p. 237.
[35] Alcina y Blecua (1989): p. 594 y pp. 663. y ss.
[36] Alarcos (1994): p. 120.
[37] Los ejemplos que se citan son del autor
[38] Aunque ninguno de los autores lo señala, es de destacar también la diferente posición del numeral respecto del nombre en ambos casos.

En Martínez (1999) los numerales cardinales se tratan como cuantificadores y se señala que por su función son adjetivos. De los ordinales se dice que son adjetivos (no cuantificadores) cercanos a los calificativos por su función, *pero estrechamente emparentados en lo léxico con los cardinales.*

Fernández y Anula (1995), siguiendo a Eguren (1989)[39] sitúan entre los determinantes sólo a los numerales cardinales, mientras que los ordinales quedan fuera de esta categoría.

Otro ejemplo de la proximidad entre ordinales y calificativos es que ambas clases de palabras pueden coordinarse[40], cosa que no ocurre entre determinantes y ordinales ni entre determinantes y calificativos, tal como se muestra en los ejemplos 2.7.

(2.7) (a) *Una* **segunda e importante** *característica de la conductividad del complejo...* (dc3).
(b) *la* **primera, y más simple**, *solución en este sentido fue la adición a un polimero...* (dc3).

En cuanto a las propiedades gramaticales de los adjetivos (ya sean calificativos, ya ordinales), los autores mencionan como morfemas característicos (que no necesariamente exclusivos) los de género y número. Estos morfemas, que también presenta el sustantivo, tienen aquí un valor distinto: *los morfemas del adjetivo no añaden ninguna información nueva [...] son meros índices funcionales de la relación que el adjetivo contrae con el sustantivo cuando este no los manifiesta explícitamente* (Alarcos (1994): § 98).

En el *Esbozo* los adjetivos calificativos se clasifican, atendiendo al género, en tres grupos:

1. genéricamente invariables,
2. femenino en -a, masculino en -o,
3. femenino en -a, masculino que no es -o

En este sentido, los ordinales formarían parte del segundo grupo, puesto que todos presentan la alternancia *o/a* para el género.

En cuanto al número, las formas son singular–plural y se dan muy pocos casos de neutralización (o morfema cero) de este morfema.

2.2.3.2. Criterios y etiquetación adoptados

Como hemos visto, existen suficientes características definitorias para considerar el adjetivo como una categoría independiente. Resumimos a continuación estas características de acuerdo con todo lo comentado en 2.2.2.2, haciendo especial hincapié en los aspectos formales o morfosintácticos y dejando de lado deliberadamente sus características semánticas:

1. admite la anteposición del artículo neutro *lo*;
2. tiene función secundaria;
3. es una palabra morfológicamente variable;

[39] L. Eguren "Algunos datos del español en favor de la hipótesis de la Frase determinante" en *Revista Argentina de Lingüística*, 5, 1/2, 163–203.

[40] Aunque parece que la coordinación sólo es posible si ambos adjetivos preceden al nombre, seguramente porque el ordinal exige esta posición.

4. presenta rasgos morfológicos concordantes;

5. aunque no siempre, admiten ir antepuestos o pospuestos al nombre.

Estos rasgos definen una categoría que incluye dos subclases de palabras: los calificativos y los ordinales. La única diferencia entre ambos es que los calificativos, como señala Bosque (1991) admiten ser predicados de construcciones absolutas, mientras que los ordinales no pueden aparecer en dicha construcción.

Otro hecho que nos ha llevado a considerar como miembros de la misma clase calificativos[41] y ordinales es el hecho de que pueden coordinarse, cosa que no sucede entre otras clases de adjetivos (tal como ya se vio anteriormente).

Por lo tanto, y según todo lo anteriormente comentado, proponemos tratar de modo diferenciado la **categoría** adjetivo respecto de la de sustantivo, y establecer dos **tipos**, el calificativo y el ordinal.

La etiquetación que adoptamos para el adjetivo queda recogida en el cuadro 2.1.

Atributo	**Valor**	**Código**
Categoría	Adjetivo	A
Tipo	Calificativo	Q
	Ordinal	O
Apreciativo	Sí	A
Género	Masculino	M
	Femenino	F
	Común	C
Número	Singular	S
	Plural	P
	Invariable	N
Participio	Sí	P

Cuadro 2.1: Etiquetas para los adjetivos

En la propuesta Eagles, el tercer atributo para los adjetivos es el de grado. En español, la gradación es un procedimiento más sintáctico que morfológico, ya que a excepción del superlativo absoluto, que se forma con el sufijo *-ísimo*, los grados comparativo y superlativo relativo se expresan sintácticamente, con adverbios y conjunciones. Sólo unos pocos adjetivos tienen formas especiales para el comparativo de superioridad y para el superlativo absoluto[42]. De entre los ordinales, sólo *primero* y *último* admiten la derivación superlativa: *primerísimo – ultimísimo* pero no con su significado ordinal. Dado pues que el grado no es un procedimiento esencialmente morfológico, hemos optado por no marcarlo en el adjetivo, y por reservar esta posición de la etiqueta para marcar, en un futuro, los **apreciativos**

[41] Entendemos por calificativos tanto los adjetivos calificativos propiamente dichos como los llamados relacionales (cf. Bosque (1991): cap. 5).

[42] Algunos ejemplos son:
bueno–mejor–óptimo / malo–peor–pésimo / grande–mayor–máximo / pequeño–menor–mínimo

(diminutivos, aumentativos y despectivos), mucho más productivos en español. En la actualidad, este dígito aparece siempre con valor *0*, aunque más adelante se tratará este caso particular de derivación en el módulo de sufijos del analizador.

Ambas clases de adjetivos presentan variación de género y número. En cuanto al **género**, el adjetivo presenta formas masculinas y femeninas. Pero esta diferenciación no siempre aparece de forma explícita. Se da en los grupos 2 y 3 que propone el *Esbozo* y en todos los ordinales, pero no en el grupo 1. En este último caso, el atributo de género tendrá el valor *común* que, como ya se ha indicado en la página 57, significa que es imposible determinar desde un punto de vista estrictamente morfológico el género de la palabra; sólo el contexto permite la asignación de género masculino o femenino.

Algunos ejemplos de adjetivos (calificativos) con indeterminación del morfema de género son los siguientes:

	azul	*civil*	*eficaz*
abominable	*azulgrana*	*comunista*	*ejemplar*
aborigen	*belga*	*conveniente*	*espiritual*
actual	*breve*	*difícil*	*estéril*
antisindical	*caliente*	*dominante*	*falaz*
armamentista	*capaz*	*dulce*	*feliz*

La variación que afecta al **número** es la alternancia singular–plural. De modo semejante a lo que ocurre en el tratamiento del morfema de género, se dan casos (aunque muy pocos y que afectan fundamentalmente a palabras de procedencia extranjera) en que es imposible determinar a nivel puramente morfológico el número de un adjetivo. Para estos casos está previsto el valor *invariable*: *grunge, kitsch, seudo, pop, unisex, isósceles, rubiales.* Aquí también, la interpretación como singulares o plurales se infiere del contexto en que las palabras aparecen.

Un caso particular de adjetivos (calificativos) lo forman los **participios verbales** cuando no se utilizan ni en los tiempos compuestos de la conjugación verbal, ni en la voz pasiva, ni en las llamadas construcciones absolutas. En tanto que formas variables, con género y número dependientes del nombre al que complementan o refieren, su comportamiento morfosintáctico se asimila al del adjetivo: *Los participios atribuidos a un sustantivo desempeñan una función adjetiva igual a la de cualquier adjetivo complementario de un nombre.*[43] De modo similar se expresa E. Alarcos, cuando dice que

> *los llamados* infinitivo, gerundio *y* participio *[son] considerados, no sin razón, como* formas nominales del verbo. *En realidad, son unidades derivadas del signo léxico de los verbos y que funcionan, respectivamente, en los papeles de los sustantivos, de los adverbios y de los adjetivos. Sin embargo, tales unidades derivadas conservan en parte las posibilidades combinatorias admitidas por el signo léxico verbal. Es decir, [...] son susceptibles de llevar adyacentes análogos a los que el verbo recibe en la oración*[44].

[43] RAE (1973): p. 496.
[44] Alarcos (1994): pp. 142-143.

La alternancia entre participios y adjetivos puede observarse en los siguientes ejemplos[45]:

La niña viene cansada	$\sim$	*La niña viene contenta*
Me quedé aturdido	$\sim$	*Me quedé cabizbajo*
El árbol caído	$\sim$	*El árbol viejo*

Adjetivo y participio aparecen también en las llamadas *construcciones absolutas*[46]:

(2.8) *Oídos los reos, el juez dispuso...*
Limpias las armas, ...

En ocasiones, la misma forma de palabra se interpreta sólo como adjetivo (ejemplos 2.9 (a)) o sólo como verbo (ejemplos 2.9 (b)), en función del contexto, tal como lo demuestran los siguientes ejemplos[47]:

(2.9) *hombre resuelto, mujer ocupada*
problema resuelto, territorio ocupado

En nuestro sistema de etiquetación, los participios aparecen con una doble etiqueta: por un lado la de verbo–participio (cf. sección 2.2.7.2) y, por otro, la de adjetivo calificativo. De este modo podemos dar cuenta de la "doble naturaleza" de estas formas. Hay que destacar, sin embargo, que sólo consideramos participio la forma que aparece en los tiempos verbales y en las construcciones de participio absoluto. Sin embargo, puesto que los adjetivos de origen verbal pueden mantener los complementos subcategorizados por el verbo, hemos incluido un atributo más, de tipo booleano, llamado **Participio** que, con el valor *P*, se utilizará para marcar estos casos, mientras que en los restantes su valor será *0*[48]. En la práctica esto significa que no podremos establecer las diferentes interpretaciones de que habla I. Bosque para los casos de *resuelto, ocupado* y que nuestra etiquetación será, en ambos casos, la de *aq0msp*.

Se dan algunas excepciones a este hecho: la primera está constituida por aquellos casos en que hay alternancia léxica adjetivo–participio, como los pares *lleno–llenado*, *sito–situado*, *limpio–limpiado* o *seco-secado* en que la forma en *-ado* tiene sólo etiqueta verbal.

A continuación pueden observarse distintos ejemplos de etiquetación de los adjetivos, atendiendo a los elementos anteriormente mencionados.

Glosa	**Etiqueta**	**Forma**
1 terminación	aq0cn0	burdeos, kitch
2 terminaciones	aq0cp0.	alegres
	aq0cs0.	alegre
4 terminaciones	aq0fp0.	bonitas, primeras
	aq0fs0.	bonita, primera
	aq0mp0.	bonitos, primeros
	aq0ms0.	bonito, primer / primero
participio	aq0msp.	llegado

[45]Las frases de la izquierda son del *Esbozo*: p. 494.
[46]Ejemplos tomados de RAE (1973): p.498.
[47]Extraídos de Bosque (1991): p.166.
[48]Por ejemplo, la etiqueta para "cansado" será *aq0msp* y la de "rojo" será *aq0ms0*.

2.2.4. Adverbio.

2.2.4.1. Definición, clasificación y propiedades

La clase de los adverbios es
la peor definida en las gramáticas
(Bosque (1991): p.127.)

La categoría **adverbio** no aparece tratada como tal en el *Esbozo*, aunque en diversos párrafos se mencione, como por ejemplo en el § 3.4.9 donde se dice que *los medios más usuales para expresar [las] relaciones circunstanciales son [...]: Adverbios o locuciones adverbiales*[49] entre otros.

Para Bello[50], el adverbio es la palabra que modifica al verbo o al adjetivo, aunque también puede modificar a otro adverbio. Una definición muy similar es la que aparece en la *Gramática española* de Alcina-Blecua donde el adverbio se considera una clase de palabras constituida por elementos *que actúan como términos terciarios con relación a verbos o adjetivos (términos secundarios) y a otros adverbios*[51]. Los problemas aparecen a la hora de inventariar los elementos que componen esta clase, ya que:

> *(a) De las palabras tradicionalmente incluidas entre los adverbios sólo una parte puede modificar a verbos, adjetivos y adverbios. Frente a esta parte, otra sólo conoce la referencia al verbo que se confunde con la situación de todo el enunciado en una determinada circunstancia.*
> *(b) Algunos adverbios, que aportan una información de tipo circunstancial al verbo o al enunciado total, tienen una manera de significar semejante a la de los pronombres.*
> *(c) Mientras una parte de adverbios, que admiten gradación, son de origen adjetivo y se forman por neutralización de los categorizadores de género y número, otra parte está en estrecha relación con preposiciones y otras categorías. Por otra parte, algunos de estos adverbios pasan fácilmente a habilitarse como marcas sintácticas de subordinación.*
> *(d) Por último, no se ha elaborado un criterio suficiente para marcar el límite entre el adverbio y el complemento circunstancial*[52].

Para Emilio Alarcos *adverbio* designa

> *una clase de palabras invariables en su significante y a menudo indescomponibles en signos menores, destinadas en principio a cumplir por sí solas el papel de adyacente de un adjetivo o de otro adverbio distinto*[53].

[49] RAE (1973): pp. 375-376.
[50] Bello (1847): § 64
[51] Alcina y Blecua (1989): p. 700.
[52] Alcina y Blecua (1989): pp.701–703.
[53] Alarcos (1994): p. 129.

Menciona también la adverbialización de los adjetivos, que se produce cuando éstos inmovilizan su flexión. Como ejemplos cita *temprano, claro, pronto, sólido, medio, bastante, mucho*. Otro procedimiento de adverbialización se produce, según este autor, cuando se añade al adjetivo en la forma del femenino singular el sufijo *-mente*.

I. Bosque[54] parte del hecho de que *la clase de los adverbios es la peor definida en las gramáticas* y considera que *adverbio* abarca demasiados elementos. Más adelante señala que *los adverbios son por lo general 'circunstantes' que sitúan la significación del verbo en unas coordenadas espaciales o temporales o que añaden información que completa la estructura argumental del predicado.*

Hay dos capítulos de la obra de Bosque y Demonte (1999) donde se habla de los adverbios: Pavón (1999) y Kovacci (1999). El primero, *Clases de partículas*, compara tres clases de palabras: preposiciones, conjunciones y adverbios. El segundo está dedicado exclusivamente al adverbio.

En Pavón (1999) se señalan como características comunes de las *partículas* el hecho de que son invariables desde un punto de vista morfológico; que son *elementos sintácticos encargados de establecer relaciones entre oraciones o entre partes de la oración*[55] y que los constituyentes encabezados por estos elementos suelen realizar la función de complemento circunstancial. La definición que propone de adverbios es la siguiente: *clase de palabras que modifica al verbo (o a la oración), al adjetivo o a otros adverbios [...] si bien existen ciertos adverbios (los de la clase de* incluso, casi, *etc.) que pueden modificar prácticamente a cualquier tipo de categoría gramatical*[56].

En Kovacci (1999) se destaca que *el aspecto morfológico es insuficiente por sí solo para caracterizar a [esta] categoría. Desde el punto de vista sintáctico, en cambio, es posible establecer un ordenamiento sistemático del adverbio, considerando las estructuras de las que forma parte*[57].

2.2.4.2. Criterios y etiquetación adoptados

Hemos considerado adverbios todas aquellas palabras que, por una parte, no tienen morfemas flexivos, bien porque en ningún caso los poseen, bien porque en el uso se fijan en la forma no marcada (caso de los adjetivos adverbializados: *claro, alto, primero...*); y que, por otra, son elementos que no establecen ningún tipo de relación entre elementos oracionales, lo que los opone a las preposiciones y conjunciones (que también son invariables). Además, son elementos intransitivos, en el sentido de que no necesitan obligatoriamente llevar complementos (al contrario que las preposiciones y conjunciones, que son elementos transitivos en este sentido).

El cuadro 2.2 muestra estos elementos definitorios del adverbio con respecto a otras

[54] Bosque (1991), Cap. 6.
[55] Pavón (1999): p. 567.
[56] Pavón (1999): p. 567.
[57] Kovacci (1999) p. 722.

clases de palabras con las que comparte ciertas propiedades[58].

	Adverbio	**Adjetivo**	**Preposición**	**Conjunción**
Flexión	-	+	-	-
Transitividad	-	-	+	+

Cuadro 2.2: Características del adverbio

Las clasificaciones que se han realizado de los adverbios suelen responder a criterios más semánticos que morfosintácticos, aunque también se ha propuesto clasificarlos según criterios distribucionales como en Bosque (1991) o Kovacci (1999). Por nuestra parte, hemos establecido dos tipos de adverbios basándonos en su comportamiento sintáctico: por una parte el adverbio *no* y por otra el resto de adverbios. Esta distinción se basa en el distinto comportamiento sintáctico del adverbio negativo respecto de los demás adverbios.

La etiquetación adoptada es la que aparece en el cuadro 2.3.

Atributo	**Valor**	**Código**
Categoría	Adverbio	R
Tipo	General	G
	Negativo	N

Cuadro 2.3: Etiquetas para los adverbios (1)

En el estándar Eagles se proponía un atributo más para el adverbio en el nivel L1: el grado, que no utilizamos aquí porque, al igual que ocurría con el adjetivo, la gradación es sintáctica y muy pocos adverbios utilizan la sufijación[59].

Finalmente, en el nivel L2 de Eagles, se propone la distinción entre los interrogativos y relativos por una parte (*adverbios-wh*) y el resto por otra. Los *adverbios-wh* quedan incluidos, en nuestra propuesta, en la categoría de los pronombres (interrogativos y relativos, respectivamente). Ambas clases poseen propiedades sintácticas especiales (son elementos anafóricos y pueden introducir subordinadas) y nos ha parecido más coherente tratarlas junto con el resto de pronombres relativos e interrogativos, respectivamente (cf. sección 2.2.5).

Las consecuencias de la clasificación aquí adoptada son que, aun incluyendo un gran número de palabras con propiedades distribucionales y semánticas diversas, la clase del adverbio queda homogéneamente definida por los criterios de invariabilidad morfológica y de no-relacionante de elementos oracionales.

Hemos de señalar que dentro de esta categoría hemos incluido también locuciones, es decir, grupos de palabras que funcionan unitariamente como un adverbio. El criterio para considerarlas ha sido la indivisibilidad de la secuencia y la invariabilidad morfológica.

Algunos ejemplos de adverbios y locuciones adverbiales son los siguientes:

[58]La caracterización de *preposición* y *conjunción* se comentará en las secciones 2.2.9.1 y 2.2.8.1 respectivamente.

[59]Se pueden mencionar, sin embargo, bien–mejor / lejos–lejísimos / cerca–cerquísima.

	despacio	*entonces*	*afortunadamente*
debajo	*después*	*abreviadamente*	*altamente*
debido	*detrás*	*absolutamente*	*alternativamente*
delante	*encima*	*actualmente*	*amargamente*
dentro	*enfrente*	*adicionalmente*	
deprisa	*enseguida*	*admirablemente*	

	a_cubierto	*a_la_larga*
a_borbotones	*a_cuestas*	*a_la_perfección*
a_bordo	*a_diario*	*a_la_postre*
a_buen_seguro	*a_fin_de_cuentas*	*a_la_vez*
a_cambio	*a_flote*	
a_ciegas	*a_fondo*	
a_continuación	*a_gusto*	

2.2.5. Pronombres y Determinantes.

En la bibliografía, pronombres y determinantes aparecen tratados de muy diversas formas: algunos autores los consideran como una única clase de palabras (Bello (1847), RAE (1973) y Alcina y Blecua (1989) y la lingüística moderna) mientras que otros los consideran sólo adjetivos, como Alarcos (1994). Por otra parte, tampoco hay acuerdo sobre las subclases que deben considerarse ni sobre el inventario de formas que cada una contiene.

Por todo ello, y aunque ya se ha establecido anteriormente que se tratan aquí como dos clases distintas, los presentamos en la misma sección; y, al igual que hemos hecho con las clases anteriores, tras una presentación teórica general detallamos la solución aquí adoptada.

2.2.5.1. Definición, clasificación y propiedades

Para Bello (1847) los pronombres son *nombres que significan primera, segunda o tercera persona, ya expresen esta sola idea, ya la asocien con otra* (§ 229), y se dividen, como el nombre, en sustantivos y adjetivos. Considera este autor tres clases de pronombres, a saber, los personales (de primera y segunda persona), los posesivos y los demostrativos; dentro de estos últimos se incluyen los relativos que *pasan a interrogativos acentuándose* (§ 320).

Las formas de los pronombres **personales**, *significan la idea de persona por sí sola*[60]. Los **posesivos** son *los que a la idea de persona determinada [...] juntan la de posesión, o más bien, pertenencia* e incluyen las series de *mi, tu, su* en sus formas plenas y apocopadas (§§ 248–253). Por último, los **demostrativos** son *aquellos de que nos servimos para mostrar los objetos señalando su situación respecto de determinada persona*[61]; pero

[60] Las formas de estos pronombres son *yo, nosotros, nosotras, tú, vosotros, vosotras, nos, os, me, te, mí, ti, conmigo, contigo*, más sus variantes estilísticas *nos, vos* (§§ 230–235). Las formas de *él, ella, ellos, ellas*, que este autor no incluye en la categoría de pronombre, resultan de la sustantivación del artículo y se declinan por casos (§§ 277 y 279 respectivamente).

[61] Sus formas son las de *este, ese, aquel, esto, eso, aquello* (§§ 254–263) y *tal, tanto* (§§ 339–339).

además se incluyen como demostrativos los relativos y los interrogativos: *Llamánse* relativos *los demostrativos que reproducen un concepto anterior, y sirven especialmente para enlazar una proposición con otra* (§ 304); además, como se ha mencionado anteriormente, *los pronombres relativos pasan a interrogativos acentuándose* (§ 320)[62]. En el capítulo XVIII, Bello menciona como sustantivos neutros las palabras *todo, mucho, más, menos, demasiado, bastante, asaz, harto, poco, algo, nada, nonada, uno, otro, ál* señalando que provienen de adjetivos. Y en el § 85, al hablar de las recategorizaciones que pueden afectar a algunas palabras, señala que *algo* y *nada* pueden ser sustantivos o adverbios (según el contexto en que aparezcan) y que *mucho, poco* y *más* pueden ser sustantivos, adjetivos o adverbios.

En el *Esbozo* se habla de una sola clase de palabras, el pronombre:

> *los pronombres constituyen en español una clase extensa de palabras dotadas de caracteres morfológicos y sintácticos, algunos de los cuales comparten con sustantivos y adjetivos, o exclusivamente con una de estas clases, pero otros son específicamente pronominales. Por otro lado no todos los pronombres participan por igual en dichos caracteres*[63].

A pesar de esta heterogeneidad, todas las formas tratadas como pronombres presentan una característica común,

> *que no es ni morfológica ni propiamente sintáctica, aunque tenga consecuencias de orden sintáctico. Son nulos o escasos los contenidos semánticos del pronombre*[64].

Por otra parte, todos los pronombres son o bien *deícticos*, porque señalan inconceptualmente a lo que vemos o recordamos o bien *anafóricos*, porque remiten a lo que se acaba de enunciar[65].

Las clases de pronombres que se consideran en esta obra son las siguientes: **personales, posesivos, demostrativos, relativos, interrogativos, indefinidos** y **cuantitativos**.

Los **pronombres personales** son los que designan a la tres personas del discurso. Presentan los morfemas de persona, género, número y caso y se reparten en formas acentuadas y formas inacentuadas, tal como puede observarse en el cuadro 2.4. A estas formas, presentadas en el capítulo dedicado a la Morfología[66], se añaden, cuando se trata el pronombre desde la sintaxis[67] las formas tónicas de caso preposicional *sí, consigo*.

Los **pronombres posesivos** comparten características con los personales, aunque no tienen caso y *se caracterizan [...] por la propiedad sintáctica de aparecer siempre, fuera de su función como predicados, en construcciones atributivas, a diferencia de los pronombres*

[62] Las formas de los relativos (e interrogativos) son *que, quien, cuyo, cual, cuanto* (Caps. XVI–XVII).
[63] RAE (1973): p. 202.
[64] RAE (1973): p. 202.
[65] RAE (1973): p. 202.
[66] RAE (1973): p. 204.
[67] RAE (1973): p. 422.

			Nominativo	Preposicional	Acusativo	Dativo
1	Sing.		*yo*	*mí, conmigo*	*me*	*me*
	Plur	Masc.	*nosotros*	*nosotros*	*nos*	*nos*
		Fem.	*nosotras*	*nosotras*	*nos*	*nos*
2	Sing.		*tú*	*ti, contigo*	*te*	*te*
	Plur	Masc.	*vosotros*	*vosotros*	*os*	*os*
		Fem.	*vosotras*	*vosotras*	*os*	*os*
3	Sing.	Masc.	*él*	*él*	*lo (le)*	*le, se*
		Fem.	*ella*	*ella*	*la*	*le (la), se*
		Neut.	*ello*	*ello*	*lo*	*le, se*
	Plur.	Masc.	*ellos*	*ellos*	*los (les)*	*les, se*
		Fem.	*ellas*	*ellas*	*las*	*les (las), se*

Cuadro 2.4: Pronombres personales según el *Esbozo*

personales, que están privados de esta propiedad[68]. Desde un punto de vista sintáctico, los posesivos funcionan exclusivamente como adjetivos[69]. Las formas de los posesivos expresan persona, género y número, y se dividen en formas inacentuadas y formas acentuadas; además semánticamente se dividen entre los que designan un solo poseedor y los que designan varios poseedores. Las formas de los posesivos son *mi, tu, su*, con sus respectivos plurales; y *mío, tuyo, suyo, nuestro, vuestro*, con sus respectivos femeninos y plurales.

Los **demostrativos** son palabras que realizan distintos tipos de señalamiento (y en este sentido se acercan al artículo). Tienen variación de género y número y pueden realizar una doble función sintáctica: sustantivo y adjetivo (aunque las formas neutras sólo puedan desempeñar la primera de estas dos funciones)[70].

Los **pronombres relativos** *realizan señalamientos anafóricos a palabras o complejos sintácticos del contexto. Se diferencian de todos los restantes pronombres por el hecho de funcionar simultáneamente, en la mayor parte de los casos, como nexos de subordinación*[71]. En el seno de la subordinada, unos funcionan como sustantivos (*que, quien, el que, el cual*) y otros como adjetivos (*cuyo*)[72].

[68] RAE (1973): p. 209.

[69] Casos como *Su mundo no es el nuestro* se explican no por la sustantivación de la forma del posesivo sino por el valor anafórico del artículo que reproduce el sustantivo *mundo*. El único caso de sustantivación de los posesivos que se señala es el de posesivos en plural acompañados del artículo y con significación de persona: *los suyos = sus partidarios, sus adeptos* (RAE (1973): p. 211.).

[70] Las formas de los demostrativos son: *este, ese aquel* con la variaciones de género y número, más la serie neutra *esto, eso, aquello* que sólo aparece en singular. Se incluyen también en el capítulo dedicado a los demostrativos las formas *tal, tanto*, que, como los otros, pueden funcionar indistintamente como adjetivos o sustantivos, aunque *tienen una sintaxis más complicada que los restantes demostrativos. Además son también adverbios, lo que da lugar con frecuencia a interferencias entre las dos categorías* RAE (1973): p. 217.

[71] RAE (1973): p. 218.

[72] Las formas de los pronombres relativos son: *que, quien, cual, cuyo, cuanto*. A las que hay que añadir las formas de los relativos correlativos, a saber, *tal ... cual, tanto ... cuanto*.

Los **pronombres interrogativos** son palabras que poseen acento de intensidad, marcado por la tilde gráfica. Lo que los distingue de los restantes pronombres *es el hecho de que sirven primordialmente como instrumentos a la función apelativa del lenguaje*[73]. Las funciones que pueden desempeñar, según las formas, son la de sustantivo y/o adjetivo[74].

Los **pronombres indefinidos** *poseen componentes conceptuales, lo que explica el hecho de que sus radicales entren más frecuentemente que el de los restantes pronombres en el mecanismo de la derivación y la composición [...]. Por otro lado, la mención que realizan deja sin identificar personas y cosas, bien porque no importa o no conviene o no es posible esta operación*[75]. Para el *Esbozo* el nombre **indefinidos** sirve para designar un amplio grupo de palabras que incluye también a los numerales. Las funciones que pueden realizar los indefinidos son o bien sólo la de sustantivo o bien, indistintamente, las de adjetivo o sustantivo. Las propiedades morfológicas de este grupo de palabras son el género (masculino, femenino y neutro) y el número (singular y plural)[76]. En lo referente a los numerales, son pronombres los cardinales, entre los que se incluyen las formas de *ambos*; pero no lo son los ordinales, que son nombres empleados la mayoría de las veces como adjetivos. Algunas de las formas de los indefinidos presentan variación de género y número; otras sólo de número; por último algunas son invariables.

Para Alcina–Blecua los pronombres (bajo cuya denominación se incluyen tanto determinantes como pronombres propiamente dichos) son palabras que se definen según los siguientes criterios morfológicos (a-c), sintácticos (d) y semánticos (e):

> *(a) forman una serie de sistemas morfológicos cerrados; (b) la mayor parte de ellas reciben morfemas de género o número como los nombres; algunas conocen el género neutro; (c) en determinados usos pueden neutralizar la oposición de género en singular; (d) funcionan en el discurso indistintamente de manera semejante a los sustantivos, adjetivos sustantivados, adjetivos o adverbios [...]; algunos de ellos, sin embargo, actúan específicamente en una sola determinada función; (e) semánticamente, su significado no es pleno hasta que no se les relaciona con el contexto lingüístico o extralingüístico en que son utilizados*[77].

La clasificación que proponen de los pronombres es la que aparece en el cuadro 2.5.

Los **pronombres personales** aluden a los *actuantes del discurso* y cubren tanto la *mención directa* como la *indirecta*, además distinguen la función sintáctica[78].

Las formas de los **posesivos** que proponen son las mismas que proponía el *Esbozo*; ambas obras coinciden también en decir que los posesivos son siempre adjetivos, aunque mientras el *Esbozo* no hablaba propiamente del sustantivación del posesivo con el artículo, Alcina-Blecua sí lo hacen: por una parte afirman que *los bisílabos posesivos aparecen*

[73] RAE (1973): p. 224.

[74] Una subclase de los interrogativos son los exclamativos. Las formas son *qué, quién, cuál, cuánto, cúyo*; la primera es invariable; las dos siguientes tienen variación de número, y las dos últimas de género y número.

[75] RAE (1973): pp. 226–227.

[76] El inventario de indefinidos que el *Esbozo* propone es el siguiente: *uno; alguno, ninguno; algo, nada; alguien, nadie; cualquiera; quienquiera; todo; más, menos; mucho, poco; otro; demás; cada; bastante, demasiado; varios.*

[77] Alcina y Blecua (1989): pp. 589–590.

[78] Alcina y Blecua (1989): pp. 596–599.

Indiciales de campo	Personales	
	Posesivos	
	Demostrativos	
	Locativos	Espaciales
		Temporales
Determinativos	Cuantitativos	Gradativos
		Existenciales
		Intensivos
	Numerales	Cardinales
		Ordinales
		Múltiplos
		Partitivos
		Distributivos
	Identificativos	
Relativos	Enunciativos	
	Interrogativos	
	Exclamativos	

Cuadro 2.5: Clases de pronombres según Alcina-Blecua

agrupados con el artículo formando construcciones sustantivas. El artículo concordante es anafórico siempre y alude claramente al sustantivo al que el posesivo determina[79]; luego señalan *el artículo del posesivo sustantivado...*[80].

En el grupo de los **demostrativos**, además de las formas correspondientes a *este, ese, aquel* con sus respectivas variaciones de género y número y a *esto, eso, aquello*, aparecen también las formas de *otro*. Todas estas formas pueden desempeñar la función sustantiva o adjetiva.

Los **pronombres locativos** son palabras que *funcionan como término terciario referidos a la totalidad del enunciado en que aparecen, semánticamente expresan circunstancia y formalmente no seleccionan morfemas concordantes*[81]. Las formas que incluyen son, entre otras, *aquí, ahí, allí; acá, allá; ahora, entonces; hoy, ayer, anteayer, mañana.*

Las formas de los **pronombres cuantitativos** se dividen en tres grupos: gradativos, que expresan *la gradación de cantidad, número o intensidad*[82]; existenciales, que *introducen en el discurso lo que no existe -serie negativa- o lo que existe y o no tiene nombre o se desconoce o no se quiere nombrar*[83]; y, por último, los intensivos que, según los autores

[79] Alcina y Blecua (1989): pp. 615–616.

[80] Alcina y Blecua (1989): p. 616. Y al hablar de la forma *lo* neutra lo llaman *transpositor de sustantivación* (Alcina y Blecua (1989): p. 569.)

[81] Alcina y Blecua (1989): pp. 629–630.

[82] Cuyas formas son: *mucho, poco, bastante, demasiado, harto, todo* (Alcina y Blecua (1989): p: 636).

[83] Las formas de estos pronombres son: *alguien, nadie, alguno, ninguno, algo, nada* (Alcina y Blecua (1989): p. 647).

se relacionan con los gradativos [...] en cuanto si éstos representan la gradación absoluta respecto de una totalidad, los intensivos al intensificar adjetivos, adverbios, verbos o la realidad aludida lo hacen por comparación[84].

Los **pronombres numerales** se dividen en cardinales, los que nombran la serie natural de los números enteros: *uno, dos, tres, ...* a los que hay que añadir los colectivos *dúo, ambos*; ordinales, que nombran la situación u orden dentro de la sucesión de los números enteros: *primero, segundo, ...*; los partitivos, que nombran las partes de una unidad y que se hallan repartidos en dos series, la primera formada por la estructura <ordinal + *parte* + *de*> y la segunda formada por derivación con el sufijo *-avo*; los multiplicativos, que nombran el resultado de multiplicación por los números naturales de una determinada realidad y que se forman por derivación con los sufijos *-ble, -ple, -plo*; y, por último, los distributivos, que *presuponen una especialización de cada una de las unidades componentes de un conjunto o una correlación entre cada uno de los componentes de un conjunto y otro u otros nombres* y cuyas formas son *sendos* y *cada*[85].

La serie de los identificativos comparte el hecho de expresar coincidencia de lo que se menciona con una realidad distinta y está formada por *mismo, igual, propio, tal, así, mientras, sí, también, no* y *tampoco*[86].

Los **pronombres relativos** se caracterizan por su comportamiento afín en la articulación del discurso y se distribuyen en dos series, tónicos y átonos. Las formas tónicas se utilizan como interrogativos y exclamativos, mientras que las átonas se utilizan como relativos enunciativos[87].

Emilio Alarcos, en su *Gramática de la lengua española*[88] no contempla la existencia de una categoría llamada pronombre. Tal como ya se mencionó anteriormente[89], este autor establece dos clases de adjetivos. Todas las formas agrupadas en la categoría adjetivo (tanto los de tipo uno como los de tipo dos) pueden sustantivarse. En el caso concreto de los demostrativos, por ejemplo, dice el autor: *el doble papel del demostrativo ha inducido a distinguir entre adjetivos y pronombres demostrativos. No es necesario, por cuanto todos los adjetivos, mediante la sustantivación, son capaces de cumplir en el enunciado la función de sustantivos.[...] Los demostrativos son, pues, una subclase de los adjetivos caracterizados porque para su sustantivación no requieren la aparición del artículo, ya que en su significado contienen el valor de identificación propio del artículo*[90]. Y esto se aplica, además de a los demostrativos, a posesivos, indefinidos y numerales.

Para este autor, bajo *la denominación de* **pronombres personales** *se agrupan varias palabras, en número limitado, cuyo contenido se refiere a la noción de persona*

[84] Las formas de los intensivos son: *más, menos, tanto, tan* (Alcina y Blecua (1989): p. 652).
[85] Alcina y Blecua (1989): pp. 668-669.
[86] Alcina y Blecua (1989): p. 675.
[87] Las formas son las mismas, unas acentuadas y con tilde; las otras inacentuadas y sin tilde: *que, cual, quien, cuyo, cuanto, cuando, como, donde, do*, mientras que las formas compuestas con el verbo 'querer': *comoquiera, dondequiera, quienquiera, cualquiera* son exclusivamente enunciativas. Alcina y Blecua (1989): pp. 687 y ss.
[88] Alarcos (1994).
[89] Cf. página 65.
[90] Alarcos (1994): p. 89.

gramatical[91]. Sin embargo, las divide en dos grupos (que presenta en dos capítulos separados) atendiendo a la tonicidad de unos y a la atonicidad de otros: *sustantivos personales* e *incrementos átonos del verbo*[92] respectivamente. Sobre los primeros dice que *constituyen en realidad una subclase de los sustantivos, puesto que coinciden con éstos en su función, y, al menos, parcialmente, entrañan unos mismos tipos de accidentes o morfemas (el número y el género)*[93]. Las formas que incluyen son las mismas que propone el *Esbozo*.

Las formas de los **demostrativos** son *este, ese, aquel* con sus correspondientes variaciones de género y número. La serie neutra sólo aparece en las funciones propias del sustantivo.

Los **posesivos** son unidades *de comportamiento funcional vario. Todas cumplen al menos una de las dos funciones propias de los adjetivos: la de adyacentes de un sustantivo (u otro elemento sustantivado) en un grupo nominal unitario, y la de atributo de un verbo*[94]. Las formas se clasifican en tres series: (i) unidades dependientes que exigen la presencia de un sustantivo[95]; (ii) unidades autónomas que solas pueden ser atributo de un núcleo verbal o adyacentes del sustantivo[96]; y (iii) los que cumplen las dos funciones del adjetivo[97].

Los **relativos** son palabras que se agrupan en una clase por motivos funcionales: *son capaces de trasponer o degradar (al menos bajo determinadas condiciones) los enunciados llamados oraciones [...] a la función de adyacente dentro de un grupo nominal unitario*[98].

Las formas de los **interrogativos** coinciden básicamente con las de los relativos, pero son palabras tónicas, por lo que llevan tilde gráfica y funcionan como elementos autónomos. La forma *cúyo* está actualmente en desuso; el relativo *el cual* tiene como equivalente la forma *cuál*; aparece en esta clase una nueva forma, aunque arcaizante: *cuán*.

Los **indefinidos** son palabras *con función sustantiva o adjetiva o con ambas alternativamente, cuyo rasgo común es de índole semántica*[99].

En Rigau (1999) se presentan por una parte los determinantes que incluyen *grosso modo* lo que la gramática tradicional considera artículo, demostrativos y posesivos y por otra los cuantificadores, que se corresponderían con los indefinidos y los numerales. Todas estas formas se comentan en diferentes capítulos de Bosque y Demonte (1999): el artículo en Leonetti (1999), los demostrativos en Eguren (1999), los posesivos en Picallo y Rigau (1999), los cuantificadores en Sánchez (1999) y Marcos (1999). Además, los pronombres personales, se tratan en Fernández (1999b).

[91] Alarcos (1994): p. 70.
[92] Capítulos VI y XV respectivamente.
[93] Alarcos (1994): p. 71.
[94] Alarcos (1994): p. 93.
[95] Cuyas formas son *mi, tu, su, mis, tus, sus*.
[96] Que incluye la siguiente serie: *mío, mía, míos, mías, tuyo, tuya, tuyos, tuyas, suyo, suya, suyos, suyas*.
[97] A saber *nuestro, nuestra, nuestros, nuestras, vuestro, vuestra, vuestros, vuestras*.
[98] Alarcos (1994): p. 98. Las formas incluidas en esta clase son: *que, el cual, quien, cuyo, como, donde, cuando, cuanto* con las variaciones que género y número que algunas poseen.
[99] Alarcos (1994): p. 114. Son indefinidos exclusivamente sustantivos las formas *alguien, algo, nadie, nada, quienquiera*. Los adjetivos son: *uno, alguno, ninguno, cualquiera; más, menos; mucho, poco, bastante; sendos, cada; todo; mismo*, a los que hay que añadir los numerales cardinales (el resto de numerales, se comportan según el autor, como adjetivos del grupo primero, esto es, como calificativos).

Para Rigau (1999), los determinantes y los cuantificadores son las palabras que capacitan a los sintagmas nominales para actuar semánticamente como un argumento del predicado oracional porque sólo con determinante o cuantificador pueden los sintagmas nominales expresar propiedades extensionales[100].

En Fernández (1999b) se definen los pronombres **personales** por sus relaciones con otras clases de palabras. Esta clase de pronombres *desempeña las mismas funciones sintácticas que el sustantivo (que los sintagmas nominales)* [101]. Se distingue del nombre común por el hecho de no tener rasgos semánticos inherentes, sino de adquirir significado según el contexto en que aparece. Con los nombres propios comparte el hecho de admitir el mismo tipo de adyacentes aunque sólo el pronombre admite la cuantificación y tiene el rasgo de persona. Sobre este rasgo, algunos autores han señalado que sólo es realmente pertinente para la primera y la segunda, mientras que la tercera es, en realidad, la *no–persona*[102]. Las formas de los pronombres personales son las que aparecen en el cuadro 2.6[103].

	Serie tónica		**Serie átona**	
	Sujeto	Objeto	Acusativo	Dativo
1s	*yo*	*mí, conmigo*	*me*	
2s	*tú*	*ti, contigo*	*te*	
3s	*él, ella, ello*		*lo, la*	*le*
1p	*nosotros, –as*		*nos*	
2p	*vosotros, –as*		*os*	
3p	*ellos, ellas*		*los, las*	*les*

Cuadro 2.6: Pronombres personales según Fernández (1999b)

Los rasgos que presentan los pronombres personales son los de persona, género, número y caso[104], aunque no todas las formas de los pronombres manifiestan explícitamente esta variación.

Los **demostrativos** forman parte, según Eguren (1999), de las expresiones deícticas y son términos abiertos, *cuya referencia no está fijada de antemano ni se mantiene constante, sino que se establece, crucialmente, cada vez que cambian el hablante, el oyente o las coordenadas espacio-temporales de los actos de enunciación*[105]. Los deícticos incluyen pronombres y adverbios demostrativos. Las formas de los pronombres son *este, ese, aquel*, con las correspondientes variaciones de género (masculino – femenino) y de número (singular – plural) y funcionan indistintamente como determinantes o pronombres, aunque *esto, eso, aquello*, sólo funcionan como pronombres. *Tal, tanto* son formas neutras que, en algunos de sus usos, pueden funcionar como pronombres (con valor identificativo) mientras que en otros funcionan como determinantes o adverbios (con valor cualitativo o cuantita-

[100] Rigau (1999): pp. 313–314.
[101] Fernández (1999b): p. 1211.
[102] Fernández (1999b): p. 1213.
[103] Fernández (1999b): pp. 1219 y 1221.
[104] Fernández (1999b): p. 1219.
[105] Eguren (1999): p. 931.

tivo)[106]. Para Rigau (1999) son determinantes demostrativos las formas: *este, ese, aquel* y *tal* y añade que las formas neutras no se combinan con sintagmas nominales[107].

Según Picallo y Rigau (1999) los **posesivos** son pronombres que *como los demás pronombres personales, distinguen formas acentuadas e inacentuadas. Presentan, además, formas apocopadas proclíticas y formas plenas*[108]. Las formas de estos pronombres son *mi, tu, su* con variación de número y *mío, tuyo, suyo, nuestro, vuestro* con variación de género y número. Se señala asimismo que si estas formas aparecen antepuestas al núcleo nominal actúan como determinantes, mientras que si aparecen pospuestas, el posesivo forma parte de un sintagma nominal con núcleo elíptico, lo mismo que ocurre en casos como los de 2.10[109]:

(2.10) (a) *Mi ordenador y* el tuyo *están estropeados.*
(b) *Fue una idea brillante* la suya.
(c) *Juan se sentía satisfecho d*el suyo.

Frente a estas tres clases de determinantes, aparecen los **cuantificadores** que se definen como grupo independiente por sus propiedades semánticas, no sintácticas. Son elementos *que dicen qué cantidad de individuos u objetos de un dominio dado tienen una determinada propiedad, o en qué medida una propiedad es poseída por un individuo u objeto*[110]. Según (Sánchez, 1999) la diferencia entre determinación y cuantificación es que la determinación de la referencia se hace, en el primer caso, por identificación del referente, mientras que en el segundo se determina por el tamaño del conjunto o por el número de *individualidades referidas*[111]. Y sin embargo, si bien la clase de cuantificadores se define por criterios semánticos, la cuantificación es un fenómeno sintáctico *que consiste en desencadenar la interpretación cuantitativa de ciertos elementos*[112]. Hay dos clases de cuantificadores según se clasifiquen por su capacidad implícita o explícita de denotar cantidad del elemento al que modifican: los propios y los focales o presuposicionales. Las formas y funciones de los cuantificadores propios que propone este autor son las que aparecen en el cuadro 2.7 donde $Det_{(P)}$ significa que son determinantes que pueden ocasionalmente funcionar como pronombres (ser núcleo de sintagma); $(pre)Det_{(P)}$ indica que además puede funcionar como predeterminante. Los elementos de la última columna señalan la variación de género (G), de número (N) o la invariabilidad morfológica (–).

Las formas de los cuantificadores focales son *también, incluso, hasta, tampoco, ni siquiera* y *sólo, al menos, apenas.* Todas funcionan como adverbios, pero mientras las del primer grupo son *incluyentes*, esto es, *presuponen otros valores posibles para el argumento cuantificado*, las del segundo son *excluyentes*, es decir, *niega[n] la presuposición*

[106]Por su parte, se tratan como adverbios demostrativos locativos las formas *aquí, ahí, allí, acá, allá,* temporales *ahora, entonces, hoy, ayer, mañana, anoche* además de algunas locuciones como *antes de ayer, pasado mañana, etc.* y, finalmente, como adverbio demostrativo de manera aparece la forma *así.*
[107]Rigau (1999): pp. 328–329.
[108]Picallo y Rigau (1999): p. 975, nota 1.
[109]Los ejemplos son de las autoras, p. 992.
[110]Sánchez (1999): p. 1027.
[111]Sánchez (1999): p. 1027.
[112]Sánchez (1999): p. 1029.

Clase	**Subclase**	**Formas**	**Función**	**Morfología**
NUMERALES	Cardinales	*uno, dos, mil*	$Det_{(P)}$	G–N
	Ordinales	*primero, décimo*	Adjetivo	G–N
	Partitivos	*mitad, tercio*	Nombre	N
	Multiplicativos	*doble, triple*	Nombre	–
	Distributivos	*sendos*	Det	G
INDEFINIDOS	Universales	*todo*	(pre)$Det_{(P)}$	G–N
		cada	Det	–
		cada uno	Pron.	G
		ambos	$Det_{(P)}$	G
		cualquiera	$Det_{(P)}$	N
	No-universales	*algo/alguien*	Pron.	–
		uno/alguno	$Det_{(P)}$	G–N
		varios/pocos	$Det_{(P)}$	G
		mucho/demasiado	$Det_{(P)}$	G–N
		bastante	$Det_{(P)}$	N
		nada/nadie	Pron.	–
		ninguno/alguno	$Det_{(P)}$	G
GRADATIVOS	Comparativos	*más/menos*	$Det_{(P)}$ – Adv	–
		tanto	$Det_{(P)}$ – Adv	G–N
	Proporcionales	*algo, (un) poco, mucho*	Adv	–
		bastante, demasiado		
		todo, nada		

Cuadro 2.7: Formas y funciones de los cuantificadores según Sánchez (1999)

de existencia de otr[os elementos][113]. Estas diferencias pueden apreciarse en los ejemplos siguientes[114]:

(2.11) (a) *Sólo Juan compró un coche.*
(b) *También Juan compró un coche.*

A todo esto hay que añadir la consideración que se hace en cada obra del artículo. Los autores coinciden, por lo general, en incluir bajo esta denominación sólo las formas *el, lo, la, los, las*. Sin embargo difieren en las consideraciones sobre este elemento. Para Bello (1847)[115] y RAE (1973)[116] estas formas están muy cerca de los demostrativos. Para Alcina y Blecua (1989) y Alarcos (1994), las formas del artículo son *el, la, lo, los, las*, pero no constituyen una clase de palabras, sino que son morfemas libres del sustantivo.

[113]Sánchez (1999): p. 1106.
[114]Tomados del autor, p. 1106
[115]Capítulo XIV.
[116]§ 2.6.

En Leonetti (1999) se sitúa al artículo definido entre los determinantes, puesto que restringe y define la referencia de los sintagmas nominales a la vez que ocupa una posición prenominal. El paradigma del artículo definido está formado por las formas *el, la, los, las, lo*; asimismo se critica el tratamiento que se ha dado a veces al artículo como morfema, puesto que *(a) es posible insertar sintagmas completos entre artículo y nombre [...]. (b) Es posible emplear el artículo en ausencia de núcleos nominales explícitos. (c) El artículo en español no es un elemento que se adjunta al nombre, sino una marca que caracteriza las propiedades referenciales de todo el SN y que, por lo tanto, es sensible a los rasgos aportados también por constituyentes distintos del núcleo nominal*[117].

Sobre el artículo indefinido, se dice que *si se acepta que los pronombres son esencialmente determinantes [...] no será necesario distinguir el pronombre indefinido* uno *del artículo o del numeral.*

2.2.5.2. Criterios y etiquetación adoptados

Revisadas hasta aquí las principales propuestas en torno al estatus de los pronombres y determinantes así como las distintas clasificaciones, cabe ahora hacer las siguientes observaciones:

1. Los criterios de definición difieren entre los autores. Sánchez (1999), Bello (1847) y RAE (1973) utilizan criterios semánticos; Alarcos (1994) propone una definición basada en la sintaxis; Alcina y Blecua (1989) se sirven de criterios morfosintácticos y semánticos; finalmente, RAE (1973) recurre a criterios pragmáticos.

2. No todos los autores tratan estas palabras como pertenecientes a dos categorías diferentes: para Bello, la RAE y Alcina-Blecua son pronombres que pueden funcionar como sustantivos o como adjetivos; para Alarcos son determinantes que admiten, en ocasiones, la sustantivación; en Bosque y Demonte (1999) se habla de determinantes y cuantificadores como dos clases distintas que pueden aparecer en sintagmas nominales con núcleo elíptico o explícito.

3. Las clasificaciones de los pronombres y/o de los determinantes que se hacen desde la teoría lingüística no coinciden. Para Bello sólo existen tres clases de pronombres: los personales, los posesivos y los demostrativos (que incluyen como subtipos los relativos y los interrogativos); mientras el *Esbozo* propone siete clases de pronombres, para Alcina y Blecua (1989) hay tres grandes clases que llegan a descomponerse en diez subclases posteriormente subdivididas; Alarcos (1994) considera 5 clases de pronombres, entre las que no se hallan los personales; por último en Bosque y Demonte (1999) se habla de *Determinantes* (artículo definido, demostrativos y posesivos) por un lado y de *cuantificadores* por otro. El caso del artículo definido merece especial atención, puesto que para algunos autores (Alcina y Blecua (1989) y Alarcos (1994)) es un morfema libre del sustantivo.

4. En ocasiones, una misma palabra aparece en clases o subclases diferentes; así por ejemplo, la palabra *mismo* es considerada un adjetivo por RAE (1973) mientras que

[117] Leonetti (1999): pp. 807–808.

en otras obras aparece como determinante. Otro ejemplo podría ser el de la palabra *él* que Bello considera un artículo sustantivado; que en RAE (1973) y Alcina y Blecua (1989) se trata como un pronombre personal; y que en Alarcos (1994) aparece como un sustantivo personal.

5. La categoría de los cuantificadores, como se ha visto, es más semántica que formal. Además los cuantificadores pueden desempeñar distintas funciones (determinantes, pronombres o adverbios).

A estas divergencias en el plano teórico hay que añadir que desde Eagles se plantean distintas formas de tratar los determinantes: o bien quedan incluidos en la categoría adjetivo, o forman una clase única. Lo que sí se sugiere es diferenciar pronombres de determinantes.

En nuestro sistema de etiquetación se tratarán los pronombres y los determinantes como categorías independientes atendiendo a criterios distribucionales y funcionales (recuérdese lo mencionado en la sección 2.2.2.2) a propósito de la clasificación de las palabras[118]. En primer lugar presentamos la descripción de los pronombres y a continuación la de los determinantes.

2.2.5.3. Pronombres

Las palabras que consideramos pronombres son aquellas que presentan las características que aparecen en la figura 2.9.

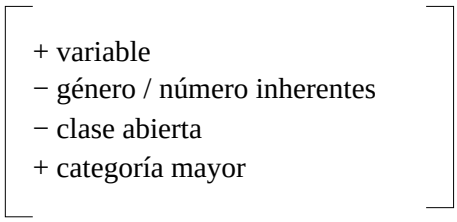

Figura 2.9: Características del pronombre

[118]Uno de los planteamientos que surgió fue el de tratar como una categoría independiente la de los **cuantificadores**. Los tipos podrían ser los que aparecen en el cuadro de la página 83: cardinales, partitivos, multiplicativos,.... Como atributos podrían señalarse el género y el número, puesto que algunos de los cuantificadores son palabras variables. Un último atributo hubiera podido ser el de la función, y tener como valores determinante, pronombre o adverbio. En este caso, la problemática es otra vez a misma: ante la palabra *todo* habría que decidir si es un cuantificador en función de determinante, pronombre o adverbio; del mismo modo que se plantea con la solución propuesta aquí.

Por otro lado, si no se incluyera este último atributo, lo que se conseguiría es dejar más ambigüedad en el nivel de análisis morfológico y, por tanto, habría que resolver el problema en la sintaxis.

Desde la perspectiva del PLN es conveniente distinguir determinantes de pronombres en los primeros niveles de análisis. Las estructuras en las que intervienen ambas clases de palabras son distintas, y si la correcta etiquetación puede hacerse en los niveles morfológicos en vez de esperar a los niveles sintácticos, mejor, porque se reduce la ambigüedad en los niveles de análisis posteriores.

Evidentemente, se dan casos de duplicación con formas tratadas aquí también como determinantes, pero las condiciones morfosintácticas en que ambas clases de palabras aparecen permiten la diferenciación y la desambiguación[119].

En el cuadro 2.8 puede observarse el tratamiento que hacemos de estos elementos.

Atributo	**Valor**	**Código**
Categoría	Pronombre	P
Tipo	Personal	P
	Demostrativo	D
	Posesivo	X
	Interrogativo	T
	Exclamativo	E
	Relativo	R
	Indefinido	I
	Numeral	N
Persona	Primera	1
	Segunda	2
	Tercera	3
Género	Masculino	M
	Femenino	F
	Neutro	N
	Común	C
Número	Singular	S
	Plural	P
	Invariable	N
Caso	Nominativo	N
	Acusativo	A
	Dativo	D
	Oblicuo	O
Poseedor	Singular	S
	Plural	P
Politeness	*Polite*	P

Cuadro 2.8: Etiquetas para el pronombre

La **categoría** pronombre presenta varios **tipos**: *personal, demostrativo, posesivo, interrogativo, exclamativo, relativo, indefinido*, y *numeral*. La caracterización de cada uno de estos tipos de pronombre se detalla en las secciones siguientes.

El atributo **persona** se utiliza para marcar la referencia a las tres personas del discurso. Este atributo aparece diferenciado en dos tipos de pronombre: los personales y los posesivos. En los restantes casos, el valor siempre es *0* (es decir, todos los pronombres excepto los personales y los posesivos, tienen inespecificado el atributo de persona, puesto que ninguno

[119] Véase el capítulo 3 para los criterios de desambiguación de estas dos clases de palabras.

de ellos refiere a las personas del discurso).

El **género** tiene cuatro valores: *masculino, femenino, neutro* y *común*. Los valores *masculino* y *femenino* marcan aquellas palabras en que hay una distinción morfológica explícita de género como en *vosotros–vosotras, los–las*, por ejemplo. El valor *común* se utiliza en aquellos casos en que es imposible determinar a nivel estrictamente morfológico el género de una forma, como en el caso de *le, tal, dos, que, quien, consigo, ustedes* entre otros. Por último, el valor *neutro* sólo aparece en las formas *esto, eso, aquello*, *lo_suyo* y *ello*, que claramente no pueden asignarse al género común (según la definición de género común que hemos establecido).

El **número** tiene dos valores fundamentales: *singular* y *plural*, que se utilizan para aquellos casos en que hay distinción formal de número: *cual, bastante, este, ella, el_suyo* o en aquellos casos en que una forma sólo admite el singular (*nadie, nada, mí*) o plural (*varios, tres*). El valor *invariable* marca aquellos casos en que una misma forma puede aparecer en contextos singulares o plurales: *qué, sí, consigo*.

El atributo de **caso** sólo es pertinente para el pronombre personal, aunque, como se comentará más adelante, no todas las formas de estos pronombres aparecen con un caso marcado. En el resto de pronombres, este atributo queda inespecificado.

Utilizamos el atributo **poseedor** sólo con los posesivos. Sus dos posibles valores son *singular* y *plural*, según el poseedor tenga número singular o plural. Este atributo es el que permite la distinción entre las formas *el_mío* y *el_nuestro*. Sin embargo, queda inespecificado para las formas de *el_suyo*, en las cuales es imposible distinguir a nivel puramente formal el número del referente.

Por último, el atributo booleano **politeness** se especifica sólo con los pronombres personales, en concreto en las formas de *usted, ustedes* y *vos*. Este atributo permite tratar los pronombres de cortesía como elementos de segunda persona a la vez que su presencia permite la diferenciación con respecto a las formas *tú / vosotros*.

Antes de pasar a un breve comentario sobre las particularidades de cada uno de los tipos de pronombres, vamos a resumir lo dicho hasta aquí con un cuadro (2.9), donde se muestran las relaciones entre los tipos de pronombres y los atributos pertinentes para cada tipo.

A continuación comentamos detalladamente todos y cada uno de los tipos de pronombres aquí tratados.

2.2.5.4. Pronombres personales

Como se ha visto anteriormente, no hay unanimidad a la hora de considerar las formas que se incluyen en el paradigma de los pronombres personales ni a la hora de considerar la definición de esta subclase de pronombre.

Consideramos pronombres personales las palabras que poseen rasgos de concordancia en persona, número, género y caso, *gracias a los cuales denotan inequívocamente al referente*

	Pers.	Gén.	Núm.	Caso	Poseed.	Polit.
Personal	+	+	+	+	-	+
Demostrativo	-	+	+	-	-	-
Posesivo	+	+	+	-	+	-
Interrogativo	-	+	+	-	-	-
Exclamativo	-	+	+	-	-	-
Relativo	-	+	+	-	-	-
Indefinido	-	+	+	-	-	-
Numeral	-	+	+	-	-	-

Cuadro 2.9: Relación tipo-pronombre – atributos

al que señalan en el contexto sintáctico en el que se insertan[120]. Las formas coinciden con las que propone (Fernández, 1999b) a las que añadimos los pronombres reflexivos.

La etiquetación de todos los pronombres personales y formas afines pueden observarse en el cuadro 2.10.

Comentamos a continuación cada uno de los atributos de los pronombres personales así como las formas que los poseen.

1. **Persona**

   La distribución de los distintos valores del atributo de persona es la siguiente:

   a) formas de primera persona: *yo, mí, conmigo, nosotros, nosotras, me, nos*;

   b) formas de segunda persona: *tú, ti, contigo, vosotros, vosotras, te, os, usted, vos, ustedes*;

   c) formas de tercera persona: *lo, la, los, las, sí, consigo, le, les, él, ella, ellos, ellas, ello, se*.

2. **Género**

   Como hemos comentado anteriormente, hemos considerado cuatro géneros en el sistema pronominal:

   a) Masculino
   para las formas *él, ellos, lo, los, nosotros, vosotros*

   b) Femenino
   para las formas *ella, ellas, la, las, nosotras, vosotras*

[120] Fernández y Anula (1995): p. 227.

Etiqueta	Forma	Etiqueta	Forma
pronombres personales			
nominativos			
pp1csn00.	yo	pp2csn00.	tú
acusativos			
pp3msa00.	lo	pp3mpa00.	los
pp3fsa00.	la	pp3fpa00.	las
pp3cna00.	lo		
oblicuos			
pp1cso00.	mí	pp1cso00.	conmigo
pp2cso00.	ti	pp2cso00.	contigo
pp3cno00.	sí	pp3cno00.	consigo
dativos			
pp3csd00.	le	pp3cpd00.	les
con caso inespecificado			
pp1mp000.	nosotros	pp1fp000.	nosotras
pp2mp000.	vosotros	pp2fp000.	vosotras
pp3ms000.	él	pp3fs000.	ella
pp3mp000.	ellos	pp3fp000.	ellas
pp3cn000.	se	pp3ns000.	ello
pp1cs000.	me	pp1cp000.	nos
pp2cs000.	te	pp2cp000.	os
pp2cs00p.	usted	pp2cp00p.	ustedes
pp2cn00p.	vos		
casos particulares			
p010s000.	me	p010p000.	nos
p020s000.	te	p020p000.	os
p0300000.	se	p0000000	se

Cuadro 2.10: Etiquetas para el pronombre personal

c) Común
para las formas *conmigo, consigo, contigo, le, les, lo, me, mí, nos, os, se, sí, te, ti, tú, ustedes, usted, vos, yo*

d) Neutro
para la forma *ello*

Debe comentarse aquí el hecho de que la forma *lo* aparezca dos veces, una con género masculino y otra con género común. El primer caso es el que corresponde a los ejemplos de 2.12 (a-b), esto es cuando este pronombre funciona como complemento directo; mientras que si funciona como atributo, ejemplos de 2.12 (c-d), entonces su

género es común, ya que esta forma sirve para referenciar elementos masculinos y femeninos indistintamente.

(2.12) (a) *Le encargó al multimillonario míster Chaplin que* **lo** *hiciera por él* (a28)
(b) *Y* **lo** *va consiguiendo* (d2).
(c) *Ésa es la bandera nacional. Pero oficialmente no* **lo** *era cuando Chibana la quemó* (a15).
(d) *No ha estado a la moda. [...] y tampoco* **lo** *estaría ahora* (a28).

3. **Número**

Los valores del atributo de número son tres:

a) Singular
Las formas de número singular son: *conmigo, contigo, él, ella, ello, la, le, lo, me, mí, te, ti, tú, usted, yo.*

b) Plural
Las formas plurales son: *ellas, ellos, las, les, los, nosotras, nosotros, nos, os, ustedes, vosotras, vosotros.*

c) Invariable
Por último, las formas de número invariable (indistinto para el singular o el plural) son: *consigo, lo, se, sí, vos.*

La forma *lo* aparece también en dos ocasiones, como singular (la forma masculina), y como invariable (la forma de género común). Los motivos son similares a los anteriormente aducidos al comentar el género: la forma singular refiere a elementos singulares, mientras que la invariable hace referencia tanto a singulares como plurales, tal como lo muestra el siguiente ejemplo:

(2.13) *Se ha deslizado en la mente de los españoles la convicción de que no somos 'refinados'. Es cierto que buena parte de nuestra cultura y de nuestra vida real no* **lo** *son; pero la otra parte* **lo** *es, y en muchos casos extremadamente* (a10).

4. **Caso**

Aparecen a continuación las formas asignadas a cada uno de los casos (nominativo, acusativo, dativo y oblicuo). Hemos marcado el caso sólo cuando las formas son totalmente inambiguas. Si una forma expresa dos o más casos (neutralización del morfema) hemos inespecificado el valor de este atributo, que será por tanto *0*. Otra observación que debemos hacer es que sólo tenemos en cuenta los usos rectos de los pronombres, es decir, no consideramos los casos de leísmo, loísmo o laísmo[121].

a) Nominativo: *yo, tú.*

b) Acusativo: *la, las, lo* (masculino singular), *lo* (común invariable), *los.*

[121] Más adelante, en el capítulo 3 página, 166, se detallará cómo se han etiquetado estos usos en el corpus.

c) Dativo: *le, les.*

d) Oblicuo: *conmigo, consigo, contigo, mí, sí, ti.*

e) Inespecificado: *él, ella, ellas, ello, ellos, me, nosotras, nosotros, nos, os, se, te, ustedes, usted, vosotras, vosotros, vos.*

A continuación aparecen ejemplos de algunas de las formas en que el valor del atributo de caso está inespecificado y se muestran las formas en contextos diferentes correspondientes a cada uno de los casos, lo que justifica que se mantengan con este atributo inespecificado. Los ejemplos 2.14 (a-d) muestran distintas apariciones de la forma *él*; en contextos similares pueden aparecer los pronombres *ella, ellas, ello, ellos, nosotras, nosotros, ustedes, usted, vosotras, vosotros, vos.* Las oraciones (e-f) ejemplifican los dos casos posibles de la forma *me* (acusativo y dativo) que son los que corresponden también a las formas *nos, os, se, te.*

(2.14) (a) *supe por qué* **él** *había permanecido todo el tiempo inmóvil en su silla* (a11). Caso Nominativo.
(b) *no lo hizo para que no lo ahorcasen a* **él** (t2). Caso Acusativo.
(c) *se pasa la vida buscando en los demás lo que a* **él** *le falta* (a23). Caso Dativo.
(d) *Si de* **él** *careciéramos, ¿para qué unas tareas que requieren esfuerzo [...]?* (a1). Caso oblicuo.
(e) *cualquiera que* **me** *conozca un poco sabe que yo solita, pintura incluida, soy un teatro lírico* (c1). Caso acusativo.
(f) *O mejor dicho,* **me** *gustaban* (a29). Caso dativo.

5. **Politeness**

Este atributo se utiliza para diferenciar las formas de tratamiento de cortesía del resto. Se trata de un booleano que sólo aparece marcado con el valor *P* en las formas *ustedes, usted, vos.*

Casos particulares

Los estudios sobre la forma *se* y, por extensión, sobre las formas *me, te, nos, os* son extensos; baste citar, a título de ejemplo, Alcina y Blecua (1989): pp. 907–923; Alarcos (1987): pp. 213–222; Marsá (1984): pp. 113–119 o Gómez (1992). En el primer caso se distinguen tres valores básicos: construcciones reflexivas (y recíprocas), construcciones de reflexivo medial (que incluyen casos como los de los verbos *levantarse, atreverse, confesarse de, caerse*) y, por último, construcciones de *se* de indeterminación de agente, es decir, pasivas-reflejas e impersonales reflejas. Para E. Alarcos puede hablarse de cinco valores distintos: (i) implementación o complementación reflexiva; (ii) alternancia de incrementación reflexiva, como en los casos de *acordar / acordarse; tratar / tratarse*; (iii) reflexivo enfático en función de complemento, como en la oración *me como una chuleta*; (iv) construcciones llamadas pasiva–refleja; y (v) construcciones impersonales.

Por último, para Gómez (1992) hay tres valores básicos: (i) pronombre personal; (ii) reflexivo, que incluye el pronombre reflexivo y recíproco con función nominal y el pronombre reflexivo sin función nominal, componente de verbo pronominal (como en los casos

de *dormirse, negarse, moverse*); y (iii) SE no pronominal, componente verbal (como en los casos de *se me antojó un pastel; se produjo una enorme explosión*) u oracional, que corresponde a los casos de las oraciones de pasiva refleja y a las oraciones impersonales.

En nuestra propuesta, además del caso anteriormente comentado de los pronombres personales, sean o no reflexivos o recíprocos (que se ejemplifican en 2.15) la forma *se* y, por extensión, las formas *me, te, nos, os* reciben otra etiqueta: *me_P010S000, nos_P010P000, os_P020P000, se_P0300000, te_P020S000*. Se trata de aquellos casos en que estas palabras no funcionan como auténticos pronombres personales, con función sintáctica de complemento directo o indirecto (casos acusativo y dativo, respectivamente), sino que aparecen como morfemas verbales de los verbos pronominales, como en los ejemplos 2.16.

(2.15) (a) *En aquel chaval de regate prodigioso* **se** *vio reflejado Emilio* (d1).
(b) *... sé que mientras católicos y protestantes* **se** *matan en Irlanda...* (c5)

(2.16) (a) *Con el personal* **me** *llevo bien* (n1).
(b) *Eres el líder, nos miras desde arriba, pero* **te** *rebajas hasta donde haga falta para seguir ahí* (d1).
(c) *Poner pasión en el trabajo es nobilísima conducta : quienes* **nos** *dedicamos a estos menesteres no podemos trabajar sin entusiasmo* (a1).
(d) *Lo amarillo es un precipicio cuyo vértigo* **se** *acerca a los despeñamientos de la luz y su circuito frenético* (a26).

En el caso de 2.16 (a) se trata del verbo pronominal *llevarse bien*; en (b), de *rebajarse*; en (c) el verbo es *dedicarse a*; y en (d) se trata de *acercarse a*. No puede afirmarse, desde un punto de vista lingüístico, que estos casos sean comparables a los presentados en 2.14 (e-f), por lo que les hemos otorgado una etiqueta distinta.

Además de todo ello, la forma *se* puede tener otro valor, lo que nos ha llevado a introducir otra etiqueta (*se_P0000000*) para aquellos casos en que se trata de oraciones de pasiva-refleja (ejemplos de 2.17 (a-b)) o impersonales (ejemplos (c-d)):

(2.17) (a) *Una sutil indiferencia hacia su suerte* **se** *extiende por el Bernabeu* (d1).
(b) *Por el contrario, la verdadera heroicidad* **se** *construye calladamente* (a14).
(c) **se** *compra a plazos* (c5).
(d) *En realidad,* **se** *trata de viejos conocidos que tuvieron un origen común* (dc10).

2.2.5.5. Pronombres demostrativos

Consideramos pronombres demostrativos aquellos elementos deícticos pertenecientes a una clase cerrada que tienen una función nominal en la oración. Reciben, pues, la etiqueta de pronombre demostrativo las formas correspondientes al paradigma de *este, ese, aquel* en sus variaciones de género (masculino, femenino y neutro) y número (singular y plural). Puesto que las normas de la RAE sobre acentuación[122] señalan que los pronombres demostrativos sólo se acentuarán obligatoriamente cuando exista riesgo de ambigüedad (lo que implica que en otros casos no necesitan tilde), se ha optado por otorgar esta etiqueta

[122] RAE (2000): p. 49.

tanto las formas sin tilde gráfica como las que la llevan. Además se etiquetan también como demostrativos las formas *tal, tales*, con la única variación de número[123]. Así pues, las formas etiquetadas como pronombres demostrativos son las que aparecen en el cuadro 2.11.

Formas	**Etiqueta**
este, ese, aquel éste, ése, aquél	PD0MS000
esta, esa, aquella ésta, ésa, aquélla	PD0FS000
estos, esos, aquellos éstos, ésos, aquéllos	PD0MP000
estas, esas, aquellas éstas, ésas, aquéllas	PD0FP000
esto, eso, aquello	PD0NS000
tal	PD0CS000
tales	PD0CP000

Cuadro 2.11: Formas y etiquetas para los demostrativos

Ejemplos tomados del corpus para esta categoría son:

(2.18) (a) **Tal** *fue el caso, por ejemplo, de la ya mencionada salazón o salvación eterna* (a23).
(b) **Tales** *son algunas de las más usuales invocaciones que se le hacen* (e1).
(c) *Y* **eso** *no es sólo una Operación Limpieza;* **eso** *es, señora Aguirre, limpieza étnica* (c1).
(d) *Es* **ésta** *una cultura llena de náufragos cuyas naves se fueron a la mierda* (c4).
(e) *Además, según los ligandos utilizados,* **éstos** *pueden o no intervenir en el proceso de conducción* (dc3).
(f) *Había resultado ser el único lujo a su alcance de entre todos* **aquellos** *con los que soñara de soltera* (t6).

La forma *tanto*, que Eguren (1999) y Alcina y Blecua (1989) sitúan entre los demostrativos aparece también entre los cuantificadores gradativos, según la propuesta de Sánchez (1999): *duerme* **tanto** *como quiere* (p. 1037). Admitiendo estos dos usos de la palabra, como demostrativo y como cuantificador, no se puede establecer ningún mecanismo puramente formal para su desambiguación, y dado que su uso más frecuente en corpus es el de cuantificador, hemos decidido no etiquetarlo como demostrativo. A esto podemos añadir que los usos residuales o poco frecuentes de las palabras pueden dar problemas a efectos de etiquetación. Por otra parte, las diferencias semánticas entre ambos usos tampoco son muy claras, tal como señala Eguren (1999).

[123]Tanto Rigau (1999) como Eguren (1999), entre otros, incluyen estas formas en el paradigma de los demostrativos.

2.2.5.6. Pronombres posesivos

Los posesivos son los únicos pronombres en los que se especifica el atributo *poseedor*, aunque se deja inespecificado para las formas *el_suyo*, que equivalen tanto a *de él* (que correspondería al singular) como a *de ellos* (que se asociaría al plural). Todos los posesivos presentan los atributos de persona (primera, segunda o tercera); género (masculino, femenino o neutro) y número (singular o plural). Reciben esta etiqueta todas las formas del paradigma de *el_mío, el_tuyo, el_suyo, el_nuestro, el_vuestro* con variación de género y número. En el cuadro 2.12 pueden observarse algunos ejemplos de pronombres posesivos con sus respectivas etiquetas.

Formas	Etiqueta
el_mío	PX1MS0S0
la_suya	PX3FS000
lo_suyo	PX3NS000
el_suyo	PX3MS000
la_nuestra	PX1FS0P0
los_nuestros	PX1MP0P0
los_vuestros	PX2MP0P0
los_suyos	PX3MP000

Cuadro 2.12: Etiquetas para los pronombres posesivos

Si hemos incluido el artículo en las formas de los pronombres posesivos ha sido porque su aparición es obligatoria en ausencia de un núcleo nominal[124]. Una consecuencia de esta etiquetación es que, en los casos de posesivos pospuestos (*un amigo mío*), que Picallo y Rigau (1999) consideran también pronombres, nosotros hemos optado por la etiquetación como determinantes posesivos (que aparecen pospuestos).

2.2.5.7. Pronombres interrogativos

Son pronombres interrogativos aquellas formas tónicas que aparecen en las oraciones interrogativas parciales: *cuál, cuánto, quién, qué, adónde, dónde, cómo, cuándo*, que quedan etiquetadas como aparece en el cuadro 2.13:

Como se ha comentado anteriormente, el atributo para la persona permanece inespecificado. Las formas de *cuánto* presentan variación de género y número; las de *cuál* y *quién* sólo de número, por lo que el género es común; la forma *qué* tiene género común y número invariable. Las últimas formas de los interrogativos tampoco presentan especificación de los atributos de género ni de número.

Al definir la categoría de adverbio y los elementos que la componen (cf. página 73), ya

[124]La inclusión del artículo en la forma del posesivo, hace problemática su etiquetación si aparecen tras las preposiciones *a, de* en las formas del masculino singular por la contracción entre preposición y artículo. Una prueba de ello aparece en la siguiente frase: *Nos deja realmente atónitos este mundo, simple y a la vez complejo, de los virus, por otra parte, inseparable* del nuestro (dc10). Por ello, y sólo en los casos masculino–singular, se ha etiquetado también la forma posesiva sin el artículo como pronombre.

Formas	Etiqueta
cuánto	PT0MS000
cuánta	PT0FS000
cuántos	PT0MP000
cuántas	PT0FP000
cuál, quién,	PT0CS000
cuáles, quiénes,	PT0CP000
qué	PT0CS000
adónde	PT000000
dónde	PT000000
cómo	PT000000
cuándo	PT000000

Cuadro 2.13: Etiquetas para los pronombres interrogativos

se mencionó que no se trataban los llamados *adverbios-wh* sino que estas palabras quedaban incluidas en la categoría de pronombre.

2.2.5.8. Pronombres relativos

Las palabras etiquetadas como pronombres relativos son las que aparecen encabezando las oraciones relativas. La definición de esta clase es más sintáctica que morfológica. Los pronombres relativos equivalen a nombres, adjetivos o adverbios a la vez que introducen oraciones subordinadas. Su etiquetación se muestra en el cuadro 2.14.

Formas	Etiqueta
cuanto, cuyo	PR0MS000
cuanta, cuya	PR0FS000
cuantos, cuyos	PR0MP000
cuantas, cuyas	PR0FP000
cual, quien	PR0CS000
cuales, quienes	PR0CP000
que	PR0CN000
donde	PR00000
como	PR00000
cuando	PR00000

Cuadro 2.14: Etiquetas para los pronombres relativos

Sobre la asignación de valores a los atributos de persona, género, número y persona, pueden hacerse las mismas observaciones que las referidas a los interrogativos (cf. sección 2.2.5.7).

2.2.5.9. Pronombres indefinidos

Las clase de los pronombres indefinidos es, sin duda, la que presenta mayores discrepancias entre los diferentes autores. Aquí hemos considerado indefinidos aquellos pronombres que expresan cantidad de forma imprecisa. Esta definición de subclase es semántica y no morfosintáctica.

En el *Esbozo* se incluyen en la clase de los indefinidos los cardinales y las formas de *algo, alguien, alguno, bastante, cada, cualquiera, demasiado, demás, menos, mucho, más, nada, nadie, ninguno, otro, poco, quienquiera, todo, uno, varios*. Muchas de estas formas son, para Alcina y Blecua (1989) pronombres cuantitativos (todas excepto *cada, cualquiera, demás, otro, quienquiera, uno, varios*[125], pero este grupo incluye además otras palabras como *harto, tan, tanto*. Las formas que E. Alarcos sitúa entre los indefinidos[126] coinciden con las que propone el *Esbozo*, pero añade a la lista las formas de *mismo* y de *sendos*. Por último, (Sánchez, 1999) menciona como cuantificadores con función pronominal los cardinales y las formas *todo, cada uno, ambos, cualquiera, algo, alguien, uno, alguno, varios, pocos, mucho, demasiado, bastante, nada, nadie, ninguno, más, menos, tanto*. Entre todos los autores, las formas incluidas entre los indefinidos (o equivalentes[127]) son: *algo, alguien, alguno, ambos, bastante, cada, cada uno, cualquiera, demasiado, demás, harto, menos, mismo, mucho, más, nada, nadie, ninguno, otro, poco, quienquiera, sendos, tal, tan, tanto, todo, uno, varios*.

Las formas que nosotros tratamos como pronombres indefinidos son: *algo, alguien, alguno, bastante, cualquiera, demasiado, demás, mismo, mucho, nada, nadie, ninguno, otro, poco, quienquiera, sendos, tanto, todo, uno, varios*, a las que hemos añadido: *cual, naide, todito*.

Han quedado, por tanto, fuera de la clase de los pronombres indefinidos las siguientes formas (en la columna de la izquierda aparece la forma y en la de la derecha la categoría que le hemos otorgado):

ambos	cardinal (entra en distribución con *dos*)
cada	determinante (necesita acompañar a un nombre)
cada uno	deter. + pron.
harto	adverbio (invariable; acompaña a cualquier clase de palabra)
menos	adverbio (invariable; acompaña a cualquier clase de palabra)
más	adverbio (invariable; acompaña a cualquier clase de palabra)
tal	demostrativo (véase sección 2.2.5.5)
tan	adverbio (invariable; acompaña a cualquier clase de palabra)

En el cuadro 2.15 aparecen estas formas clasificadas según los morfemas de género y número que presentan:

[125] Estos autores sitúan *cada* entre los numerales distributivos; *otro* entre los demostrativos y *uno* entre los cardinales.

[126] Recordemos que para Alarcos estas palabras son adjetivos de tipo segundo que, aislados, cumplen la función de sustantivo.

[127] Tanto si los tratan como una o dos categorías.

PI0MS000	alguno, demasiado, mismo, mucho, ninguno otro, poco, tanto, todito, todo, uno
PI0FS000	alguna, demasiada, misma, mucha, ninguna otra, poca, tanta, toda, todita, una
PI0MP000	algunos, demasiados, mismos, muchos, ningunos otros, pocos, tantos, toditos, todos, unos, varios
PI0FP000	algunas, demasiadas, mismas, muchas, ningunas otras, pocas, todas, toditas, unas, varias
PI0CS000	algo, alguien, bastante, cual, cualquiera nada, nadie, naide, quienquier, quienquiera
PI0CP000	bastantes, cualesquiera, demás, quienesquiera

Cuadro 2.15: Etiquetas para los pronombres indefinidos

2.2.5.10. Pronombres numerales

Tal como se indicó en la sección 2.2.3.2 hemos dividido los numerales en dos grupos, ordinales (que se tratan como adjetivos) y el resto, que quedan incluidos en la categoría de los pronombres numerales. Hemos incluido bajo esta denominación tanto los cardinales *uno, dos, tres, etc.*, como los denominados partitivos (*catorceavo*), los multiplicativos (*cuádruplo*) y las formas de *ambos*.

Algunas formas presentan doble variación de género y número, como *onceavo, trezavo*; otras sólo la variación de género como *uno, ambos, setecientos*; por último hay formas que sólo son plurales y de género común, como *cuatro, doce, dos, noventa*.

2.2.5.11. Los determinantes

Los determinantes son palabras pertenecientes a una clase cerrada que presentan variación de género y número concordante y que tienen que aparecer necesariamente en un sintagma nominal con núcleo no elíptico.

Los elementos que recogemos como determinantes son aquellos que presentan las características que aparecen en la figura 2.10[128].

En el cuadro 2.16 puede observarse el sistema de etiquetación adoptado para los determinantes.

Dentro de la **categoría** determinantes quedan incluidos los **tipos** *demostrativo, posesivo, interrogativo, exclamativo, indefinido, numeral y artículo*. Los tipos coinciden con los de los pronombres, si exceptuamos el artículo, que sería el equivalente del pronombre personal.

[128] Algunas teorías no consideran que los determinantes sean una categoría menor, tal como ya se mencionó anteriormente, puesto que pueden ser núcleo de sintagma. Para PLN resulta muy interesante mantener esta distinción con los pronombres.

Atributo	**Valor**	**Código**
Categoría	Determinante	D
Tipo	Demostrativo	D
	Posesivo	P
	Interrogativo	T
	Exclamativo	E
	Indefinido	I
	Numeral	C
	Artículo	A
Persona	Primera	1
	Segunda	2
	Tercera	3
Género	Masculino	M
	Femenino	F
	Común	C
	Neutro	N
Número	Singular	S
	Plural	P
	Invariable	N
Poseedor	Singular	S
	Plural	P

Cuadro 2.16: Etiquetas para los determinantes

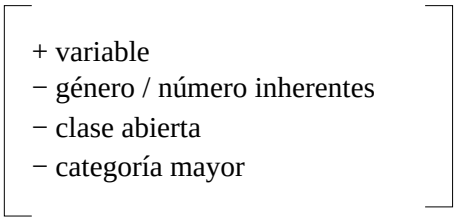

Figura 2.10: Características del determinante

En la categoría de los determinantes, el atributo de **persona** sólo presenta variación de valores en los posesivos. Los restantes se han marcado con el valor *0*, igual que ocurría con los pronombres.

El **género** puede tener los valores *masculino* (*muchos*), *femenino* (*muchas*), *común*, que es el de aquellas formas equivalentes para el masculino y el femenino (*bastante*)[129] y, por último, *neutro*, que aparece sólo con la forma *lo* del artículo.

El **número** puede tomar tres valores: *singular* (*mucho*), *plural* (*muchos*) o *invariable*; este último sólo se utiliza para la forma *qué* interrogativa-exclamativa. El valor 'invariable' no se aplica a aquellas formas determinativas que sólo existen en singular (*cada*) o en plural (*sendos*), que tienen explicitado el valor correspondiente (S / P).

El atributo **poseedor** se utiliza sólo con los determinantes posesivos, y toma los valores *singular, plural*, pero queda inespecificado para los casos de *suyo* porque no puede distinguirse el referente singular o plural.

2.2.5.12. Determinantes demostrativos

Los determinantes demostrativos son aquellas palabras de significado deíctico, pertenecientes a una clase cerrada que aparecen en sintagmas nominales con núcleo nominal explícito. Suelen ocupar una posición prenuclear, aunque también pueden aparecer detrás del nombre.

El total del formas etiquetadas como determinantes demostrativos es el que se muestra en el cuadro 2.17.

Los atributos de los demostrativos que aparecen especificados son los de género, que puede ser masculino (*este*), femenino (*esta*) o común (*semejante*); y el de número, que toma los valores singular o plural.

Tanto Bello (1847) como RAE (1973), Rigau (1999), y Eguren (1999) incluyen la forma *tal* entre los demostrativos. Este uso demostrativo puede observarse en frases como las siguientes:

(2.19) (a) *Pero tampoco podemos descartar completamente* **tal** *posibilidad* (dc2).

[129] Se trata de aquellos casos en que no es posible determinar a nivel puramente morfológico el género de la palabra.

Forma	Etiqueta	Forma	Etiqueta
aquel	DD0MS0	aquella	DD0FS0
aquellas	DD0FP0	aquellos	DD0MP0
cual	DD0CS0		
esa	DD0FS0	esas	DD0FP0
ese	DD0MS0	esos	DD0MP0
esta	DD0FS0	estas	DD0FP0
este	DD0MS0	estos	DD0MP0
semejante	DD0CS0	semejantes	DD0CP0
tales	DD0CP0	tal	DD0CS0

Cuadro 2.17: Etiquetas para los determinantes demostrativos

(b) *precisamente este tipo de países son los que [...] podrían sentirse más tentados de usar* **tales** *armas* (dc2).

Ninguna de las obras que hemos consultado menciona *semejante* como perteneciente a la clase de los demostrativos; pero, en ejemplos tomados de corpus aparecen estos usos de la palabra:

(2.20) (a) *ningún animal atesora en su piel* **semejante** *cúmulo de funciones.* (dc1)
(b) **Semejante** *renuncia nos ha defraudado un poco* (d2).
(c) *propiedades eléctricas* **semejantes** *a los metales* (dc3).
(d) *estos compuestos son* **semejantes** *a los metales orgánicos* (dc3).
(e) *apariencias fantasmagóricas* **semejantes** *a nubes movidas por el viento* (dc1).

El uso de la palabra *semejante* es distinto en los ejemplos de 2.20 (c-e), donde ocupa una posición posnominal y equivale a *parecido*, que en los de (a-b), en posición prenominal y equivaliendo a *tal, este*. Por este motivo, hemos establecido una doble etiqueta para esta palabra: determinante demostrativo, que podrá utilizarse para los casos de (a-b), y la de adjetivo calificativo, que es la que recibirá en los ejemplos de (c-e).

Por último, debemos explicar la inclusión de *cual* entre los demostrativos. La categoría de esta palabra es, por lo general, la de pronombre relativo (así aparece en 29 de las 31 apariciones en el corpus). Pero en la siguiente frase no corresponde a esta categoría sino a la de demostrativo, en coordinación con *tal*:

(2.21) *a ratos daban cierto repeluco y hasta nos proporcionaban tal o cual supuesto* (a29).

Aquí aparece en coordinación con otro demostrativo por lo que hemos incluido *cual* en la serie de estos determinantes[130].

[130]El otro ejemplo de *cual* es el siguiente: *Visto lo cual, Antonia le invitó a un café* (t5), que hemos etiquetado como pronombre indefinido.

2.2.5.13. Determinantes posesivos

Los determinantes posesivos son la formas de *mi, tu, su* con variación de género y número que aparecen en el sintagma nominal con núcleo explícito, normalmente antepuestos a dicho núcleo. Los determinantes posesivos se distribuyen en dos series: una con formas plenas y la otra con formas apocopadas. Esta últimas deben obligatoriamente preceder al núcleo.

Los atributos que presentan son los siguientes: persona gramatical, con distinción entre la primera (*mi*), la segunda (*tu*) y la tercera (*su*); género masculino (*mío*), femenino (*mía*) o común (todas las formas apocopadas); número, singular o plural; y poseedor singular (*mío*) o plural (*nuestro*). Este último atributo no se específica para las formas *su, suyo* puesto que la diferencia sólo es apreciable en contexto. Todas estas formas aparecen en el cuadro 2.18.

Forma	**Etiqueta**	**Forma**	**Etiqueta**
mía	DP1FSS	mías	DP1FPS
mi	DP1CSS	mis	D13CPS
mío	DP1MSS	míos	DP1MPS
nuestra	DP1FSP	nuestras	DP1FPP
nuestro	DP1MSP	nuestros	DP1MPP
sus	DP3CP0	su	DP3CS0
suyas	DP3FP0	suya	DP3FS0
suyos	DP3MP0	suyo	DP3MS0
tus	DP2CPS	tu	DP2CSS
tuyas	DP2FPS	tuya	DP2FSS
tuyos	DP2MPS	tuyo	DP2MSS
vuestras	DP2FPP	vuestra	DP2FSP
vuestros	DP2MPP	vuestro	DP2MSP

Cuadro 2.18: Etiquetas para los determinantes posesivos

2.2.5.14. Determinantes interrogativos

Las palabras incluidas en esta subclase son los elementos interrogativos que aparecen encabezando oraciones interrogativas y precediendo a un nombre o equivalente. Al contrario de lo que ocurría con los pronombres, esta clase contiene sólo dos formas básicas: *qué* y la serie de *cuánto*. La persona está inespecificada; sólo la serie de *cuánto* presenta variación de género (masculino y femenino) y de número (singular y plural); mientras que la forma *qué* presenta género común y número invariable. Estas formas aparecen en el cuadro 2.19.

2.2.5.15. Determinantes exclamativos

Las formas que incluimos como determinantes exclamativos son las mismas que aparecen como determinantes interrogativos. con la etiqueta **DE**.

Forma	Etiqueta	Forma	Etiqueta
cuánta	DT0FS0	cuántas	DT0FP0
cuánto	DT0MS0	cuántos	DT0MP0
qué	DT0CN0		

Cuadro 2.19: Etiquetas para los determinantes interrogativos

2.2.5.16. Determinantes indefinidos

El problema de la delimitación de formas en el caso de los determinantes indefinidos es el mismo que aparecía al tratar los pronombres.

Las formas que aquí hemos considerado determinantes indefinidos son las siguientes: *alguno, bastante, cada, cierto, cualquiera, cuanto, demasiado, demás, diferentes, distintos, escaso, escasísimo, mismo, mucho, ninguno, otro, poco, propio, sendos, tanto, todo, uno, varios.*

La mayoría aparece en las obras consultadas excepto: *cierto, escaso, escasísimo* y *diferentes, distintos, diversos* (sólo en plural). Si hemos considerado estas formas como determinantes ha sido porque hemos encontrado ejemplos en el corpus que corroboran este hecho. Ello no significa que siempre sean determinantes: sólo cuando preceden al núcleo nominal. Los ejemplos de 2.22 muestran estas palabras en función de determinante, por contraposición a los ejemplos de 2.23 donde aparecen pospuestas o fuera de un sintagma nominal (y, por tanto, como adjetivos).

(2.22) (a) *tratamiento médico de* **ciertas** *enfermedades* (dc1).
(b) *venciendo la* **escasa** *resistencia de los debilitados defensores* (dc2).
(c) *semejantes a los que presentan* **diferentes** *tipos de insectos* (dc1).
(d) *estamos introduciendo nuevos parámetros como las* **distintas** *estaciones del año* (dc1).
(e) *con* **escasísimas** *creencias morales* (a21).

(2.23) (a) *si admitimos como* **cierta** *la idea antropológica que afirma que...* (dc3).
(b) *esa dignidad en el comportamiento público [...] que cada día resulta más* **escasa** *entre nosotros* (a4).
(c) *un millón de tipos de anticuerpos* **diferentes** (dc10).
(d) *no es sino la totalización de [...] categorías* **distintas** (e2).

Las formas de los determinantes indefinidos están inespecificadas con respecto a la persona; la mayoría presenta variación de género (masculino–femenino) y número (singular–plural), aunque algunas no manifiestan género explícito (*bastante*) y otras aparecen sólo en singular (*cada*) o en plural (*demás*).

Finalmente, debemos mencionar que *alguno, ninguno* aparecen, en masculino singular, tanto en la forma plena como en la apocopada. En 2.24 aparecen ambas formas como

determinantes: en (a) pospuesto a causa de la presencia del adverbio negativo *no*; en (b) antepuesto, en una frase enunciativa afirmativa.

(2.24) (a) *quiso mirar a su alrededor y no advirtió cambio alguno* (t6).
(b) *con Antonio en casa siempre había algún quehacer* (t5).

En el cuadro 2.20 aparecen todas las formas etiquetadas como determinantes indefinidos:

DI0MS0	algún, alguno, cierto, cuanto, demasiado escasísimo, escaso, mismo, mucho, ningún, ninguno otro, poco, propio, tanto, todo, un
DI0FS0	alguna, cierta, cuanta, demasiada, escasa escasísima, misma, mucha, ninguna, otra poca, propia, tanta, toda, una
DI0MP0	algunos, ciertos, cuantos, demasiados distintos, escasísimos, escasos, mismos, muchos ningunos, otros, pocos, propios, sendos, tantos todos, unos, varios
DI0FP0	algunas, ciertas, cuantas, demasiadas distintas, escasas, escasísimas, mismas, muchas ningunas, otras, pocas, propias, sendas, tantas todas, unas, varias
DI0CS0	bastante, cada, cualquier
DI0CP0	bastantes, cualesquiera, cualesquier, demás, diferentes

Cuadro 2.20: Etiquetas para los determinantes indefinidos

2.2.5.17. Determinantes numerales

Las formas etiquetadas como determinantes numerales son las mismas que aparecían como pronombres numerales, con los mismos valores para los atributos de persona, género y número. En el cuadro 2.21 aparecen algunos ejemplos.

Etiqueta	**Forma**	**Forma**
dn0fp0.	ambas	cuatrocientas
dn0cp0.	cuatro	cinco
dn0mp0.	cuatrocientos	céntuplos
dn0ms0.	cuádruplo	dieciochavo

Cuadro 2.21: Etiquetas para los determinantes cardinales

2.2.5.18. Artículo

Hemos incluido el llamado artículo definido en la categoría de los determinantes. Sobre el llamado artículo indefinido, además de que no hay acuerdo sobre si debe considerarse

una clase especial de determinante, diferente al indefinido o al numeral, tampoco es posible diferenciar a nivel formal entre estos usos; incluso a nivel semántico resulta especialmente compleja la distinción entre indefinido y artículo. Por ello, hemos tratado las formas de *un* con los indefinidos y hemos considerado artículo sólo el definido.

La variación morfológica que presenta es la de género masculino (*el, los*), femenino (*la, las*) y neutro (*lo*) y número singular (*el, la, lo*) y plural (*los, las*). Las etiquetas completas aparecen en el cuadro 2.22.

Etiqueta	Forma
da0ms0.	el
da0fs0.	la
da0mp0.	los
da0fp0.	las
da0ns0.	lo

Cuadro 2.22: Etiquetas para los artículos

2.2.6. Nombre.

2.2.6.1. Definición, clasificación y propiedades

No hay ninguna [palabra] que tan fácilmente se reconozca y distinga, ni que sea tan a propósito para guiarnos en el conocimiento de las otras. (Bello (1847): § 41).

En la sección 2.2.3 ya hemos comentado que algunos autores incluyen sustantivo y adjetivo en una misma superclase: el nombre. También hemos señalado las diferencias que establece Ignacio Bosque entre ambas categorías. Por ello, nos limitaremos aquí a reproducir lo concerniente exclusivamente al sustantivo.

Bello define sustantivo como

> *la palabra esencial y primaria del sujeto, el cual puede también componerse de muchas palabras, dominando entre ellas un sustantivo, a que se refieren todas las otras [...]. El SUSTANTIVO es, pues, una palabra que puede servir para designar el sujeto de la proposición. Se dice que* puede servir*, no que* sirve*, porque además de esta función, el sustantivo ejerce otras*[131].

Los nombres se clasifican, según este autor, en *propios* y *apelativos*. Los primeros incluyen sólo sustantivos, mientras que los apelativos agrupan sustantivos y adjetivos. En § 100, el autor define el nombre propio como *el que se pone a una persona o cosa individual para distinguirla de las demás de su especie o familia*, mientras señala que el apelativo *es el que conviene a todos los individuos de una clase, especie o familia, significando su naturaleza o las cualidades de que gozan.*

[131] Bello (1847): § 41.

En cuanto a su significado, *los sustantivos no significan sólo objetos reales o que podamos representarnos como tales aunque sean fabulosos o imaginarios [...] sino objetos también en que no podemos concebir una existencia real, porque son meramente las cualidades que atribuimos a los objetos reales, suponiéndolas separadas o independientes de ellos*[132].

En el *Esbozo* se señala que el nombre sustantivo posee los morfemas de género masculino y femenino y los de número singular y plural. Se hace una distinción entre apelativos y propios (que no se explicita porque, se dice, *nada tiene que ver con la Gramática*[133]) que se corresponde con la tradicional distinción entre nombres comunes y propios. Desde un punto de vista sintáctico, se señala

> *el nombre sustantivo puede desempeñar en la oración los oficios de núcleo del sujeto y de complemento predicativo en el predicado nominal; puede formar modos adverbiales y ser también complemento de otro nombre, de un adjetivo y de un verbo*[134]

Para Alcina y Blecua (1989) las palabras con función primaria son sustantivos, mientras que las que desempeñan una función secundaria son adjetivos. También se ha comentado anteriormente (cf. sección 2.2.3) que estos autores consideran sustantivos y adjetivos como pertenecientes a una misma clase de palabras: el nombre, que se caracteriza porque todas las palabras incluidas en ella *admiten los categorizadores que se denominan género, número y artículo (o uno de ellos por lo menos) en su realización en el mensaje*[135]. En la clasificación que establecen de esta clase, se mezclan criterios morfosintácticos y semánticos. Su clasificación es la siguiente:

nombres de tipo (a): *España, Azorín*;
nombres de tipo (b): *caos, cenit*;
nombres de tipo (c): *aceite*;
nombres de tipo (d): *francés, verde, sabio*;
nombres de tipo (e): *filósofo, físico, viajero*.

Los del tipo (a) no admiten artículo; los de tipo (b) no admiten número plural; los de tipo (c) matizan el significado al pasar de singular a plural en la relación de materia a clase de dicha materia (*Hay aceite en Andalucía; los aceites andaluces son famosos*; los de los tipos (d) y (e) pueden aparecer como términos primarios o secundarios. Las diferencias existentes entre los dos últimos tipos de nombres no aparecen de modo muy claro: los nombres de (d) y algunos de (e) admiten gradación, aunque también otros como *hombre, niño, torero* a los que puede anteponerse el adverbio *muy*. Por otra parte, la derivación adverbial con el sufijo *-mente* sólo es posible en algunos casos de los nombres de tipo (d).

Para Emilio Alarcos, que trata el sustantivo como una categoría diferenciada del adjetivo, es sustantivo

[132] Bello (1847): § 103.
[133] RAE (1973): p. 172
[134] RAE (1973): p. 401.
[135] Alcina y Blecua (1989): p. 497.

> *toda palabra capaz de cumplir en los enunciados llamados oraciones la función de sujeto explícito [...] o la de objeto directo [...] sin necesidad de ningún otro elemento*[136].

Según este autor, todo sustantivo comporta un morfema de género y, en general, variación de número. En lo referente a las clases de sustantivos, señala los propios y los comunes o apelativos:

> *Frente a los sustantivos comunes o apelativos, que clasifican los objetos de la realidad física o mental como pertenecientes a una determinada clase, los nombres propios identifican con su etiqueta a un objeto dado, que resulta inconfundible para los interlocutores*[137].

En Bosque (1999) los sustantivos se dividen en comunes y propios:

> *El sustantivo llamado 'común' o 'apelativo' es la categoría gramatical que expresa la pertenencia de las cosas a alguna clase. El 'nombre propio' es la categoría que distingue o identifica una cosa entre los demás elementos de su misma clase*[138].

Se deduce de esta definición (semántica) que los nombres comunes y los propios pertenecen a clases distintas. En la misma línea se manifiesta Fernández (1999a) cuando señala que:

> *Los nombres propios [...] constituyen una categoría no exclusivamente lingüística; su carácter marginal deriva de la dificultad que supone su delimitación mediante las relaciones intrínsecas entre los signos que constituyen el sistema de una lengua: es una clase de palabras desprovista de contenido léxico codificado*[139].

En esta obra se señala que las características morfológicas, sintácticas y semánticas del nombre propio son distintivas pero no exclusivas; esto es, ningún criterio permite establecer una disyunción clara entre el nombre propio y las otras clases de palabras a partir de estos elementos. Así, las principales características del nombre propio son:

1. desde el punto de vista morfológico:
   - *a*) flexión fija (aunque con algunos matices y con casos de vacilación);
2. desde el punto de vista sintáctico:
   - *a*) por lo general, ausencia de determinante;
   - *b*) incompatibilidad con complementos restrictivos o especificativos, ya que no definen una clase léxica;

[136] Alarcos (1994): p. 60.
[137] Alarcos (1994): p. 68.
[138] Bosque (1999): p. 5.
[139] Fernández (1999a): p. 79.

3. desde el punto de vista semántico:

   a) unicidad referencial;

   b) falta de significado léxico;

   c) imposibilidad de traducción.

Además de todo ello, a nivel gráfico los nombres propios se distinguen porque se escriben en mayúsculas.

2.2.6.2. Criterios y etiquetación adoptados

Consideramos nombres aquellas palabras que poseen flexión inherente de género. Los clasificamos en comunes y propios. Los comunes admiten la anteposición del indefinido *un*. A nivel semántico, los nombres comunes categorizan objetos, mientras que los propios los identifican.

La etiquetación que proponemos para el nombre es la que queda reflejada en el cuadro 2.23.

Atributo	**Valor**	**Código**
Categoría	Nombre	N
Tipo	Común	C
	Propio	P
Género	Masculino	M
	Femenino	F
	Común	C
Número	Singular	S
	Plural	P
	Invariable	N
–	–	0
–	–	0
Apreciativo	Sí	A

Cuadro 2.23: Etiquetas para el nombre (1)

Dentro de la **categoría** nombre tratamos dos **tipos**: los comunes y los propios. A pesar de que las propiedades formales de unos y otros no son las mismas creemos que nada impide tratarlos como dos subtipos de una misma categoría, dado que también comparten algunas características, especialmente sintácticas: el nombre propio puede realizar las mismas funciones que el sintagma nominal. Una característica tipográfica importante es la mayúscula, que distingue al nombre común del propio. En nuestro sistema de etiquetación, las etiquetas para los nombres propios no incorporan información sobre el género y el número; los valores de estos atributos se asignan posteriormente, mediante reglas sintácticas, en función de la *trigger-word* que los acompaña[140].

[140] Véase Arévalo (2001) y Arévalo et al. (2002)

En el caso de los nombres comunes, en cambio, sí se codifica esta información. Hemos considerado dos **géneros** morfológicos (masculino y femenino) y tratamos como nombres de género común aquellos que presentan la misma forma para el masculino y el femenino sin que ello implique un cambio de significado, como por ejemplo las palabras *atleta, mar*. En aquellos casos en que el cambio de género implica un cambio de significado se utilizan los valores masculino / femenino: *cólera, cometa*. Un caso particular de esta situación se produce en palabras como *guardia, policía* o *escolta* que, como femeninos, tienen dos valores: por una parte un colectivo o genérico y por otra una clase de personas concretas de sexo femenino, mientras que como masculino son un nombre concreto. Un caso similar se da con el nombre *inconsciente*: como nombre referido a persona puede ser masculino o femenino, pero como nombre abstracto sólo es masculino. Por ello, las palabras *guardia, policía, escolta* recibirán dos etiquetas, una de género común y otra de género femenino: *nccs000, ncfs000*, igual que *inconsciente*, que recibe una etiqueta con género común y otra con género masculino: *nccs000, ncms000*.

En otro orden de cosas, los sustantivos femeninos que empiezan por *a* tónica y que, por cuestiones de eufonía, aparecen con algunos determinantes en la forma masculina (*el agua fría*), aparecen etiquetados como femeninos.

El atributo de **número** toma los valores singular–plural–invariable. Etiquetamos como invariables aquellos nombres que presentan la misma forma para el singular que para el plural: *crisis, análisis, cortapapeles, limpiabotas*. Una vez más, *invariable* significa que no es posible determinar morfológicamente el número singular o plural de la palabra. En los casos de *tijeras–tijera, pantalones–pantalón*, como se admiten ambas formas, hay doble etiqueta singular–plural.

Los dígitos quinto y sexto se han utilizado en una aplicación concreta (Arévalo (2001)) para clasificar desde un punto de vista semántico los nombres propios. Esta clasificación no está plenamente incorporada al analizador, por lo que por el momento, estos atributos aparecen siempre subespecificados.

El atributo **apreciativo**, que aparece en lugar del de *grado* propuesto por Eagles, se utilizará en un futuro para tratar los apreciativos en los nombres comunes: diminutivos, aumentativos y despectivos.

En el cuadro 2.24 aparecen distintos ejemplos de etiquetación de nombres.

Etiqueta	Forma
nccp000.	mares, oyentes
nccs000.	mar, oyente
ncfs000.	cometa, chica
ncms000.	cometa, chico
ncfn000.	tesis
ncfp000.	chicas
ncmp000.	chicos
ncmn000.	cortapapeles
np00000.	Medardo_Fraile

Cuadro 2.24: Etiquetas para el nombre (2)

2.2.7. Verbo.

Según cierto moderno filólogo, los verbos son
"aquellas palabras que significan (o en otro tiempo significaron)
el acto de ejecutar los movimientos materiales y (por extensión)
las operaciones de los espíritus". Este definición
tiene el pequeño inconveniente de contradecirse a sí misma.
Si las palabras que en otro tiempo significaron movimiento y ya no,
son todavía verbos ¿no se sigue que
varios verbos no significan hoy movimiento?
(Bello (1847) Nota III.)

2.2.7.1. Definición, clasificación y propiedades

Para Bello (§§ 35-40) el verbo es la principal palabra del atributo entendido como la parte de la proposición que da a conocer lo que pensamos acerca del sujeto.

> *El VERBO es [...] una palabra que denota el atributo de la proposición, indicando juntamente el número y persona del sujeto y el tiempo del mismo atributo*[141].

Bello no trata infinitivos, gerundios y participios como formas propiamente verbales, sino como derivados verbales[142] ya que *derivan inmediatamente de algún verbo y [...] le imitan en el modo de construirse con otras palabras.*

Sobre los tiempos compuestos señala que *propiamente no pertenecen a la conjugación material*[143]. Señala que las formas simples están formadas por las inflexiones del verbo y las compuestas son *frases en que está construido el participio sustantivado del verbo con cada una de las formas simples de* haber *[...]; el infinitivo del verbo con cada una de las formas simples de* haber, *mediando entre ambos la preposición* de *[...]; o el gerundio del verbo con una de las formas simples de* estar *[...].* Haber *y* estar *se llaman, por el uso que se hace de ellos en estas frases, verbos* auxiliares[144].

En § 476 señala que el verbo indica también el modo del atributo. Considera Bello que los modos son *las inflexiones del verbo en cuanto provienen de la influencia o régimen de una palabra o frase a que esté o pueda estar subordinado*[145]. Los modos verbales que distingue son el Indicativo (*formas que son o pueden ser regidas por los verbos* saber, afirmar, *no precedidos de negación*[146]), y Subjuntivo; éste se divide en dos: subjuntivo común (*formas que se subordinan o pueden subordinarse a los verbos* dudar, desear[147]) y subjuntivo hipotético que presenta *un constante significado de condición o hipótesis*[148]. El subjuntivo común *presta sus formas* al subjuntivo optativo empleado en *proposiciones*

[141] Bello (1847): § 40.
[142] Bello (1847): Capítulo XX y nota IX.
[143] Bello (1847): § 487.
[144] Bello (1847): § 617.
[145] Bello (1847): § 450.
[146] Bello (1847): § 455.
[147] Bello (1847): § 459.
[148] Bello (1847): § 469.

independientes para significar el deseo de un hecho positivo o negativo[149]. Este modo tiene una forma particular en el imperativo ya que las formas *di, ven, hablad, escribid* son, según el autor, abreviaciones de *quiero que digas, deseo que vengas, que habléis, que escribáis*[150].

Finalmente, el indicativo presenta cinco tiempos verbales en las formas simples (*presente, pretérito, futuro, co-pretérito, pos-pretérito*), y el subjuntivo tres: presente, pretérito y futuro.

Según el *Esbozo*,

> *el verbo, por sus caracteres formales, es aquella parte de la oración que tiene morfemas flexivos de número, como el nombre y el pronombre, morfemas flexivos de persona, como el pronombre personal, y, además, a diferencia del nombre y del pronombre, morfemas flexivos de tiempo y modo*[151].

Esta definición no incluye las llamadas formas no personales del verbo[152] que se definen como sigue:

> *el infinitivo es un sustantivo verbal; el gerundio, un adverbio verbal; y el participio, un adjetivo verbal. [...] Además, por ser formas no personales, tienen en común el no expresar por sí mismas el tiempo en que ocurre la acción, el cual se deduce del verbo de la oración en que se hallen, de los adverbios que los acompañen y de otras circunstancias de la elocución. Son aptas, en cambio, para la expresión de la pasiva y del aspecto perfecto o imperfecto del hecho que significan*[153].

Las formas verbales finitas (formas con flexión) toman los siguientes morfemas: persona (1, 2, 3); número (singular, plural); tiempo (presente, pretérito imperfecto, pretérito perfecto simple, futuro y condicional); y modo (indicativo, subjuntivo e imperativo). Las formas infinitas o no personales carecen de los morfemas de persona y número.

En principio, las formas compuestas quedan incluidas dentro de la flexión verbal (*la flexión de los verbos españoles comprende formas simples y formas compuestas*), aunque luego añade: *si nos atenemos a los principios lingüísticos más rigurosos, estas formas llamadas compuestas no constituyen tema propio de la Morfología, sino de la Sintaxis, ni más ni menos que otras perífrasis verbales*[154].

Alcina-Blecua presentan el verbo en el capítulo 5 de su *Gramática* y, sin embargo, no ofrecen una definición de esta categoría. Desde un punto de vista morfológico, las formas verbales pueden dividirse en dos grandes grupos: las formas personales y las formas no personales. Las primeras poseen morfemas auxiliares y concordantes; las segundas carecen de estos últimos. Son morfemas auxiliares el de tiempo (pasado, presente, futuro, potencial),

[149] Bello (1847): § 464.
[150] Bello (1847): § 467.
[151] RAE (1973): p. 249.
[152] Que se tratan, en el *Esbozo* en capítulo aparte y dentro de la Sintaxis.
[153] RAE (1973): p. 483.
[154] RAE (1973): p. 252.

el de modo (imperativo, indicativo y subjuntivo), y el de aspecto (perfecto e imperfecto); son morfemas concordantes el de persona (1, 2, y 3) y el de número (singular y plural). En lo referente a las formas compuestas, las presentan como oposición de las formas simples para cualquier verbo[155]; en ellas el verbo *haber* y el morfema auxiliar del participio *aportan la información de tiempo, modo y aspecto*[156].

Emilio Alarcos define el verbo como

> *una clase de palabras que funcionan como núcleo de la oración, y que, en consecuencia, son susceptibles de aparecer representándola sin necesidad de otras unidades*[157].

En cuanto a la caracterización morfológica de estas formas, el autor señala que los verbos poseen, por un lado, los morfemas de persona y número (que no son exclusivos de esta categoría), y, por otro, y de manera exclusiva, los de modo (imperativo, indicativo, condicionado y subjuntivo), tiempo (presente, pasado y futuro) y aspecto (terminativo y no terminativo).

Las formas compuestas de los verbos quedan totalmente integradas en la conjugación: *si bien separados sus dos componentes en la grafía, son unidades globales en cuanto al sentido* y se oponen a las simples en la expresión de la anterioridad[158].

Los infinitivos, gerundios y participios los trata como formas derivadas del verbo (o *formas nominales*) que tienen como características particulares el no poder funcionar como núcleo de la oración y el carecer de los morfemas propios del resto de formas verbales.

Según se define en Alcoba (1999): p. 4917,

> *el verbo es una clase de palabras que significan un evento, una acción, proceso o estado. Son núcleos predicativos y núcleos de complementación sintáctica.*

A esta definición semántico-sintáctica, añade más adelante otra morfológica:

> *los verbos se manifiestan en distintas formas léxicas, se conjugan, para significar diferencias de modalidad en la consideración del evento por parte del hablante; diferencias de aspecto en la forma de desarrollarse o producirse la acción, acabada o no; diferencias de momento: presente, pretérito o futuro; y diferencias en cuanto a las personas que intervienen en la realización del evento de que se trata y su número.*

Los morfemas que presenta el verbo son, según se cita, la vocal temática, el tiempo, modo y aspecto, y la persona y el número. La variación en la vocal temática está motivada por factores estrictamente morfológicos ya que depende de la conjugación del verbo y de los valores de los restantes morfemas, mientras que los restantes están determinados extraléxicamente, por la sintaxis, según la oración o enunciado en que aparecen[159].

[155] Alcina y Blecua (1989): p. 738.
[156] Alcina y Blecua (1989): p. 766.
[157] Alarcos (1994): p. 137.
[158] Alarcos (1994)
[159] Alcoba (1999): p. 4919.

El tiempo, según Rojo y Veiga (1999): p. 2879, es una *categoría gramatical deíctica mediante la cual se expresa la orientación de una situación.* Una definición muy similar se da en de Miguel (1999), donde el tiempo se opone al aspecto: así, mientras el tiempo *es una categoría deíctica [que] localiza el evento verbal en un tiempo externo, orientándolo bien en relación con el momento del habla, bien en relación con el tiempo en que tiene lugar otro evento*, *el* aspecto flexivo *es la información relativa al modo en que tiene lugar un evento* y *se ocupa del tiempo como una propiedad inherente o interna del propio evento [...] sin hacer referencia al momento del habla*[160].

En Alcoba (1999) se indica que los infinitivos, gerundios y participios sólo presentan vocal temática y acaso el morfema de aspecto, pero carecen de los restantes morfemas de las formas flexivas. Los considera variantes de la conjugación verbal; más en concreto variantes flexivas (que no derivativas) por dos motivos: el primero, porque aparecen como núcleos de perífrasis verbales y, el segundo, porque cuando hay metábasis, en algunos casos el cambio categorial queda restringido[161]. Una opinión contraria se manifiesta en Rojo y Veiga (1999), donde se afirma que estas formas quedan fuera del conjunto formado por las formas flexivas por su comportamiento sintáctico[162].

En lo concerniente a las formas compuestas, se señala, en Alcoba (1999): p. 4921, que *morfológicamente no se puede considerar que una forma compuesta sea una variante flexiva del verbo, sino más bien una construcción, un tipo particular de perífrasis verbal: con una forma fija del participio y la forma flexionada del verbo* haber, *en función auxiliar.* En el mismo sentido se manifiesta Cartagena (1999) cuando señala: *todos los tiempos compuestos formados por la perífrasis* <haber + *participio*>.... En cambio, en Rojo y Veiga (1999) se afirma que *existen razones para considerar [...] que las formas verbales compuestas no constituyen complejos gramaticales disociables en dos elementos* y que *los significados gramaticales expresados por formas simples y compuestas se integran en un mismo conjunto estructurado*[163], a diferencia de los que ocurre con las perífrasis.

A modo de resumen, presentamos el cuadro 2.25 que esquematiza los distintos aspectos comentados acerca del verbo. La segunda columna hace referencia al hecho de incluir o no las llamadas formas no finitas del verbo dentro del paradigma verbal; la tercera a si los tiempos compuestos (TC) pertenecen propiamente a la conjugación verbal; la última, a la variación morfológica que presenta el verbo.

2.2.7.2. Criterios y etiquetación adoptados

Consideramos verbos aquellas palabras con variación morfológica de persona, número, tiempo y modo. Las formas con esta variación son las formas flexivas, pero además también incluimos en la categoría de verbo las formas no personales, esto es, infinitivo, gerundio y participio, atendiendo al hecho de que a nivel sintáctico tienen un comportamiento similar

[160] de Miguel (1999): p. 2989.

[161] Como ocurre por ejemplo en el uso adjetivo del participio del verbo *corromper*: no podemos utilizar la forma *corrompido*, sino *corrupto* (Alcoba (1999): p. 4923.)

[162] Rojo y Veiga (1999): p. 2871.

[163] Rojo y Veiga (1999): pp. 2870-71

	{inf,ger,part} ∈ V	**TC ∈ V**	**Morfemas verbales**
(Bello, 1847)	+	+	Núm, Pers, Tpo, Mod
(RAE, 1973)	-	+	Núm, Pers, Tpo, Mod
(Alcina y Blecua, 1989)	+	+	Núm, Pers, Tpo, Mod, Asp
(Alarcos, 1994)	-	+	Núm, Pers, Tpo, Mod, Asp
(Rojo y Veiga, 1999)	-	+	
(Cartagena, 1999)		-	
(Alcoba, 1999)	+	-	Núm, Pers, Tpo, Mod, Asp

Cuadro 2.25: Cuadro resumen de la caracterización lingüística del verbo

al de las formas flexivas: reciben complementos típicamente verbales y son núcleos de perífrasis.

Nuestra propuesta de codificación de las formas verbales es la que aparece en el cuadro 2.26.

A nivel de etiquetación morfológica tratamos sólo las formas simples del verbo. Las formas compuestas las tratamos, igual que las perífrasis verbales, en una fase de análisis posterior, entre la morfología y la sintaxis[164]. La base de esta decisión es tanto teórica como aplicada. Teórica porque consideramos, junto con Alcoba (1999), Cartagena (1999) e Yllera (1999) que los tiempos compuestos pueden tratarse del mismo modo que las perífrasis. A nivel aplicado y teniendo en cuenta las herramientas de que disponemos en la actualidad, hemos optado por llevar a cabo, en la primera fase, una tarea de reconocimiento de elementos y reservar para la fase siguiente su agrupación, de modo que tanto los tiempos compuestos del verbo como las perífrasis verbales se tratarán posteriormente.

La **categoría** verbo presenta tres **tipos**: *principal, semiauxiliar y auxiliar*; consideramos que todos los verbos excepto *ser, haber* son principales. Las formas correspondientes a *ser* son semiauxiliares, y las de *haber* auxiliares. La justificación para esta diferenciación es de índole práctica: distinguir, en fases posteriores de análisis, los tiempos compuestos (que aparecerán siempre con el auxiliar *haber*) y las formas pasivas (que aparecerán con el semiauxiliar *ser*) del resto de formas verbales.

Al tratar el **modo** verbal la distinción más habitual es la de imperativo, indicativo y subjuntivo (RAE (1973), Alcina y Blecua (1989), Ridruejo (1999))[165]. A esta diferenciación, puede añadirse la de formas finitas y formas no-finitas, que tradicionalmente no se ha incluido dentro del modo, cosa que sí se ha hecho en el ámbito de la lingüística sajona. Lo que proponemos en nuestro sistema es seguir las propuestas Eagles y considerar simultáneamente ambas clasificaciones, lo que recogemos en la figura 2.11.

El atributo **tiempo** presenta cinco valores: pasado–imperfecto–condicional–presente–futuro. El pasado y el condicional sólo se utilizan en el modo indicativo.

En la **persona** se establecen la primera, la segunda y la tercera. Por último, en el

[164]Cf. capítulo 4.

[165]Recordemos brevemente que Bello (1847) distinguía dos tipos de subjuntivo, el común y el hipotético; por su parte Alarcos (1994) añade a los tres mencionados el modo condicionado.

Atributo	**Valor**	**Código**
Categoría	Verbo	V
Tipo	Principal	M
	Semiauxiliar	S
	Auxiliar	A
Modo	Indicativo	I
	Subjuntivo	S
	Imperativo	M
	Infinitivo	N
	Gerundio	G
	Participio	P
Tiempo	Presente	P
	Imperfecto	I
	Condicional	C
	Futuro	F
	Pasado	S
Persona	Primera	1
	Segunda	2
	Tercera	3
Número	Singular	S
	Plural	P
Género	Masculino	M
	Femenino	F

Cuadro 2.26: Etiquetas para el verbo (1)

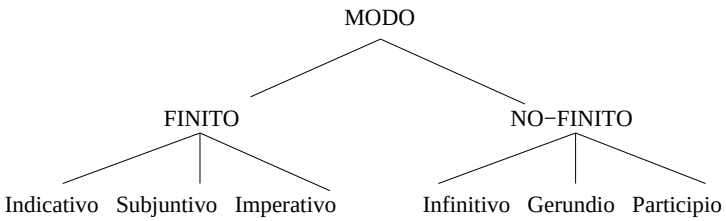

Figura 2.11: Los modos verbales

número se distingue singular de plural, y el **género**, masculino y femenino[166].

En los casos en que uno de estos atributos no resulta relevante, se le asigna el valor '0'[167].

Sobre el aspecto verbal, que, como se vio anteriormente, todos los autores consideran morfema del verbo, hay que decir que su expresión va más allá de la morfología estricta de la palabra, en el sentido de que todas las formas simples, excepto el pretérito indefinido, tienen aspecto imperfectivo mientras que todas las formas compuestas, además del pretérito indefinido, tienen aspecto perfectivo. Como en esta fase sólo tratamos las formas simples, no hemos considerado el morfema de aspecto, que se confunde con el de tiempo.

Ejemplos de etiquetación de formas verbales pueden observarse en el cuadro 2.27.

Verbos Principales		**Verbos (Semi)auxiliares**			
Etiqueta	**Forma**	**Etiqueta**	**Forma**	**Etiqueta**	**Forma**
vmip1s0.	como, amo	vaip1s0.	he	vsip1s0.	soy
vmip2s0.	comes, amas	vaip2s0.	has	vsip2s0.	eres
vmip3s0.	come, ama	vaip3s0.	ha	vsip3s0.	es
vmip1p0.	comemos, amamos	vaip1p0.	hemos	vsip1p0.	somos
vmip2p0.	coméis, amáis	vaip2p0.	habéis	vsip2p0.	sois
vmip3p0.	comen, aman	vaip3p0.	han	vsip3p0.	son
vmii1s0.	comía, amaba	vaii1s0.	había	vsii1s0.	era
vmii2s0.	comías, amaba	vaii2s0.	habías	vsii2s0.	eras
vmii3s0.	comía, amaba	vaii3s0.	había	vsii3s0.	era
vmii1p0.	comíamos, amábamos	vaii1p0.	habíamos	vsii1p0.	éramos
vmii2p0.	comíais, amabais	vaii2p0.	habíais	vsii2p0.	erais
vmii3p0.	comían, amaban	vaii3p0.	habían	vsii3p0.	eran

Cuadro 2.27: Etiquetas para el verbo (2)

[166]Esta distinción sólo afecta a los participios.

[167]Esto ocurre para los atributos de tiempo, persona y número en infinitivos, gerundios y participios, en el caso del atributo de género en todas las formas verbales excepto en el participio, y en el atributo de tiempo del imperativo.

2.2.8. Conjunción.

De esta recíproca permuta de oficios no se infiera
que sería mejor reducir esas tres clases de palabras a una sola.
Son esencialmente distintos los oficios del adverbio,
de la preposición, de la conjunción;
la palabra que pasa de una clase a otra varía de sintaxis y aún de significado.
(Bello (1847) Cap. L.)

2.2.8.1. Definición, clasificación y propiedades

Según Bello, *la conjunción sirve para ligar dos o más palabras o frases análogas, que ocupan un mismo lugar en el razonamiento, como dos sujetos de un mismo verbo [...]*[168]. En otro párrafo (§ 77) indica que además de unir elementos análogos de una proposición, también une proposiciones enteras. Sin embargo no proporciona un inventario completo de las formas que son conjunciones, aunque en el Capítulo L de su gramática titulado "Observaciones sobre el uso de algunos adverbios, preposiciones y conjunciones", menciona como conjunciones (o como palabras que perteneciendo originalmente a otra clase se comportan como tales): *ahora bien, ahora pues, antes, antes que, apenas ... cuando, apenas si, así que, así es que, aunque, bien que, como, como que, con que, luego, más, ni, o, pero, empero, porque, pues, puesto que, si, sino, y.*

La primera definición que el *Esbozo* proporciona de conjunción es que son palabras que enlazan oraciones dentro del período, aunque más adelante añade que también pueden enlazar elementos de la oración simple a condición de que éstos sean de la misma clase; ésta es una característica que la distingue de la preposición que siempre subordina a su término[169]. El primer tipo de conjunciones serían las subordinantes o hipotácticas, mientras que las segundas serían coordinantes o paratácticas. La diferencia entre ambos tipos de relación se explica del siguiente modo:

> *las oraciones coordinadas se enlazan en el período y expresan relaciones variadas entre sí; pero no se funden hasta el punto de que una de ellas pase a ser elemento sintáctico de otra. [...] Las subordinadas, en cambio, son elementos incorporados formalmente a la oración principal o subordinante, como sujeto, predicado o complemento de cualquier clase*[170].

En lo referente a las formas de las conjunciones y locuciones conjuntivas que se mencionan en esta obra, son las siguientes:

---

[168]Bello (1847): § 74.
[169]RAE (1973): p. 501.
[170]RAE (1973): p. 503.

y	e	ni	que	o
u	mas	pero	empero	sino
aunque	sin_embargo	no_obstante	antes_bien	con_todo
más_bien	fuera_de	excepto	salvo	menos
más_que	antes_que_no	si		
según(_que)	como_para	como_que	como_si	según_y_como
según_y_conforme	igual...que	lo_mismo_que	más...que(de)	
a_que	para_que	a_fin_de_que	pues	pues_que
porque	puesto_que	supuesto_que	de_que	ya_que
como(_quiera)_que	luego	conque	por_consiguiente	por_(lo_)tanto
por_esto(eso)	así_que	así_pues	tanto/tan...que	tal...que
así...que	de_modo_que	en_grado_que	cuando	como
siempre_que	caso_(de_)que	con_tal_que	con_solo_que	con_que
aunque	aun	de_manera_que		

Para Alcina–Blecua, el inventario de conjunciones es muy reducido y no aceptan la distinción entre conjunciones coordinantes y subordinantes, puesto que estas últimas están estrechamente relacionadas con los adverbios y los pronombres. Hablan sólo de conjunciones coordinantes (elementos que enlazan tanto oraciones como elementos o constituyentes de elementos): *y, e, ni, o, u, mas, pero, sino, empero, pues, luego* y de transpositores. Éstos últimos se definen como

> *marcas que advierten del encajamiento de una oración como elemento de una oración compuesta o como constituyente de un elemento de una oración compuesta*[171].

La clase de transpositores está formada por un reducido grupo de palabras átonas que tienen una función sustantiva, adjetiva o adverbial o sólo la de marcas. Los transpositores son los pronombres relativos enunciativos e interrogativos y las *marcas* **que** y **si**.

Emilio Alarcos señala que

> *con el término de* conjunciones *se reunen en una misma categoría las unidades lingüísticas que permiten incluir oraciones dentro de un mismo enunciado. Se distinguen las de* coordinación *y las de* subordinación. *[...] Las primeras son* conectores *que funden en un único enunciado dos o más oraciones que de suyo podrían manifestarse aisladas como enunciado. [...] Las conjunciones de subordinación, en cambio, degradan (al igual que los relativos) la oración en que se insertan y la transponen funcionalmente a una unidad de rango inferior que cumple alguna de las funciones propias del sustantivo, del adjetivo o del adverbio*[172].

Entre las conjunciones se incluyen palabras como *y, e, ni, o, u, o_bien, sea, pero, sino, mas, que, si, como, como_si, si_bien* y lo que el autor llama "conjunciones, locuciones o construcciones" transpositoras como *para_que, ya_que, pues, a_que, a_fin_de_que, aunque, y_eso_que, bien_que, mal_que* o *a_pesar_de(_que).*

[171] Alcina y Blecua (1989): p. 977.
[172] Alarcos (1994): p. 227.

En Pavón (1999) las conjunciones se definen como *una clase de palabras cuya misión es relacionar oraciones o elementos de una oración*[173]. Hay dos tipos de relaciones que pueden establecer: de coordinación o de subordinación. Los elementos relacionados han de ser obligatoriamente oraciones en el segundo caso, mientras que en el primero han de ser elementos análogos (oraciones o partes de la oración)[174]. Las conjunciones coordinantes se dividen en copulativas, disyuntivas, adversativas y distributivas; mientras que las subordinantes incluyen, por un lado, la conjunción *que* (*que introduce oraciones sustantivas que funcionan como complemento de verbos, nombres, adjetivos, adverbios y preposiciones*[175]) y, por otro, las conjunciones que introducen las llamadas adverbiales impropias. Estas conjunciones se caracterizan por tener contenido semántico, por expresar relaciones semánticas que también pueden expresarse con preposiciones (o locuciones preposicionales), y por que el elemento que encabezan funciona como adjunto en la oración[176]. Entre las conjunciones que introducen adverbiales impropias cabe señalar los elementos léxicos *si, como, aunque, porque* y, mayoritariamente, locuciones, como por ejemplo *para que, antes de que, siempre que,* etc.

2.2.8.2. Criterios y etiquetación adoptados

Hemos considerado dentro de la **categoría** conjunción aquellas palabras y locuciones morfológicamente invariables capaces de enlazar estructuras oracionales, ya sea subordinándolas, ya coordinándolas, y hemos establecido dos **tipos**: las coordinantes y las subordinantes, siguiendo a Pavón (1999) y Alarcos (1994). Las conjunciones coordinantes no sólo unen elementos oracionales, sino también palabras y/o sintagmas.

En el cuadro 2.28 aparecen las etiquetas destinadas a las conjunciones.

Atributo	Valor	Código
Categoría	Conjunción	C
Tipo	Coordinante	C
	Subordinante	S

Cuadro 2.28: Etiquetas para la conjunción

Al igual que ocurría con los adverbios, hay secuencias que en ocasiones pueden funcionar como locuciones y a veces no, como por ejemplo *esto_es, así_es_que*.

2.2.9. Preposiciones.

2.2.9.1. Definición, clasificación y propiedades

Para Bello las preposiciones son aquellas palabras cuya función es *anunciar* un término con el cual forma el *complemento* o expresión que sirve *para completar la significación de*

[173]Pavón (1999): p. 568.
[174]Pavón (1999): p. 621.
[175]Pavón (1999): p. 621.
[176]Pavón (1999): p. 624.

la palabra a que se agregan[177]. Sobre las relaciones que expresan, señala este autor que mientras algunas se aplican a muy diversas relaciones (por ejemplo *de*), otras, en cambio, expresan tipos de relación muy concretos (como por ejemplo, *sobre*).

Las palabras que son preposiciones son: *a, ante, bajo, con, contra, de, desde, en, entre, hacia, hasta, para, por, según, sin, sobre, tras.* A estas formas se añaden *so, cabe* (anticuado), *mientras, pues.* Estas dos últimas palabras, dice, *dejan a menudo el oficio de preposiciones*[178]. Por último, hay una serie de palabras (a saber: *afuera, adentro, arriba, abajo, adelante, atrás, antes, después*) que *toman el carácter, aunque no el lugar de la preposición, posponiéndose al nombre*[179].

Según el *Esbozo* la preposiciones son *partículas proclíticas (salvo* según*) que encabezan un complemento nominal de otra palabra y lo subordinan a ella*[180]. El inventario de las preposiciones es el siguiente: *a, ante, bajo, cabe, con, contra, de, desde, en, entre, hacia, hasta, para, por, pro, según, sin, so, sobre, tras.*

Las características definitorias de las preposiciones son, según Alcina–Blecua, las siguientes:

– son palabras que expresan más o menos vagamente una relación;
– *marcan a un nombre o constituyente que haga sus veces*
(Alcina y Blecua (1989): p. 827.);
– convierten a este constituyente en complemento de otra palabra;
– no tienen un uso independiente;
– se emplean siempre antepuestas a otra palabra;
– forman grupo acentual con la secuencia siguiente;
– todas las preposiciones (con excepción de *entre, hasta y según*) sólo aceptan las formas *mí, ti* de los pronombres personales de primera y segunda persona del singular.

Las palabras que forman el inventario de las preposiciones son: *a, ante, bajo, con, contra, de, desde, en, entre, hacia, hasta, para, por, según, sin, sobre, tras, excepto, salvo, durante, mediante, obstante, embargante, incluso, inclusive* además de ciertos usos de *cuando* y *donde.*

Según Emilio Alarcos,

> *las* preposiciones *son unidades dependientes que incrementan a los sustantivos, adjetivos o adverbios como índices explícitos de las funciones que tales palabras cumplen bien en la oración, bien en el grupo unitario nominal*[181].

Las palabras que propone como preposiciones son: *a, ante, bajo, con, contra, de, desde, en, entre, hacia, hasta, para, por, sin, sobre, tras. Según* queda excluida del inventario por ser palabra tónica y por poder aparecer aislada.

[177] Bello (1847): § 67.
[178] Bello (1847): § 1182.
[179] Bello (1847): § 1182.
[180] RAE (1973): p. 438.
[181] Alarcos (1994): p. 214.

En Pavón (1999) se define la preposición como *una clase de palabras encargada de establecer una relación de modificación o subordinación entre dos constituyentes*[182]. En de Bruyne (1999) se estudian las preposiciones *ante, cabe, contra, desde, en, entre, hacia, hasta, para, por, sin, so, sobre, tras* desde un punto de vista semántico. Otras preposiciones mencionadas son *a, de* que no se tratan en este capítulo así como las que se llaman *preposiciones dudosas*[183] y que son *pro, salvo, según, vía, versus.*

2.2.9.2. Criterios y etiquetación adoptados

Consideramos preposiciones aquellas palabras pertenecientes a una categoría mayor, morfológicamente invariables y que establecen relaciones entre constituyentes sintagmáticos en el seno de la oración (que son por tanto transitivas).

Las palabras que tenemos etiquetadas como preposiciones son: *a, ad, al, ante, aparte, apud, bajo, cabe, con, contra, de, del, desde, durante, en, entre, excepto, hacia, hasta, mediante, para, por, salvo, según, sin, sobre, so, tras, versus.*

En la etiquetación de las preposiciones hemos tenido en cuenta no sólo unidades léxicas simples, sino también locuciones. Otro elemento destacable es el hecho de que en algunos casos el artículo masculino singular se contrae con la preposición. Por ello, los atributos para las categoría preposición son los que aparecen en el cuadro 2.29.

Atributo	**Valor**	**Código**
Categoría	Adposición	S
Tipo	Preposición	P
Formación	Simple	S
	Contracción	C
Género	Masculino	M
Número	Singular	S

Cuadro 2.29: Etiquetas para las preposiciones

Respetando el estándar de anotación y teniendo en cuenta que en las diversas lenguas hay preposiciones y posposiciones, nos ha parecido pertinente mantener la denominación de *adposición* para el nombre de la **categoría**; sin embargo, consideramos un solo **tipo**, la *preposición.* El tercer atributo es el que da cuenta de las formas simples y de las contracciones, mientras que los dos últimos atributos (**género** y **número**) sólo son válidos para las contracciones y siempre aparecen con los mismos valores: *masculino, singular.*

Ejemplos:
spcms para las contracciones como *al, del a_partir_del, pese_al,* etc,
sps00 para el resto de preposiciones: *por, en, contra, a_partir_de,* etc.

[182] Pavón (1999): p. 567.
[183] de Bruyne (1999): pp. 696–698.

2.2.10. Interjecciones.

2.2.10.1. Definición, clasificación y propiedades

A. Bello considera las interjecciones como una clase de palabras independiente. La define como la *palabra en que parece hacernos prorrumpir una súbita emoción o afecto, cortando a menudo el hilo de la oración*[184].

En el *Esbozo* se mencionan pero no se definen[185]. Se clasifican en propias (como por ejemplo *¡Oh!, ¡Ah!*) y derivadas (como *¡Bueno!, ¡Diablo!*).

Alcina y Blecua (1989) las consideran también como clase de palabras aunque admiten que resulta muy difícil su delimitación. Entre las notas características de las interjecciones destacan estos autores las siguientes:

1. algunas carecen de contenido semántico;
2. pueden enriquecerse con secuencias fonemáticas extrañas al sistema español;
3. necesitan una entonación específica;
4. no tienen función primaria en la enunciación.

Además, al igual que el *Esbozo*, las clasifican en propias o primarias (*ordenaciones de fonemas, sancionadas por el uso e incorporadas a la lengua con cierta fijeza*[186]) e impropias o secundarias (*constituidas por palabras de diversas clases que por transposición se emplean con la misma intención que las anteriores*[187]).

Por su parte, E. Alarcos define la interjección como *una clase de palabras autónomas que, a diferencia de los sustantivos, los adjetivos, los verbos y los adverbios, no se insertan funcionalmente dentro de la oración y constituyen por sí solas enunciados independientes*[188].

En Alonso-Cortés (1999) la interjección de define como

> *una palabra constituida generalmente por una sola sílaba en cuyo ataque y coda pueden aparecer fonemas que no aparecen en final de palabra en el léxico patrimonial, colocada preferentemente en posición inicial, y cuyo significado es enteramente expresivo*[189].

Para este autor las principales interjecciones propias del español actual son:

	bah	*uy*	*ea*	*puf*	*fu*
ay	*eh*	*oh*	*ja*	*bo*	*hum*
aj	*ah*	*ca*	*puaf*	*bu*	*pse*

[184] Bello (1847): § 78.
[185] RAE (1973): pp. 115 y 357.
[186] Alcina y Blecua (1989): p. 820
[187] Alcina y Blecua (1989): p. 820.
[188] Alarcos (1994): p. 240.
[189] Entendiendo por acto de habla expresivo aquel cuyo propósito es manifestar que el hablante se encuentra afectado por algo. Alonso-Cortés (1999): pp. 4025 y 3996 respectivamente.

psche *tota* *uf*
psst *uhy*

Señala que, además, pueden utilizarse también como interjecciones nombres y verbos, denominados entonces *interjecciones impropias* como *caracoles, arrea, canastos.*

A nivel morfológico, las interjecciones son palabras invariables, excepto *ay* que puede lexicalizarse como sustantivo, adquiriendo en ese momento sus características morfológicas.

2.2.10.2. Criterios y etiquetación adoptados

Hemos considerado interjecciones un total de 207 expresiones, que incluyen tanto las interjecciones propias como las impropias. Todas ellas son expresiones invariables, por lo que carecen de atributos morfológicos. Las interjecciones aparecen todas con la etiqueta **I**, como en los siguientes ejemplos:
achís_I
alá_I
bravo_I
buf_I
cáscaras_I

Sin embargo, dado que muchas palabras pueden utilizarse con valor interjectivo, la inclusión de una forma en esta categoría depende mucho del contexto en que se utilice y en el corpus[190] se han etiquetado como interjecciones algunos usos de ciertas palabras.

Las onomatopeyas han quedado también recogidas bajo esta etiqueta.

2.2.11. Abreviaturas.

Consideramos abreviaturas aquellas expresiones lingüísticas, por lo general formadas por más de una palabra, con una forma reducida. Las etiquetamos como **Y**, tal como aparece a continuación:

bps_Y	bites por segundo
kwh_Y	kilovatios/hora
rpm_Y	revoluciones por minuto

Las abreviaturas que corresponden a una sola palabra reciben la etiqueta correspondiente a su categoría gramatical. Así, por ejemplo, *cm* se etiqueta como nombre común, igual que *km, ml* y los símbolos de los elementos químicos.

2.2.12. Puntuación.

Como ya se comentó anteriormente (página 62) se hace necesario etiquetar también los signos de puntuación que aparecen en los textos, ya que su valor no siempre es el mismo. Nuestra propuesta de etiquetación es la que aparece en el cuadro 2.30.

[190] Véase capítulo 3.

Signo		Código	Signo		Código	Signo		Código
,	=>	Fc	.	=>	Fp	“	=>	Fe
...	=>	Fs	:	=>	Fd	"	=>	Fe
;	=>	Fx	%	=>	Ft	’	=>	Fe
-	=>	Fg	/	=>	Fh	–	=>	Fg
‘	=>	Fe	(	=>	Fpa	)	=>	Fpt
¿	=>	Fia	?	=>	Fit	¡	=>	Faa
!	=>	Fat	[	=>	Fca	]	=>	Fct

Cuadro 2.30: Etiquetas para los signos de puntuación

2.2.13. Fechas

2.2.13.1. Criterios y etiquetación adoptados

Detectar en los textos todo lo relacionado con las fechas (día, meses, años, siglos, incluso horas) resulta especialmente interesante, sobre todo para sistemas de extracción de información, o de pregunta–respuesta. Todas ellas recibirán la etiqueta **W**. Estos elementos se tratan en un módulo específico del analizador[191].

Ejemplos de elementos que consideramos fechas extraídos de corpus son horas (ejemplos 2.25 (a-b)), expresiones temporales que incluyen día, mes y año (ejemplos c-d), días de la semana (ejemplo e), siglos (ejemplo f) y años (ejemplos g-h):

(2.25) (a) *el funeral [...] partió de Tombstone a las* **3.30** (t2).
(b) *llegué a la oficina alrededor de las* **cinco** (a26).
(c) *me dijo el* **20 de setiembre de 1917** (t2)
(d) *atada con bramante rojo a una etiqueta: Rafael,* **7_de_febrero_de_1978** (t5).
(e) *comentar algo aparte del Hola_Raffaella del* **jueves** *y del penalti del* **sábado** (a15).
(f) *tales como la invasión musulmana del* **siglo_VIII**, *el descubrimiento de América en el* **XV** *o ...* (a28).
(g) *Elena otra vez en "The_Memory_of_Certains_Persons"(***1947***)* (a1).
(h) *He adaptado como título de estas cuartillas otro título de Erskine: "The_Moral_Obligation_to_Be_Intelligent"(***1921***)* (a1).

2.2.14. Cifras

Las cifras también se tratan en un módulo especializado del analizador morfológico. La etiqueta que reciben es **Z**. Los números escritos en cifras se etiquetan del mismo modo. A continuación aparecen ejemplos tomados de corpus:

(2.26) *Según el cine europeo de los* **60** (a15).

[191] Véase la sección 2.2.

*el ala-pívot Dwight_Stewart (**2,05**); el duro base Corey_Beck y este año el pívot de **2,10** Darnell_Robinson* (d1).
*con el 'efecto **5-0**' añadido* (d1).

En el primero de los ejemplos anteriores, la cifra hace referencia a un año, pero nada en el contexto inmediato permite reconocerlo como tal, por lo que se etiqueta como una cifra.

2.2.15. Unidades monetarias

Las unidades monetarias se tratan también en un módulo específico del analizador y reciben la etiqueta **Zm**, como en el ejemplo 2.27. Es de notar que en estos casos la expresión numérica se une a la unidad monetaria que la sigue en el texto.

(2.27) *Corcuera ha anunciado que no renunciará a la pensión de **663.126_pesetas_Zm** brutas mensuales que cobran los ex_ministros durante dos años* (r2).

Hasta aquí hemos presentado detalladamente las catorce categorías utilizadas: sus rasgos definitorios y el sistema de etiquetación que hemos seguido. Pero MACO no sólo asigna a cada palabra su(s) etiqueta(s); además, realiza la lematización del texto, de la que nos ocupamos en la sección siguiente.

2.3. Lematización

La lematización suele realizarse siempre junto con la asignación de las etiquetas morfológicas. De acuerdo con McEnery y Wilson (1996b) podemos definir la lematización como:

> *Lemmatisation involves the reduction of the words in a corpus to their respective lexemes -the head word form that one would look up if one were looking for the word in a dictionary*[192].

La importancia de la lematización está en el hecho de que facilita las búsquedas basadas en corpus, tanto desde la perspectiva de los estudios de vocabulario (con diversos fines, como por ejemplo para la enseñanza, para estudios estadísticos, etc.) como para la lexicografía.

En nuestro sistema ambos procesos (anotación y lematización) se realizan simultáneamente, ya que en el analizador, a cada palabra se le ha asociado un par lema–etiqueta, de modo que la salida del módulo morfológico es la que aparece en el cuadro 2.31:

Debemos comentar, antes de presentar los criterios de lematización, que en el estadio actual de análisis, no se trata la derivación apreciativa, que es la única que mantiene la categoría morfológica de la palabra primitiva[193].

[192] McEnery y Wilson (1996b): p. 53.

[193] Para el futuro está previsto un módulo, posterior al análisis morfológico estricto, que trate palabras como *casita, casona, casucha; muchísimo,* etc. como derivadas, respectivamente de *casa* y *mucho*, que serían los respectivos lemas.

Forma	$lema - etiqueta_1$	$lema - etiqueta_2$	$lema - etiqueta_3$	$lema - etiqueta_4$
Me	yo PP1CS000	yo P010S000		
gusta	gustar VMIP3S0	gustar VMM02S0		
la	el DA0FS0	la NCFS000	él PP3FSA00	
cultura	cultura NCFS000	culturar VMIP3S0	culturar VMM02S0	
del	del SPCMS			
pelotazo	pelotazo NCMS000			
porque	porque CS			
sacrifica	sacrificar VMIP3S0	sacrificar VMM02S0		
la	el DA0FS0	la NCFS000	él PP3FSA00	
búsqueda	búsqueda NCFS000			
de	de SPS00			
lo	el DA0NS0	lo NCMS000	él PP3MSA00	él PP3CNA00
útil	útil AQ0CS0	útil NCMS000		
en	en SPS00			
favor	favor NCMS000			
del	del SPCMS			
cultivo	cultivo NCMS000	cultivar VMIP1S0		
de	de SPS00	de NCMS000		
lo	el DA0NS0	lo NCMS000	él PP3MSA00	él PP3CNA00
admirable	admirable AQ0CS0			
.	. Fp			

Cuadro 2.31: Salida del módulo de análisis morfológico

Los criterios que hemos seguido para la lematización de las distintas clases de palabras consideradas son los que comentamos a continuación:

1. Nombre

   El lema de los nombres es su forma singular. En la definición de la clase del nombre (cf. 2.2.6.2) hemos comentado que los sustantivos tenían género inherente, esto es, cada nombre tiene un género propio que es el que exige a los elementos que concuerdan con él[194]. Siguiendo este principio, no consideramos que *niña* sea el femenino de *niño*, sino que estamos ante dos nombres distintos, de género gramatical distinto y que, por tanto, lematizamos de modo distinto. Los nombres propios tienen su misma forma como lema (en minúscula) sin diferencias de género ni número. El cuadro 2.32 muestra algunos ejemplos de lematización de sustantivos.

2. Adjetivo

   Los adjetivos presentan morfemas de género y número concordantes, es decir, que en el contexto toman los morfemas que les exige el sustantivo por la concordancia. Por ello, el lema de los adjetivos será siempre la forma masculina singular, tal como se

[194]En este sentido se manifiesta también Marsá (1984).

forma	lema
niño	niño
niños	niño
niña	niña
niñas	niña
paredes	pared
muros	muro
Medardo_Fraile	medardo_fraile

Cuadro 2.32: Lematización de los nombres

muestra en el cuadro 2.33. Si los adjetivos son de dos terminaciones, el lema asignado es la forma singular. Si el adjetivo es de número invariable, esta forma será el lema.

forma	lema	forma	lema
cuatro terminaciones			
alto	alto	alta	alto
altos	alto	altas	alto
primero	primero	primera	primero
primeros	primero	primeras	primero
dos terminaciones			
joven	joven	jóvenes	joven
número invariable			
antidisturbios	antidisturbios	burdeos	burdeos

Cuadro 2.33: Lematización de los adjetivos

Para el caso de los adjetivos que presentan formas apocopadas, como *gran, buen, primer, tercer*, el lema es la forma plena, es decir, *grande, bueno, primero* y *tercero* respectivamente.

3. Verbo

   El lema de todas las formas verbales es el infinitivo sin clíticos (cf. cuadro 2.34).

forma	lema
soñábamos	soñar
reirán	reír
cantados	cantar
saliendo	salir
odiar	odiar
diciéndoselo	decir

Cuadro 2.34: Lematización de los verbos

Los verbos que sólo tienen un uso pronominal, esto es, que aparecen siempre con un

clítico, tienen como lema la forma con el incremento pronominal, tal como se muestra en el cuadro 2.35.

forma	lema
(se) quejó	quejarse
(se) atrevían	atreverse
suicidándose	suicidarse

Cuadro 2.35: Lematización de los verbos pronominales

Cuando una misma forma verbal aparece (aun con cambio de significado) tanto en forma pronominal como en la no pronominal, el lema asignado es siempre el de la forma no pronominal (cf. cuadro 2.36):

forma	lema
(se) dirigió (a)	dirigir
dirigió	dirigir
(nos) acordamos (de)	acordar
acordamos	acordar

Cuadro 2.36: Lematización de los verbos alternantes

El problema aquí es que es imposible establecer claramente la distinción entre ambos verbos, porque la presencia de un clítico ante la forma verbal no asegura que se trate del uso pronominal[195].

Los lemas verbales con forma pronominal son 960.

4. Pronombre personal

   Los lemas correspondientes a las formas de los pronombres personales son las formas singulares de las tres personas *yo, tú, él* (cf. cuadro 2.37). Si no se ha tenido en cuenta el género, como en el caso del nombre, es porque la explicitación de este 'morfema' no es homogénea dentro del sistema pronominal: mientras la tercera persona lo manifiesta en la mayoría de las formas (pero no en todas, como en *le, les, se*), en la primera y la segunda sólo aparece en algunos plurales (*nosotros / nosotras*). Por ello, y para proporcionar una lematización sistemática, sólo se tiene en cuenta la persona singular a la hora de asignar lema a los pronombres. De esta forma también reflejamos el hecho de que hay tres formas pronominales, *yo, tú, él* con variaciones de género, número y caso.

   A la hora de la consulta a corpus o del estudio de los usos de los pronombres, la lematización que proponemos permite estudiar el uso general de los pronombres,

[195] Un ejemplo: *Se acordó una rebaja de impuestos.* Aquí no se trata del verbo *acordarse* sino del verbo *acordar* en una oración pasiva refleja. La única forma de poder hacer esta distinción de un modo automático es conociendo la estructura argumental y el esquema de subcategorización del verbo. Si esta información estuviera disponible en el diccionario, podríamos disponer de dos formas con dos lemas y la desambiguación, en el proceso posterior, debería ser posible.

mientras que si lo que se desea es llevar a cabo estudios más refinados y/o concretos, pueden realizarse búsquedas por formas o por etiquetas.

Sin embargo, hay un caso especial, en concreto la forma **se** para la cual se han establecido dos lemas distintos (**él / se**) según el contexto en que aparece. Esta doble lematización tiene un correlato con las etiquetas asignadas a esta palabra. Si esta forma es un morfema verbal (si aparece en oraciones impersonales o pasivas), entonces el lema es *se* mientras que en los casos restantes, su lema es *él*.

forma	**lema**
yo	yo
nosotros	yo
nos	yo
conmigo	yo
vosotras	tú
te	tú
ella	él
ellos	él
le / les	él
se	él/se
la / lo	él
las / los	él

Cuadro 2.37: Lematización de los pronombres personales

5. Determinativos

Todos los determinativos, tanto adjetivos como pronombres, tienen como lema las formas del masculino singular (o las formas singulares si se trata de determinativos que no distinguen género en el singular). Se tratan igual que los adjetivos.

Casos particulares:

a) Adjetivos determinativos que presentan formas apocopadas

El lema asignado es la forma masculina singular plena, tal como se muestra en el cuadro 2.38.

forma	**lema**
ningún	ninguno
ninguna	ninguno
cualquier	cualquiera
quienquier	quienquiera
algún	alguno

Cuadro 2.38: Lematización de los determinativos (1)

b) Determinativos defectivos

Existen algunos determinativos que sólo presentan formas plurales. Su lema es la forma masculina (plural), tal como se muestra en el cuadro 2.39.

forma	**lema**
sendos	sendos
sendas	sendos
varios	varios
varias	varios

Cuadro 2.39: Lematización de los determinativos (2)

Además, la forma *cada* sólo tiene singular y el lema es la misma forma.

6. Adverbios

Los adverbios tienen como lema su misma forma (cuadro 2.40):

forma	**lema**
ayer	ayer
alegremente	alegremente
amargamente	amargamente

Cuadro 2.40: Lematización de los adverbios

El mismo tratamiento se ha dado a las locuciones adverbiales.

7. Conjunciones

Tanto las conjunciones simples como las compuestas (o locuciones conjuntivas) tienen su misma forma como lema.

8. Fechas

El lema de las fechas es una estructura compleja que recoge toda la variedad de formas en que estas pueden expresarse. Recordemos que los elementos incluidos en esta categoría van desde los días de la semana, a los años aislados, las horas o las secuencias de día–mes–año. La forma del lema para las fechas, excepto en el caso de los siglos, es la siguiente: hay tres campos, separados por **:**. El primero es para los días de la semana; el segundo para los días del mes, los meses y el año; y el último para las horas. En caso de que alguno de estos campos no esté especificado, aparecen los símbolos **??**, tal como aparece en los siguientes ejemplos.
1937 [??:??/??/1937:??.??] W
1992 [??:??/??/1992:??.??] W
1987 [??:??/??/1987:??.??] W
1945 [??:??/??/1945:??.??] W
1993 [??:??/??/1993:??.??] W

octubre_de_1987 [??:??/10/1987:??.??] W
marzo [??:??/03/??:??.??] W
mes_de_octubre [??:??/10/??:??.??] W
23_de_octubre [??:23/10/??:??.??] W
21_de_noviembre [??:21/11/??:??.??] W
24_de_febrero_de_1991 [??:24/02/1991:??.??] W
jueves [jueves:??/??/??:??.??] W
sábado [sábado:??/??/??:??.??] W
20.13 [??:??/??/??:20.13] W

Los siglos se marcan con un lema más breve:
siglo_XIX [s:xix] W

9. Unidades monetarias
El lema de las unidades monetarias lo forman la cantidad, expresada en cifras, seguida de la unidad en singular:
dos_mil_pesetas 2000_peseta Zm
300_dólares 300_dólar Zm

10. Otros elementos:

El lema de los signos de puntuación es el mismo signo. El de las abreviaturas es la propia forma abreviada.

2.4. Datos

En esta sección presentamos los datos numéricos correspondientes a las cantidades de palabras que contiene el diccionario del analizador morfológico.

El número de formas que incluye es de 921051, que se corresponden con 1134441 interpretaciones y con 130390 lemas distintos. Estos datos desglosados por categorías son los siguientes:

120358 palabras reciben la etiqueta de adjetivo; de ellas 242 son ordinales y el resto calificativos.

El número de adverbios es de 234.

El número de formas pronominales del diccionario es de 773, que se dividen en los siguientes tipos[196]:

[196]Recuérdese que los posesivos incluyen el artículo, y que por tanto forman parte del módulo de locuciones, tal como se comentó en la página 49. Las cinco formas que aquí aparecen corresponden a formas masculinas singulares sin artículo que aparecen con las preposiciones contraídas

tipo	**palabras**
demostrativo	33
exclamativo	1
indefinido	60
numeral	605
personal	32
relativo	17
interrogativo	13
posesivo	5

El número de palabras categorizadas como determinantes en el diccionario morfológico es de 742, repartidas según los siguientes tipos:

artículo	5
demostrativo	17
exclamativo	5
indefinido	83
numeral	601
posesivo	26
interrogativo	5

188590 palabras tienen la interpretación de nombre;

El número palabras con etiqueta verbal en el diccionario es de 823425, repartidas en los siguientes tipos: verbo auxiliar 66 formas; verbo semiauxiliar 62; verbos principales 823297.

El número de conjunciones en el diccionario es de 31; 13 de ellas son coordinantes y las 18 restantes son subordinantes. En cuanto a las locuciones (ambiguas o no), las conjuntivas coordinantes son 4 (*sino_que, esto_es, es, es_decir, o_sea*), mientras que las subordinantes son 130.

El número de preposiciones que aparece en el diccionario morfológico es de 30, dos de las cuales son contraídas. En el fichero de locuciones se contemplan 188 preposiciones simples y 58 contraídas.

El número de interjecciones es de 207, y el de abreviaturas 48.

2.5. Conclusión

El sistema de etiquetado propuesto ha sido probado y validado mediante la etiquetación del corpus **CLiC-TALP**. Los criterios que han guiado la elaboración del sistema han sido la máxima sencillez y la mínima redundancia en la información de las etiquetas. Asimismo, se ha tratado de mantener un equilibrio entre lo que propone la teoría gramatical y lo que un sistema de análisis automático puede tratar.

Por otra parte, en la definición de las categorías morfosintácticas se ha tratado de seguir un criterio puramente formal, guiado por la morfología y la sintaxis, de las clases de palabras, y dejando al margen, cuando ello ha sido posible, el aspecto semántico.

Las diferencias más importantes entre la versión aquí propuesta del diccionario generado y la anteriormente existente afectan a la mayoría de categorías. Los objetivos fundamentales

de esta revisión han sido, por una parte, establecer unos criterios de clasificación claros y, por otra, reducir en la medida de lo posible la ambigüedad existente. Las categorías más complejas han resultado ser las categorías cerradas, especialmente, los determinantes, pronombres, numerales y artículos, donde desde un punto de vista teórico hay grandes divergencias. La categoría *adjetivo* incluía todas las formas de los determinativos, además de los calificativos, mientras que ahora sólo incluye los calificativos y los ordinales. En la categoría de *nombre* se han introducido nuevas palabras, especialmente aquellas que pueden tener una doble función nominal y adjetiva que anteriormente sólo aparecían como adjetivos. En lo que al verbo respecta, se ha introducido la distinción entre principal, auxiliar y semiauxiliar, donde antes sólo se consideraba principal y auxiliar. Se han tratado también los verbos pronominales. Las clases de los determinantes y los pronombres se han sistematizado y han pasado a incluir los numerales no ordinales y los artículos, que anteriormente aparecían como una categoría aparte. Las clases correspondientes a las palabras invariables se han revisado completamente, introduciendo nuevas palabras y sistematizando su adscripción a las distintas categorías: preposición, conjunción y adverbio.

Capítulo 3

Desambiguación morfológica

El proceso de análisis morfológico asigna a cada palabra en un texto todos los pares lema–etiqueta posibles. La desambiguación morfológica consiste en seleccionar, de entre todos esos pares, el que corresponde según el contexto. Las palabras, en el texto, no son ambiguas: dado un contexto para una palabra, a ésta le corresponde una y sólo una etiqueta morfológica y análogamente un solo lema.

En este capítulo se tratan dos aspectos del proceso de desambiguación. Por un lado se presenta la problemática de la validación manual para la construcción de un corpus. Este corpus, el corpus **CLiC-TALP**, constituye un *gold-standard* para sistemas de desambiguación automáticos basados en métodos estadísticos y de aprendizaje automático, así como un valioso recurso para los estudios estrictamente lingüísticos del español. Por otro lado, se presentan las reglas de desambiguación basadas en conocimiento lingüístico introducidas en *RELAX*, un sistema de desambiguación automática basado en aprendizaje que admite la introducción manual de reglas cuya implementación ha significado una mejora del 2 % en el nivel de calidad del desambiguador estadístico.

La figura 3.1 sitúa este trabajo en el marco de los procesos de análisis del lenguaje de CLiC-TALP.

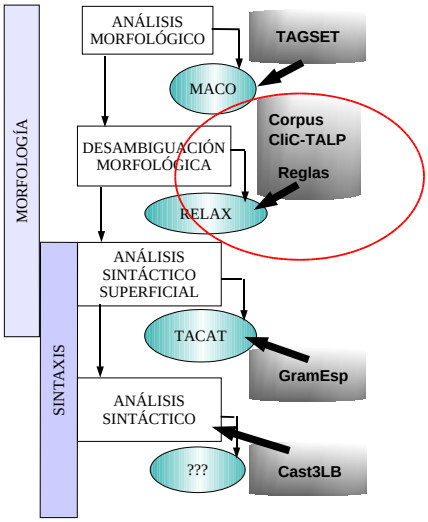

Figura 3.1: Procesos de análisis (2): corpus desambiguado y restricciones

3.1. Introducción

El proceso de desambiguación morfológica consiste en seleccionar, de entre todas las etiquetas posibles asignadas a una unidad léxica en el texto, aquélla que se corresponde con el uso de esa unidad en el contexto concreto en que aparece. Análogamente, si la lematización se realiza de modo simultáneo con la etiquetación, también se selecciona el lema adecuado para esa palabra, dado el contexto en que aparece.

A continuación aparece la frase *Como mi amiga la Susi no se anda con chiquitas, cuando le pregunté qué le haría ilusión como regalo de Reyes me dijo: Algo caro y que me dé mucho gusto* (c1). En primer lugar (cuadro 3.1), tras el proceso de análisis morfológico, con todos los pares lema–etiqueta posibles asignados a cada palabra. La primera columna corresponde a las formas del texto. En las columnas restantes, que deben leerse de dos en dos, aparecen los posibles pares lema–etiqueta para cada palabra. En segundo lugar (cuadro 3.2) aparece la misma oración tras el proceso de desambiguación.

Al igual que en el caso de la anotación morfológica, el proceso de desambiguación puede realizarse de modo manual, automático o semiautomático.

Si el proceso es **automático**, los criterios de desambiguación pueden consistir en reglas contextuales basadas en conocimiento lingüístico e introducidas por el lingüista (como por ejemplo las *Constraint Grammars* aplicadas al inglés o al euskera ((Karlsson et al., 1995), (Aduriz et al., 1997), (Aduriz, 2000)) o bien inducidas automáticamente a partir de corpus desambiguados de modo manual ((Padró, 1998), (Màrquez, 1999), por ejemplo). De una u otra forma, lo que se obtiene es un modelo del lenguaje que luego puede aplicarse para la desambiguación de nuevos textos. Según aparece en (Màrquez, Padró, y Rodríguez, 2001) las ventajas de los desambiguadores basados en conocimiento lingüístico, y en particular del sistema ENGCG (*English Constraint Grammar*) de Karlsson et al. (1995), son (i) la

```
Como comer VMIP1S0 como CS como PR000000
mi mi DP1CSS mi NCMS000
amiga amiga NCFS000 amigar VMIP3S0 amigar VMM02S0 amigo AQ0FS0
la el DA0FS0 la NCMS000 él PP3FSA00
Susi susi VMM0000 susi VMN0000 susi NC00000 susi AQ00000 susi VMP0000
no no NCMS000 no RN
se se P0000000 él P0300000 él PP3CN000
anda anda NCFS000 andar VMIP3S0 andar VMM02S0
con con SPS00
chiquitas chiquita NCFP000 chiquito AQ0FP0
, , Fc
cuando cuando CS cuando NCMS000 cuando PR000000
le él PP3CSD00
pregunté preguntar VMIS1S0
qué qué DE0CN0 qué DT0CN0 qué PE000000 qué PT0CS000 qué RG
le él PP3CSD00
haría hacer VMIC1S0 hacer VMIC3S0
ilusión ilusión NCFS000
como comer VMIP1S0 como CS como PR000000
regalo regalar VMIP1S0 regalo NCMS000
de de NCFS000 de SPS00
Reyes rey NCMP000 reyar VMSP2S0 reyes NP00000
me yo P010S000 yo PP1CS000
dijo decir VMIS3S0
: : Fd
Algo algo PI0CS000 algo RG
caro caro AQ0MS0 caro NCMS000
y y CC y NCFS000
que que CS que PR0CN000
me yo P010S000 yo PP1CS000
dé dar VMM03S0 dar VMSP1S0 dar VMSP3S0
mucho mucho DI0MS0 mucho PI0MS000 mucho RG
gusto gustar VMIP1S0 gusto NCMS000
. . Fp
```

Cuadro 3.1: Oración analizada morfológicamente

Como como CS	ilusión ilusión NCFS000
mi mi DP1CSS	como como CS
amiga amiga NCFS000	regalo regalo NCMS000
la el DA0FS0	de de SPS00
Susi susi NP00000	Reyes reyes NP00000
no no RN	me yo PP1CS000
se él P0300000	dijo decir VMIS3S0
anda andar VMIP3S0	: : Fd
con con SPS00	Algo algo PI0CS000
chiquitas chiquita NCFP000	caro caro AQ0MS0
, , Fc	y y CC
cuando cuando CS	que que PR0CN000
le él PP3CSD00	me yo PP1CS000
pregunté preguntar VMIS1S0	dé dar VMSP3S0
qué qué PT0CS000	mucho mucho DI0MS0
le él PP3CSD00	gusto gusto NCMS000
haría hacer VMIC3S0	. . Fp

Cuadro 3.2: Oración desambiguada

gran expresividad del lenguaje de representación que permite modelizar adecuadamente los fenómenos lingüísticos necesarios así como interpretar las reglas construidas; (ii) la posibilidad de construir sistemas muy precisos; (iii) la facilidad de modificación y de incorporación

de nuevo conocimiento; y (iv) la posibilidad de integración del análisis sintáctico superficial juntamente con la desambiguación. Sus principales inconvenientes son (i) el elevado coste de la construcción del modelo de reglas; (ii) la poca transportabilidad de una lengua a otras; y (iii) el hecho de que puede ser poco robusto si el modelo que se ha construido no es lo suficientemente completo.

Por lo general, la información que se utiliza para la desambiguación de una palabra dada es su contexto. Contexto se entiende aquí como la palabra anterior o la posterior, la etiqueta de la palabra anterior o de la posterior, o incluso combinaciones de palabras y etiquetas anteriores y posteriores. Por lo general el contexto es bastante local y no suele abarcar más de tres elementos a la derecha o a la izquierda. La mayoría de sistemas automáticos trabajan con combinaciones de dos o tres elementos (bigramas y trigramas, respectivamente).

Otras veces, el proceso es **semiautomático**. Así, por ejemplo, el proceso de anotación morfológica del corpus *Le Monde* en francés (Abeillé, Clément, y Kinyon, 2003) se llevó a cabo en dos fases: una primera fase automática con un conjunto de etiquetas reducido respecto del conjunto total empleado, seguida de una segunda fase de desambiguación manual con otro subconjunto de etiquetas más específicas.

Finalmente, algunos corpus, como el *Susanne* se han anotado de modo totalmente **manual** (Sampson, 1995). Para el español cabe destacar el trabajo de Martín (1999), que consiste en la anotación manual de un corpus.

Los desambiguadores automáticos presentan una tasa de error que no suele superar el 10 %. En esta tarea se puede alcanzar un grado de precisión elevado y un error del 5 % es aceptable. Estos datos quedan reflejados en el siguiente cuadro, tomado de McEnery y Wilson (1996a): p. 141[1]

Referencia	**Tasa de error (%)**
Brill (1992)	5
Cutting *et al.* (1992)	4
De Rose (1991)	4
Garside (1987)	4
Greene and Rubin (1971)	23
Voutilainen (1995)	0.7

La tasa de error de estos sistemas no tiene un valor absoluto, ya que no pueden es-

[1]Las referencias mencionadas corresponden a:
Brill, E. (1992) 'A simple rule-based part-of-speech tagger', en *Proceedings of the Third Conference on Applied Natural Language Processing* (ANLP-92), Trento, Italia.
Cutting, D., Kupiec, J., Pedersen, J. y Sibun, P. (1992) 'A practical part-of-speech tagger', en *Proceedings of the Third Conference on Applied Natural Language Processing* (ANLP-92), Trento, Italia.
De Rose, S. (1991) 'An analysis of probabilistic grammatical tagging methods', en Johanson and Stenström (eds.) *English Computer Corpora: Selected Papers and research Guide*, Berlin: Mouton de Gruyter
Garside, R. 'The CLAWS word-tagging system' en Garside, Leech y Sampson (eds.) (1987) *The Computational Analysis of English: A Corpus Based Approach*, London: Longman.
Greene, B. y Rubin, G. (1971) *Automatic Grammatical Tagging of English*, Technical Report, Department of Linguistics, Brown University, R.I.
Voutilainen, A. (1995) 'A syntax-based part of speech analyser' en *Proceedings of the 7th EACL.*

tablecerse comparaciones sin tener en cuenta el tipo de etiquetas utilizado, su número o la complejidad morfológica de la lengua en cuestión. Así, las comparaciones entre resultados no siempre son equitativas, dado que por lo general no suele especificarse el número de etiquetas que se han utilizado. Desambiguar con un conjunto de 3030 etiquetas morfológicas distintas como las utilizadas en el *Prague Dependency Treebank*[2] es mucho más difícil que hacerlo sobre las 285 del corpus *CLiC-TALP*[3], sobre las 150 del *Susanne Corpus*[4] o sobre las 48 utilizadas en el corpus del *Wall Street Journal*.

En segundo lugar, los corpus sobre los que estos sistemas se han evaluado tampoco son siempre los mismos, por lo que la comparación entre ellos se hace muy difícil. La situación la resume Padró (1998), p.85:

> *The reported accuracy of a tagger does not depend only on the tagset and the train and test corpora size, but also on the corpus itself, specially on its ambiguity ratio and on how the tagger behaves over errors in the test corpus. [...] This makes very difficult to compare systems, since they must be trained and evaluated in the same corpora to be comparable.*

Por otra parte, hay que señalar que para esta tarea es más eficiente la desambiguación automática que la puramente manual. En esta última, los errores no suelen ser consistentes y sus causas pueden ser muy variadas. En cambio, los desambiguadores automáticos realizan una anotación coherente y los errores son previsibles y, por tanto, susceptibles de ser corregidos.

Los sistemas de desambiguación basados en aprendizaje automático necesitan corpus previamente anotados de modo manual para inferir el modelo de lenguaje que se aplicará posteriormente. Por todo ello es de gran interés la documentación de los criterios seguidos para la desambiguación. Tanto si las reglas de desambiguación se escriben manualmente como si se aprenden o inducen a partir de corpus desambiguados de modo manual, los criterios de desambiguación (bien para las reglas, bien para el corpus) deben establecerse a priori. Por lo general, estos criterios se basan en la teoría lingüística, que desde sus inicios se ha ocupado de definir las clases de palabras. Pero en el uso, esas definiciones no siempre son de aplicación directa y, además, en ocasiones las teorías hacen consideraciones contradictorias o muy divergentes sobre el estatus de determinadas palabras en determinados contextos. Así por ejemplo, las clases de palabras que las distintas gramáticas han considerado no siempre han sido las mismas[5]. Por otro lado, durante el proceso de definición de los criterios de desambiguación se hace a veces patente que determinadas distinciones establecidas a nivel teórico resultan inoperativas en la práctica. Tal es el caso de la distinción entre el indicativo y el subjuntivo en francés, tal como detalla Véronis (2001a), dado que, si bien la diferenciación teórica está bien fundamentada, lo cierto que es que a un nivel puramente formal, los verbos de la primera conjugación (la que contiene un mayor número de elementos) comparten cuatro de las seis formas entre ambos modos. En este caso, la

[2] Bemova et al. (1999).

[3] Civit, Castellón, y Martí (2001a).

[4] Sampson (1995).

[5] Recuérdese todo lo mencionado en el capítulo 2 a propósito de las distintas clases de palabras en español.

distinción automática es muy difícil, y, por tanto, hay que buscar un equilibrio entre la descripción estrictamente lingüística y su aplicación.

En este capítulo presentamos los criterios utilizados para la fase de validación manual del corpus **CLiC-TALP** así como una serie de restricciones introducidas para mejorar el rendimiento del desambiguador automático RELAX (Padró, 1998).

Dado que las decisiones que se toman son específicas de cada lengua, lo que hemos intentado en este trabajo ha sido explicitar el conocimiento utilizado y evaluar las decisiones tomadas de modo que puedan utilizarse como referencia para otros trabajos, sin que ello implique un acuerdo o aceptación totales respecto de la decisión tomada, sino simplemente una propuesta o punto de partida.

3.2. RELAX

Como se comentó al inicio del capítulo, la mayoría de desambiguadores automáticos necesitan texto validado manualmente para inferir el modelo de lenguaje que les permitirá realizar la desambiguación. Por lo general, para el proceso de aprendizaje del desambiguador se utiliza una parte del corpus validado y se reserva una parte menor para el test, es decir, para la comprobación. En el caso de RELAX, se ha utilizado el 70 % del corpus **CLiC-TALP** para entrenamiento del sistema y el 30 % restante para el test o comprobación de los resultados[6]. En esta sección describimos brevemente este desambiguador.

RELAX es un desambiguador morfológico que utiliza la técnica de *relaxation labelling*. Ésta es una técnica de optimización para resolver problemas de satisfacción de restricciones (Padró (1996a), Padró (1996b) y Padró (1998)). El algoritmo halla una combinación de valores para un conjunto de variables tal que satisface, en el máximo grado posible, un conjunto de restricciones. Cada restricción es un conjunto de pares *variable–etiqueta* con un valor de compatibilidad asociado, que determina cuán compatible es este par (para más detalles, puede consultarse Civit, Martí, y Padró (2003)).

El algoritmo de relajación consiste en:

1. iniciar una asignación aleatoria de peso (probabilidades léxicas);
2. para cada variable, calcular la compatibilidad entre peso actual y los pesos actuales de las otras variables, dado un conjunto de restricciones;
3. aumentar los pesos de las etiquetas más compatibles con el contexto y disminuir el de las menos compatibles;
4. iterar el proceso hasta que los pesos ya no cambian.

El modelo de lenguaje que se obtiene es un lenguaje de restricciones, capaz de expresar el mismo tipo de patrones del formalismo de *Constraint Grammar* ((Karlsson et al., 1995))

[6]Recuérdese que uno de los objetivos de la creación de este corpus es que este desambiguador pueda inferir un modelo de lenguaje para la desambiguación.

pero aumentado para permitir que cada restricción tenga un valor de compatibilidad que indique su peso.

Estas restricciones pueden adquirirse con métodos estadísticos o bien pueden introducirse de modo manual. Para adquirir las restricciones de modo automático se necesita disponer de corpus anotado y validado manualmente. Tras la validación manual del corpus **CLiC-TALP** se realizó dicha adquisición. El modelo estadístico trabaja solamente con una ventana de dos elementos del contexto, lo que quiere decir que establece las restricciones en función de la contigüidad de dos elementos. El peso de las restricciones puede ser positivo (las categorías son compatibles) o negativo (las categorías son incompatibles). El total de restricciones aprendidas automáticamente es de 2906.

Comentamos a continuación dos de estas restricciones para mostrar su funcionamiento. La primera controla la coaparición de un determinante demostrativo seguido de un adjetivo ordinal, mientras que la segunda está relacionada con la coaparición de formas verbales.

La figura 3.2 muestra una de las restricciones aprendidas de modo automático, que afecta a la combinación de un determinante demostrativo (**DD**) que precede (posición -1) al ordinal (etiqueta **AO**) y el peso asignado a esta combinación es 1.19473621726138. Dado que es un peso positivo, ello significa que esta secuencia de palabras (determinante demostrativo seguido de adjetivo ordinal) es muy compatible.

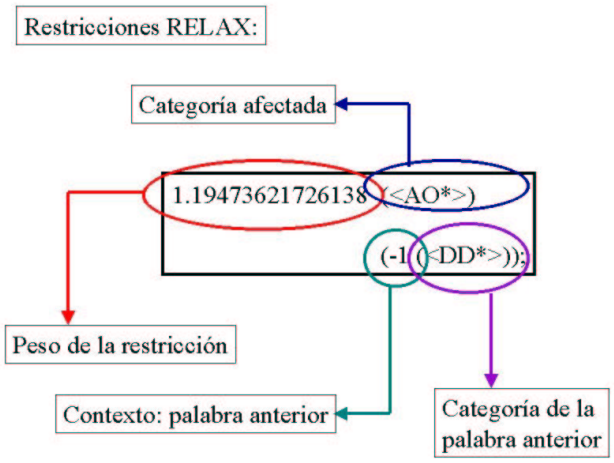

Figura 3.2: Restricciones (1)

Es de notar que las restricciones automáticas se aprenden sólo sobre las categorías y sus tipos. Así, como puede observarse en el ejemplo anterior, sólo se tienen en cuenta los dos primeros dígitos de las etiquetas morfosintácticas: **AO*** se aplica a cualquier etiqueta que empiece por estos dígitos: *AO0MS0, AO0FS0, AO0MP0, AO0FP0.*

El caso de los verbos, sin embargo, es algo distinto y se utilizan los tres primeros dígitos de la etiqueta, los que indican la categoría, el tipo (principal, auxiliar, semiauxiliar) y el modo. Por ejemplo, la siguiente restricción da un peso muy negativo a la combinación

de un verbo principal en participio (**VMP***) seguido de un verbo principal en indicativo (**VMI***) cualesquiera que sean el tiempo, la persona, el número o el género:

```
-5.540331895882  (<VMP*>)
                 (1 (<VMI*>));
```

Otra importante observación sobre el desambiguador automático es que las restricciones que utiliza sólo tienen en cuenta la compatibilidad entre pares de etiquetas (bigramas).

3.2.1. Resultados

Los resultados del desambiguador se han calculado atendiendo a dos criterios de evaluación: por una parte lo que llamamos *etiqueta corta* (**EC**) que se corresponde con los primeros dígitos de las etiquetas; por otro, a la *etiqueta larga* (**EL**), que se corresponde con la etiqueta con todos sus dígitos. **EC** hace referencia a las ambigüedades intercategoriales, mientras que **EL** incluye además la ambigüedad intracategorial[7]. Este modo de proceder nos permitirá, cuando se introduzcan las restricciones manuales, comparar los resultados de modo más apropiado con la salida del desambiguador automático. Para cada uno de estos elementos se ha calculado también el acierto respecto del lema (**L**) de la palabra, de modo que en los resultados se muestran siempre cuatro apartados: etiqueta corta (**EC**); etiqueta corta y lema (**EC+L**); etiqueta larga (**EL**); y, etiqueta larga y lema (**El+L**).

Si aplicamos las 2906 restricciones aprendidas automáticamente al propio corpus de entrenamiento, obtenemos los resultados que se muestran en el cuadro 3.3:

EC	97.29 %
EC+L	96.53 %
EL	94.48 %
EL+L	94.36 %

Cuadro 3.3: Resultados de RELAX con el corpus de entrenamiento

3.3. Validación manual del corpus CLiC-TALP

La validación manual de un corpus es una tarea que se lleva a cabo de modo puntual con dos objetivos principales. Por una parte, que el resultado pueda ser útil para la investigación lingüística y, por otra, que pueda utilizarse para que un desambiguador automático pueda inferir un modelo de lenguaje que luego utilizará para desambiguar otros textos. En las aplicaciones de PLN, la fase de desambiguación revisada manualmente queda restringida a la creación de corpus concretos, pero la desambiguación automática es la que se aplica de modo general.

Como ya se comentó en el capítulo 1, la desambiguación morfológica en el marco CLiC-TALP se desarrolla en distintas fases de procesamiento del corpus. En la fase de análisis

[7] Véanse las secciones 3.3.4 y 3.3.5 para más detalles sobre esta distinción.

morfológico se asignan todas las etiquetas morfosintácticas posibles; en la fase de desambiguación automática se selecciona una de las etiquetas en función del contexto en que la palabra aparece. En la fase de validación manual del corpus **CLiC-TALP** se han validado y corregido, cuando era necesario, las propuestas del desambiguador automático RELAX y se han introducido pequeños cambios para las etiquetas de algunas palabras (cf. sección 3.3.8). Finalmente, en el análisis por cadenas o *chunks* (cf. capítulo 4) se lleva a cabo la última fase de desambiguación, la que afecta al género y al número de los nombres y adjetivos

La figura 3.3 muestra de forma esquemática los procesos de desambiguación. Por una parte, la desambiguación automática es la que se lleva a cabo en todos los procesos de PLN[8]. En cambio, la desambiguación manual de corpus se hace de modo puntual para obtener corpus de entrenamiento para los desambiguadores así como un corpus de referencia.

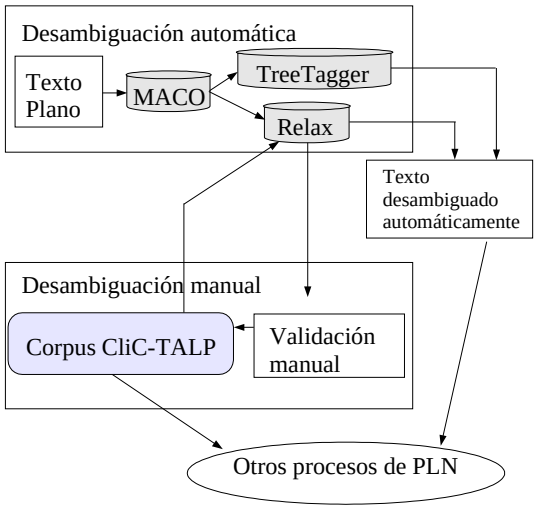

Figura 3.3: Desambiguación automática y manual

Desde el grupo Eagles se realizó una propuesta para la anotación de corpus: *Recommendations for the Morphosyntactic Annotation of Corpora*[9]. Ésta está pensada para aquellos casos en que el corpus se anota directamente, de modo manual, y por tanto, difiere de la de (EAGLES, 1996c) pensada además para lexicones.

Las principales diferencias entre la propuesta para etiquetar lexicones (EAGLES, 1996c)[10] o corpus (EAGLES, 1996b) se hallan en el número de categorías que proponen. En el primer caso se proponen sólo doce categorías principales frente a las catorce de la otra propuesta (quedan excluidos los signos de puntuación y la categoría *único / inasignado*). Además, las categorías *pronombre* y *determinante* quedan claramente diferenciadas dado que

[8]En la actualidad, el grupo CLiC-TALP dispone de dos desambiguadores automáticos, RELAX, presentado anteriormente, y TREETAGGER, basado en árboles de decisión (véase Màrquez (1999)).

[9]EAGLES (1996b).

[10]Presentada en el capítulo 2.

> *it seemed better to distinguish between different functions and, therefore, to have different categories for Pronouns and Determiners, at least at the lexical level. Lexical descriptions should be independent from applications and should aim at a general description of each language; corpus tags, depending on the capabilities of state-of-art tagging techniques, may underspecify lexical specifications, collapsing many distinctions and presenting broader categories* (EAGLES, 1996c).

En EAGLES (1996b) se presentan como una única categoría pronombres y determinantes, aunque se deja a criterio de los diseñadores del sistema de anotación el separarlas en dos categorías o mantenerlas unidas.

Sobre las subcategorías, es de destacar el hecho de que en EAGLES (1996b) las adposiciones presentan un único subtipo: *preposición*; y que los adverbios ya no están subdivididos entre *general* y *partícula*.

Otra de las diferencias destacables afecta a los atributos de género y número. Mientras en EAGLES (1996c) se propone la existencia de un género común y un número invariable, con que se recomienda etiquetar en los lexicones aquellas palabras que en la lengua no manifiestan de modo diferenciado la oposición masculino–femenino o singular–plural[11], en EAGLES (1996b) estos valores desparecen en el caso del español, dado que en un corpus, en un contexto determinado, esta indiferenciación desaparece: *posible* es femenino en 3.1 (a) y masculino en 3.1 (b)

(3.1) (a) *¿Es usted propiedad de un varón y, en_consecuencia, debo abstenerme de mirarla como* **posible** *compañera de esparcimientos sexuales?* (a11).
(b) *Deja caer alguna que otra alusión al* **posible** *idilio de Tristana* (a25).

Estas diferencias son debidas al hecho de que los objetivos que se plantean EAGLES (1996b) y EAGLES (1996c) son distintos. En el primer caso se trata de etiquetar corpus, manifestaciones de la lengua, en que las palabras tienen un uso determinado y se hallan en un contexto concreto, mientras que en el segundo se describen las palabras de la lengua aisladas de todo contexto, en un nivel abstracto, representadas en un lexicón.

3.3.1. El corpus CLiC-TALP

El **corpus CLiC–TALP** consta de 100.000 palabras tomadas del corpus LexEsp: *Léxico informatizado del español*, (Sebastián et al., 2000)[12].

El corpus CLiC-TALP se encuentra distribuido en diferentes archivos cuyos nombres contienen una letra inicial seguida de un número. La letra se corresponde con las distintas fuentes utilizadas, de modo que es posible conocer el tipo de texto. En el cuadro 3.4 se presenta la relación entre el nombre del archivo y su contenido.

El proceso de validación manual ha consistido en que una persona validaba las propuestas del tagger y otra revisaba este trabajo. Además, y con una periodicidad semanal, se discutían los criterios para la validación, con el objetivo de establecer la guía para la anotación manual de corpus (Civit, 2000).

[11] Éste sería el caso, en español, de *cantante, posible* y de *crisis, anticalvicie*, por ejemplo.
[12] Véase la sección 1.4.1.5 del capítulo 1.

a	articulistas
e	ensayo
d	prensa deportiva
dc	divulgación científica
c	suplementos de ciencia
ed	editoriales
n	noticias
r	semanarios
t	narrativa

Cuadro 3.4: Fuentes de CLiC-TALP

Para llevar a cabo la tarea de validación manual de la desambiguación se ha utilizado un interfaz que facilita la tarea al anotador lingüista. Por una parte, se ofrece en primer lugar la etiqueta propuesta por el tagger. Si el anotador considera que es la correcta no tiene más que confirmarla; en caso contrario, debe elegir una de las siguientes en la fila. Si ninguna de éstas es la adecuada, puede seleccionarla de entre el total de etiquetas utilizando un menú desplegable. Este modo de proceder tiene la ventaja de mantener la coherencia en el uso de las etiquetas y se evitan errores causados por la introducción manual de las mismas. La figura 3.4 muestra este interfaz para la anotación manual de la frase: *Apenas despertó aquella mañana, la lengua seca, y firme contra el paladar, se asustó del sabor nauseabundo de su propio aliento. Acercó lentamente a su cara la palma de...*[13].

3.3.2. Segmentación de palabras

Uno de los primeros problemas de la validación manual del corpus surge con la segmentación de las unidades léxicas, especialmente al tratar aquellas secuencias de palabras que pueden, dado un contexto, considerarse locuciones o expresiones multipalabra, pero que en un contexto diferente deben analizarse por separado. Este fenómeno se produce especialmente con lo que hemos denominado *las locuciones ambiguas*. Entendemos por *ambiguas* aquellas secuencias de palabras que en ocasiones pueden funcionar por separado y en otras como una única unidad léxica[14]. También resultan problemáticos los nombres propios, dado que en ocasiones se tratan como unidades secuencias que no lo son.

Comentamos en primer lugar lo referido a las locuciones propiamente dichas y dejamos para el final de esta sección lo concerniente al tratamiento de los nombres propios.

En este nivel de análisis morfológico no hemos considerado como unidades léxicas ni los tiempos compuestos de la conjugación verbal ni las perífrasis verbales. La agrupación se lleva a cabo en el siguiente nivel de análisis[15].

[13]Como puede observarse, las etiquetas morfosintácticas no se corresponden exactamente con las presentadas en el capítulo 2, ya que la figura corresponde a la fase previa al cambio en la etiquetación morfológica.

[14]Recuérdese a este respecto lo comentado en el capítulo 2 página 49 a propósito del módulo de locuciones del analizador.

[15]Cf. capítulo 4 para más detalles sobre este tema.

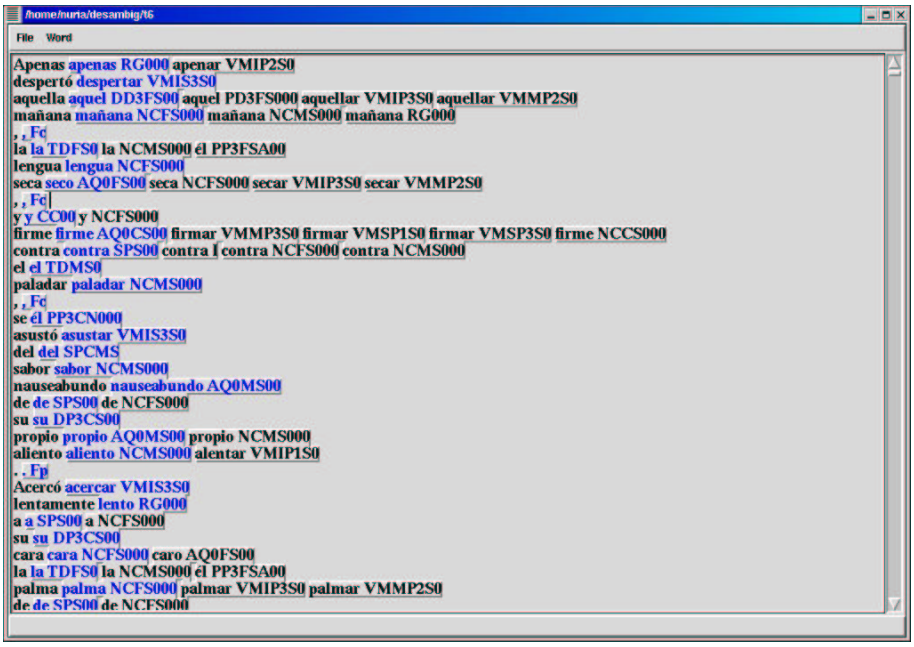

Figura 3.4: Interfaz para la validación manual

3.3.2.1. Tratamiento de las locuciones

Según Pavón (1999): *Podemos definir 'locución' como la expresión constituida por varias palabras, con una forma fija, que se utiliza en el habla como una pieza única y que presenta el comportamiento típico de una determinada categoría gramatical*[16]. Tomamos además el sentido más estricto de esta definición: *que esta no posea una estructura interna productiva, es decir, que los elementos que la componen no encabecen sus propios sintagmas.* A pesar de ello, y como también comenta la propia autora, *podemos encontrar diversos grados de fijación y gramaticalización.*

Según esta autora, los criterios para considerar que una expresión es una **locución preposicional** son en síntesis los siguientes: (i) la no existencia de un sintagma nominal; (ii) la fijación y cohesión interna de la locución; (iii) el comportamiento sintáctico paralelo al de las preposiciones (p. 579).

En lo referente a los **adverbios**, la autora señala que *la mayoría de las locuciones adverbiales están formadas a partir de una preposición que tiene como término un nombre, el cual puede estar a su vez modificado por determinantes y/o adjetivos y otros complementos [...] Una gran parte de ellas son invariables, todas están fuertemente cohesionadas y el nombre que las integra no admite expansiones propias de los sintagmas nominales* (p. 614).

Finalmente, los criterios para determinar la existencia de las **locuciones conjuntivas** no están tan claros como en los otros casos, fundamentalmente debido al hecho de que

[16] Pavón (1999): p. 568.

casi siempre aparece la conjunción *que* y, por lo general, admiten las dos interpretaciones: como locución o como una subordinada sustantiva (encabezada por *que* que es término de la preposición precedente o complemento del adverbio anterior)[17].

En nuestra definición de locución hemos seguido un criterio restrictivo, de modo que consideramos locuciones aquellas expresiones que presentan unidad morfológica, sintáctica y semántica. Además y como norma general, hemos considerado que si aparecía una locución, ésta debía tener la forma más larga posible, esto es, que debía ser la secuencia más larga de palabras. Así, por ejemplo, la secuencia *en el caso de que* se toma toda ella como una locución[18].

A continuación presentamos algunos casos representativos de esta problemática, todos ellos incluidos en la categoría de locución ambigua.

1. con_que / con que

   En el primer caso se trata de una locución conjuntiva de tipo condicional, equivalente a *con tal (de) que* (ejemplo 3.2 (a)), mientras que en el segundo se trata de una preposición seguida de un pronombre relativo (ejemplo 3.2 (b)) o de una conjunción (ejemplo 3.2 (c)):

   (3.2) (a) **Con_ que_cs** *hubieras dicho algo, todo se habría solucionado*[19].
   (b) *De esta manera cada uno de nosotros evitará que hectáreas y hectáreas de bosque sean diezmadas cada día para fabricar esos palillos pequeñoburgueses* **con_sps00 que_pr0cn000** *urgamos entre las caries* (a15).
   (c) *... el personal espera que en cualquier momento el señor Pujol salga* **con_sps00 que_cs**, *o no hay pasta para el Liceo, o de apoyo al gobierno ni mijita* (c1).

2. esto_es / esto es

   La locución conjuntiva **esto_es** (ejemplos 3.3 (a-b)) tiene un valor explicativo; suele aparecer entre comas y puede sustituirse por *es_decir*. **Esto es** (ejemplos 3.3 (c-d)) es una expresión compuesta por un pronombre demostrativo y el verbo *ser*:

   (3.3) (a) *Probablemente zamba,* **esto_ es_CC**, *mestiza de...* (a14).
   (b) *Por otra parte la cultura podía ser una adquisición acumulativa,* **esto_es_cc**, *una herencia social* (e2).
   (c) *Jesús,* **esto_pd0ns000 es_vsip3s0** *un baño turco* (t5).
   (d) **Esto_pd0cs000 es_vsaip3s0** *lo que se llama aprendizaje hebbiano* (dc1).

   Se encuentran en un caso similar expresiones como *es decir* y *o sea* que pueden ser locuciones conjuntivas coordinantes (ejemplos 3.4 (a-b)), o bien otras partes de la oración: *es_vsip3so decir_vmn0000*, *o_cc sea_vssp3s0* (ejemplos 3.4 (c-d)) .

[17] Pavón (1999): pp. 630–643.
[18] En vez de separarla por ejemplo en *en_el_caso de_que* o *en_el_caso_de que*.
[19] Ejemplo tomado de Pavón (1999): p. 631.

(3.4) (a) *Y como el Mendicutti anda en_vilo por lo del nombre, una servidora, la Susi, está tresanchantada de darle a la pen,* **o_sea_cc**, *a la pluma* (c1).
(b) *Y, sin embargo, qué espantoso ese 'aquí',* **es_decir**, *'a mi lado y calladita'* (a11).
(c) *Falso* **es_vsip3s0 decir_vmn0000** *lo que es la belleza*
(d) *Sea cierto* **o_cc sea_vssp3s0** *falso, el problema sigue existiendo*[20].

Por lo general, estas locuciones aparecen entre signos de puntuación aunque también aparecen frases como la siguiente en que el primer elemento de la puntuación se ha omitido: *... en 1765 había presentado Nicholas_Joseph_Cugnot en París su pesado triciclo movido por un motor de combustión externo es_decir, caldera y máquina de Wat, ...* (dc10).

3. al_fin / al fin

La locución adverbial *al_fin* equivale a *finalmente* (ejemplo 3.5) mientras que *al fin* equivale a *al final, al límite*[21].

(3.5) *Entendí* **al_fin_rg** *que en aquel asunto yo no había sido un elemento pasivo, un eslabón más en una cadena* (a11).

4. de_nuevo / de nuevo

De_nuevo (locución adverbial) equivale a 'nuevamente' (ejemplo 3.6); mientras que *de nuevo* es una secuencia de preposición y adjetivo calificativo[22].

(3.6) *Suspiró* **de_nuevo_rg** *intentando deshacer el agobio que tenía atravesado en el pecho* (t5).

5. así_es_que / así es que

La locución conjuntiva *así_es_que* es consecutiva y equivale a *de_modo_que* (ejemplo 3.7); en el segundo caso tenemos una secuencia de adverbio, verbo y conjunción subordinante[23].

(3.7) **Así_es_que_cs** *con cincuenta años bien cumplidos, Rosa_Regás es casi una principiante, galardonada con el Nadal, cincuenta años después que Carmen_Laforet* (c2).

[20]Los ejemplos 3.4 (c-d) no proceden del corpus, donde todas las apariciones de estas secuencias son locuciones conjuntivas.
[21]Todos los ejemplos de aparición de esta secuencia en el corpus se han analizado como locución. Un ejemplo en que no habría locución podría ser: *Nos citamos al fin de un plazo de ocho meses.*
[22]En el corpus no aparecen casos de esta secuencia como preposición y adjetivo. Un ejemplo podría ser *La llegada* **de_sps00 nuevo_aq0ms0** *armamento.*
[23]Tampoco en el corpus hay casos en que esta secuencia de palabras no sea una locución, como en *Si él lo hace así_rg es_vsip3s0 que_cs no hay otra forma de hacerlo.*

6. bien_que / bien que

   La locución conjuntiva equivale a *aunque*, mientras que en el otro caso tenemos un adverbio seguido de una conjunción o un nombre seguido de un pronombre relativo. En el corpus sólo aparecen ejemplos de adverbio seguido de conjunción[24]:

   (3.8) *Bien está el amor al trabajo bien hecho,* **bien_rg que_cs** *las criaturas a las que estamos alumbrando nos hagan sentir su viva presencia* (a1)

   Hay algunos casos de adverbio seguido de relativo, que se explican porque el adverbio *bien* está sustantivado, con lo que ya puede funcionar como antecedente del relativo:

   (3.9) *No lo hacemos todo lo* **bien_rg que_pr0cn000** *me gustaría* (d1).

7. ahora_que / ahora que

   La locución conjuntiva tiene valor concesivo[25] o constituye un modismo con el significado de *por cierto*[26]. En el segundo caso podemos estar frente a un adverbio seguido de conjunción[27] o de relativo (ejemplo 3.10).

   (3.10) *Producir, producir y producir, como los japoneses,* **ahora_rg que_pr0cn000** *los japoneses ya no están por la labor* (a18).

8. alrededor_de / alrededor de

   Hemos considerado esta secuencia como locución preposicional cuando tiene un significado temporal (ejemplo 3.11 (a)), mientras que no lo será si tiene valor locativo (ejemplo 3.11 (b)).

   (3.11) (a) *Llegué a la oficina* **alrededor_de_sps00** *las cinco* (a26).
   (b) *los enviaron de viaje* **alrededor_rg de_sps00** *todo el mundo* (dc2).

En el apéndice A puede encontrarse una lista de otras expresiones que hemos considerado locuciones así como de otras que hemos analizado como lemas independientes.

3.3.2.2. Los nombres propios

Un caso particular de unidades léxicas complejas lo constituyen los nombres propios. El módulo de nombres propios tiene en cuenta las mayúsculas y la posición de los elementos en la oración para detectarlos, pero en la fase de corrección manual ha habido que corregir algunos, principalmente para modificar la segmentación, dado que el módulo de nombres propios tiende a hacer las agrupaciones más grandes posibles, y eso en ocasiones implica errores, como en la frase 3.12 (a), donde toda la secuencia de nombres en mayúsculas más

[24] Un ejemplo de su uso como locución lo proporciona Pavón (1999): p. 643, que a su vez lo toma de Bello: *Bien que hubiese grande escasez de provisiones, no nos faltaba lo necesario.* Un ejemplo de sustantivo seguido de relativo lo encontraríamos en la oración: *hace todo el bien que puede hacer.*

[25] Este es el caso de *La casa es cómoda,* **ahora_que_cs** *no tiene ascensor*, ejemplo tomado del DRAE.

[26] Según Pavón (1999): p. 641, éste sería el caso en la oración *Ahora que lo dices, ¿sabes a quién vi ayer?*

[27] Esta secuencia se daría en la oración: *Dice* **ahora_rg que_cs** *no vendrá.*

las preposiciones aparecen agrupadas como un único nombre propio, cuando en realidad hay tres nombres propios distintos, como se muestra en 3.12 (b):

(3.12) (a) *un grupo de investigadores del departamento de Medicina_Experimental_del_Instituto_Max_Planck_de_Gottingen*
(b) *un grupo de investigadores del departamento de Medicina_Experimental del Instituto_Max_Planck de_Gottingen* (dc10).

Evidentemente el conocimiento necesario para desambiguar estos casos es conocimiento del mundo que difícilmente puede integrarse en desambiguadores automáticos.

3.3.3. La ambigüedad en el corpus CLiC-TALP

El cuadro 3.5 muestra la ambigüedad morfosintáctica que aparece en las 106126 palabras del corpus **CLiC-TALP** una vez que el texto ha sido analizado con todos los módulos anteriormente mencionados.

# **de** *types*	# **de** *tokens*	# **de interpretaciones**
698	1376	12
1	943	9
3	14	8
17	612	7
47	471	6
103	1136	5
460	4132	4
1590	14428	3
4971	32886	2
10745	50127	1

Cuadro 3.5: Ambigüedad en el corpus

Los datos sobre la ambigüedad que proporcionamos tanto en esta sección como en las siguientes deben tomarse como una simple muestra de este fenómeno. El corpus **CLiC-TALP** se utiliza aquí como ejemplificación y las listas de palabras no son exhaustivas de la lengua, sino de este corpus en concreto. A pesar de ello, creemos que lo referente a la ambigüedad verbal sí es extrapolable a la lengua española en general.

Las palabras que presentan 12 posibles etiquetas son aquellas que el analizador no ha reconocido[28]. Tras la fase manual de validación, la mayoría han resultado ser nombres

[28]Esto ocurre por ejemplo con los extranjerismos. Si se analiza la palabra *Sunday* el resultado es el siguiente: *sunday: sunday VMM0000; sunday VMN0000; sunday NC00000; sunday AQ00000; sunday VMP0000; sunday VMS0000; sunday X; sunday I; sunday VMG0000; sunday VMI0000; sunday NP00000*, es decir, se le asignan todas las etiquetas correspondientes a las categorías y tipos, pero no las correspondientes a los rasgos morfológicos. Si esta palabra se sitúa en un contexto claramente nominal (*el próximo sunday iremos a la playa*), el resultado del desambiguador automático es: *el el DA0MS0 – próximo próximo AQ0MS0* – **sunday sunday NC00000** – *iremos ir VMIF1P0 – a a SPS00 – la el DA0FS0 – playa playa NCFS000*. Como puede observarse no se proporciona información sobre los distintos atributos de la palabra, en este caso género y número, pero sí información sobre la categoría y el tipo.

propios.

La palabra *una* es la que presenta nueve posibles etiquetas, correspondientes a los siguientes pares lema–etiqueta antes de la desambiguación (cuadro 3.6).

1 Z	una NCFS000	unir VMM03S0	unir VMSP1S0
unir VMSP3S0	uno DI0FS0	uno DN0FS0	uno PI0FS000
uno PN0FS000			

Cuadro 3.6: Interpretaciones para la palabra *una*

Tras la fase de validación, las etiquetas que le corresponden son las que aparecen en el cuadro 3.7.

896	DI0FS0
20	DN0FS0
1	NCFS000
25	PI0FS000
1	PN0FS000

Cuadro 3.7: Desambiguación de la palabra *una*

Las palabras que aparecen con ocho posibles etiquetas son *suma, mueble* y *viva.*

El cuadro 3.8 recoge las palabras con siete interpretaciones (número de apariciones; palabra y etiqueta asignada tras la validación manual):

1	Papa	NP00000	2	Rosa	NP00000
1	ajenos	AQ0MP0	4	aparte	RG
1	aparte	SPS00	1	crease	VMSI3S0
9	doble	AQ0CS0	1	doble	DN0CS0
1	empaque	NCMS000	1	falla	VMIP3S0
1	fuelle	NCMS000	16	novela	NCFS000
1	novela	VMIP3S0	557	para	SPS00
1	prenda	VMIP3S0	1	prenda	VMSP3S0
2	presta	VMIP3S0	7	revista	NCFS000
3	suma	NCFS000	1	suma	VMIP3S0

Cuadro 3.8: Palabras con 7 etiquetas en el corpus

Algunas de las palabras que reciben seis etiquetas aparecen en el cuadro 3.9.

Las 1136 formas de palabra que aparecen con 5 interpretaciones en el texto corresponden a 136 palabras distintas. Hay 517 palabras diferentes que presentan cuatro posibles etiquetas.

Los casos más frecuentes de ambigüedad en el corpus aparecen con palabras que presentan dos o tres etiquetas (32886 y 14428 respectivamente).

Finalmente es de destacar que aproximadamente la mitad de las palabras del corpus no son ambiguas: 50127.

21	cinco	DN0CP0	3	cinco	PN0CP000
2	cinco	W	111	dos	DN0CP0
1	dos	NCMS000	20	dos	PN0CP000
10	este	NP00000	4	media	AQ0FS0
5	media	DN0FS0	2	media	PN0FS000
3	medio	AQ0MS0	11	medio	DN0MS0
10	medio	NCMS000	6	medio	PN0MS000
10	medio	RG	1	uno	DN0MS0
1	uno	NCMS000	100	uno	PI0MS000
2	uno	PN0MS000			

Cuadro 3.9: Palabras con 6 etiquetas en el corpus

La ambigüedad en el corpus es de **2,01** etiquetas por palabra si se tienen en cuenta todas las palabras y de **2,91** si sólo se tienen en cuenta las palabras ambiguas. Si de éstas, eliminamos las que tienen 12 etiquetas, que son una creación del sistema, la tasa de ambigüedad que resulta es del **2,64** etiquetas por palabra.

3.3.4. Clases de ambigüedad: ambigüedad intercategorial

La ambigüedad que se da en las palabras puede clasificarse en dos tipos que hemos llamado *ambigüedad intercategorial* y *ambigüedad intracategorial*. La primera es la que afecta a las palabras que pertenecen a dos o más categorías distintas, como por ejemplo la palabra *joven*, que puede ser nombre o adjetivo; la segunda es la que afecta a los rasgos morfológicos, pero no a la categoría, como en el caso de *cometa* que, siendo nombre, puede ser masculino o femenino, o *cantamos* que, siendo verbo, puede ser presente o pasado.

En esta sección se comentan las principales clases de ambigüedad intercategorial. En la sección 3.3.5 comentaremos los casos de ambigüedad intracategorial. La relación numérica de palabras que se proporciona hace referencia a *types* o clases de palabras (es decir, a apariciones únicas de las palabras) y no a *tokens* o ejemplares. Por otra parte, las clases de ambigüedad que se comentan no tienen en cuenta que algunas de las palabras mencionadas puedan además presentar otra u otras etiquetas[29].

El fenómeno de la ambigüedad presenta una doble vertiente. Por una parte, están aquellas palabras a las que corresponde un número elevado de etiquetas; y, por otra, aquellas palabras que presentan sólo dos o tres. En el primer caso, la desambiguación automática suele funcionar correctamente; sin embargo, en el segundo caso no siempre es así. El problema de algunas de las palabras que presentan pocas etiquetas es que los contextos formales locales pueden ser compartidos por todas las interpretaciones. Éstos son los casos en que más difícil se hace la desambiguación, tanto para el tagger automático, como incluso para el anotador lingüista (a este respecto, deben mencionarse los casos de las palabras *se* y *que* a los que dedicamos los apartados 3.3.6.1 y 3.3.6.2 respectivamente).

[29]Tal es el caso, por ejemplo, de la ambigüedad determinante–pronombre–adjetivo–adverbio que afecta a algunas palabras y que aquí no se comentará en conjunto, aunque sí por pares de categorías. Un trabajo interesante sobre la desambiguación de estas palabras lo constituye Bertomeu y Mayol (2003).

1. La ambigüedad Determinante – Pronombre

   122 palabras del corpus pueden pertenecer a ambas categorías, como por ejemplo los indefinidos, los posesivos, los demostrativos, los numerales, etc.

   Consideramos que estas palabras son determinantes si aparecen en un sintagma nominal con núcleo sustantivo explícito (ejemplo 3.13 (a)), y pronombres si constituyen el núcleo del sintagma nominal (ejemplo 3.13 (b)); un caso particular de esta última estructura se da cuando aparecen precediendo a una preposición (ejemplo 3.13 (c)) o a un relativo (ejemplos 3.13 (d)) .

   (3.13) (a) *... disfrutando de los videojuegos que presentan* **algunas_di0fp0** *casas comerciales* (dc1).
   (b) **Éste_pd0ms000** *es extraordinariamente refinado, y la película respeta este carácter; está hecha con lo que corresponde a él: con esmero* (a10).
   (c) *... para conseguir un autómata capaz de emular* **algunas_pi0fp000** *de las actividades propias de los seres vivos* (dc1).
   (d) *No hay sociedad ni época que no hayan conocido la envidia: no hay* **ninguna_pi0fs000** *que la haya tenido por una virtud* (a23).
   (e) *Por eso, la espiral hacia sí* **misma_di0fs0** *que describe Magüi_Mira nos sigue persiguiendo como un grito que ...* (a24).

   Si el determinativo que precede a un relativo depende de un nombre a su izquierda no se ha considerado pronombre sino determinante, como ocurre en el ejemplo 3.13 (e).

   Las palabras *lo, la, los, las* que aparecen precediendo a un relativo (ejemplo 3.14 (a)) o a un adjetivo (ejemplo 3.14 (b)) han sido consideradas determinantes (artículos) y no pronombres. Así se consideran en Alcina y Blecua (1989), RAE (1973), Alarcos (1994) que hablan en ambos casos de sustantivación de la secuencia siguiente mediante el artículo. Para Bello (1847), en cambio, las formas del artículo son formas apocopadas del pronombre.

   (3.14) (a) *Yo no necesito para nada que él me pestañee:* **lo** *que hace falta es que me pestañee un soldador* (c1).
   (b) **Lo** *importante es lo que ocurra ...* (c5).

   Del mismo modo se han considerado artículos en otras construcciones que funcionan en la sintaxis como sintagma nominal:

   (3.15) *Se saben condenados a formar parte del club vitalicio de* **los** *ex* (a12).
   Como corresponde a **los** *de su raza y oficio* (a13).
   Si llegaban al gran acuerdo nacional para_que nada se arreglase, **los** *cuevas,* **los** *redondos,* **los** *gutiérreces y* **los** *solbes se repartirían su cuota de culpabilidad* (a18).

   Cuando estas mismas palabras aparecen precediendo a formas verbales conjugadas, se han etiquetado como pronombres personales:

(3.16) *los mandatos de la conciencia* **los** *recibo por teletexto o por fax* (a12).
El mundo que fueron mis amores y la memoria que me **los** *restituye* (a13).

2. La ambigüedad Nombre propio – Nombre común

La inmensa mayoría de palabras que presentan esta clase de ambigüedad empiezan por mayúscula. El analizador morfológico no las reconoce; el hipotetizador las considera entre otras nombres comunes o propios y por ello, cuando llegan a la fase de desambiguación presentan muchas etiquetas. En concreto, de las 1006 palabras afectadas por este fenómeno, 695 se hallaban en este caso[30]. 189 presentaban solamente esta clase de ambigüedad; 87 podían recibir además otra etiqueta y 25 otras dos etiquetas. El resto, 10 palabras, podían recibir entre cinco y ocho interpretaciones.

Los criterios adoptados para la consideración de nombre propio han sido los que se proponen en Fernández (1999a) y que resumimos a continuación: (i) mayúscula inicial; (ii) monorreferencia; (iii) defecto de significación léxica; (iv) rechazo de complementos restrictivos.

(3.17) *Por_ejemplo, a* **Mario_Conde** *una carta de respaldo de* **J._P._Morgan**, *o al ministro* **Asunción** *un bonito* **Garzón**; *son los "detalles chistosos"* (c1).
De **Este** *a* **Oeste**, *de* **Norte** *a* **Sur**, *los vehículos paralizados, como lombrices muertas, tapizaban avenidas y calles, bulevares y plazas* (a25).
Fisher *ha fracasado lamentablemente en todo su reclutamiento* (d1).

En el corpus se pueden encontrar algunos casos de recategorización de nombres propios en nombres comunes (que aparecen en minúsculas) y que anotamos como tales y asignándoles los atributos morfológicos que aparecen en el determinante, tal como muestra el ejemplo (3.18).

(3.18) *Si llegaban al gran acuerdo nacional para que nada se arreglase, los* **cuevas_ncmp000**, *los* **redondos_ncmp000**, *los* **gutiérreces_ncmp000** *y los* **solbes_ncmp000** *se repartirían su cuota de culpabilidad* (a18).

También es posible encontrar el mismo nombre utilizado como común (ejemplo 3.19 (a)) y como propio (ejemplo 3.19 (b)):

(3.19)
(a) *El* **generalísimo_ncms000** *de Flandes se mostraba muy altivo ante el rey Felipe II* (a20).
(b) *Ella vivió con la cabeza alta ante el* **Generalísimo_np00000**, *que jamás pisó el Palacio de Liria* (a20).

Para todos estos casos el criterio ortográfico de la mayúscula ha sido determinante.

[30] Estas 695 palabras son en realidad palabras desconocidas a las que se asignan todas las etiquetas posibles, lo que incluye tanto la de nombre propio como la de nombre común.

3. La ambigüedad Nombre – Verbo

   Otro caso de homonimia frecuente es el que afecta a estas dos categorías gramaticales como por ejemplo, en las palabras *cuenta, vino* o en los infinitivos. En ambos casos sólo el contexto permite la distinción: los nombres suelen ir precedidos de determinantes, mientras que los verbos como tales no aparecen nunca en tal situación. Pero como los artículos coinciden en forma con los pronombres átonos es posible que una misma secuencia necesite de un contexto más amplio para desambiguarse (ejemplo 3.20 (c)).

   (3.20) (a) *Todo lo que Woody_Allen* **cuenta_vmip3s0** *y presenta pasa* (a10).
   (b) *Y en esa terrible* **cuenta_ncfs000** *no se incluyen los daños de la contaminación* (a10).
   (c) **La_da0fs0 tira_ncfs000** *pintada pasa bajo el cristal* (a1).

   Un caso particular de esta homonimia se produce en los infinitivos. En principio sólo tienen etiqueta nominal (además de verbal) en el diccionario aquellos infinitivos que ya están lexicalizados como sustantivos (es decir, que ya tienen plural en la lengua). Si en el contexto admiten plural (ejemplos 3.21 (a-c)) se han etiquetado como nombres; si no (ejemplo 3.21 (d)) como verbos.

   (3.21) (a) *Es mejor el* **atardecer_ncms000** *que el* **amanecer_ncms000** (t4).
   (b) *Pero, en realidad, le resulta difícil al* **ser_ncms000** *humano matizar sus expresiones* (a25).
   (c) *Tampoco ellos podrán jamás recuperar su antiguo* **ser_ncms000** ... (a12).
   (d) *Tus fantasmas vuelven a* **ser_vsn0000** *tu única compañía* (a13).

   Hemos utilizado dos criterios para discriminar ambas posibilidades. El primero es la capacidad del nombre de llevar un modificador adjetivo (como en *su antiguo ser*); el segundo, la capacidad de flexión plural. Por otra parte, los infinitivos verbales, además de no admitir la pluralización nominal, sólo admiten modificación adverbial[31].

   El caso del ejemplo 3.22 es asimilable al segundo caso de los que propone Bosque (1991) (cf. nota 31) por lo que se ha etiquetado como verbo.

   (3.22) *Ese continuo* **contraponer_vmn0000** *la trivialidad a la dignidad es la mejor eficacia del relato* (a1).

4. La ambigüedad Adjetivo calificativo – Participio

   Las formas verbales del participio pasado reciben dos etiquetas en el analizador morfológico, una verbal y otra como adjetivo[32]. La etiqueta verbal la utilizamos sólo para marcar estas palabras en los tiempos compuestos de la conjugación

[31] Cf. Bosque (1991), donde se comparan las estructuras (a) *el andar María* y (b) *el andar de María*. En (a) es posible *el andar lentamente María* e imposible **los andares María*, mientras que en (b) es posible *el andar lento de María* y *los andares de María*.

[32] Cf. capítulo 2.

(ejemplo 3.23 (a)), en la voz pasiva (ejemplo 3.23 (b)) y en las construcciones absolutas (ejemplo 3.23 (c)). En los casos restantes, mantienen la etiqueta de adjetivo calificativo (ejemplo 3.23 (d)) con el valor **p** para el último atributo, que es el que marca su origen verbal[33].

(3.23)

(a) *Había* **llegado_vmp00sm** *a fines del siglo anterior huyendo...* (t1)

(b) *... los gallegos no son* **considerados_vmp00sm** *unos herederos, sino unos entrometidos* (d2)

(c) *no es de extrañar, pues, que hoy,* **rotas_vmp00pf** *las cadenas que durante tanto tiempo me mantuvieron escayolado al bien común, ...* (c4).

(d) *nos hace sentir limpios,* **educados_aq0mpp**, *pertenecientes al mundo hermoso al que, al parecer...* (c2).

Hay que señalar que hemos considerado construcciones absolutas sólo aquellas que tienen un sujeto propio (y no aquellas en que el participio concuerda con el sujeto de las proposición principal). Así por ejemplo, mientras en el ejemplo anterior (3.23 (c)) tenemos auténticas construcciones de participio absoluto, en los ejemplos de 3.24, puesto que no hay sujeto propio del participio, esta forma recibe la etiqueta de adjetivo:

(3.24) *No es de extrañar, pues, que [...]* **harto_aq0msp** *ya de renunciar a la degustación de pajaritos fritos, haya sucumbido a la fascinación de...* (c4).
Invitado_aq0msp *a comer en la casa, el protagonista se entrega a una activa siesta con la señora* (a11).

5. La ambigüedad Adjetivo calificativo – Nombre común

En el corpus **CLiC-TALP** hay un total de 3386 palabras distintas que presentan esta clase de ambigüedad de las cuales 1700 presentan sólo esta clase de ambigüedad; 711 añaden otra categoría; 224 una cuarta; 31 una quinta; 13 una sexta; 7 presentan siete posibles etiquetas; cuatro, ocho interpretaciones, una once y 695 doce[34].

Para estos casos el contexto (o la distribución) suele ser suficiente para la desambiguación. En el caso de los gentilicios y los nombres de oficio y profesión, hemos considerado que la primera palabra era el sustantivo y la segunda el adjetivo. El caso de la palabra *joven* es particular, dado que si aparece con otra palabra que puede ser también sustantivo o adjetivo, lo hemos considerado adjetivo, mientras que la otra palabra ha recibido la etiqueta de sustantivo (ejemplos 3.25 (a-b))[35].

[33] El total de palabras distintas del corpus que presentan esta clase de ambigüedad es de 1997. Palabras que sólo presenten esta clase de ambigüedad son 699; 529 admiten una tercera interpretación; 101 una cuarta; 1 admite cinco interpretaciones; 2 admiten seis, y, por último, 695 admiten 12 interpretaciones. Entre estas últimas, la mayoría son palabras desconocidas para las cuales se hipotetizan todas las posibles categorías morfológicas principales, que incluyen *aq00000* y *vmp0000*.

[34] Son las mismas que aparecían en los otros casos: palabras desconocidas para las cuales se hipotetizan todas las categorías morfológicas principales. Cf. nota 30.

[35] Cf. Civit, Castellón, y Martí (2001b) para un tratamiento más en profundidad de este tema.

Si las palabras que presentan esta clase de ambigüedad son núcleo único de un sintagma con función nominal las hemos considerado sustantivos (ejemplos 3.25 (c,e)), mientras que si modifican a un sustantivo entonces reciben la etiqueta de adjetivo (ejemplo 3.25 (d,e)).

(3.25) (a) *La voluntad de tres* **jóvenes_aq0cp0 atletas_nccp000**
(b) *Una* **joven_aq0cs0 viuda_ncfs000** *con una niña ...*
(c) *Sólo faltaría que a los* **jóvenes_nccp000** *se les hablase de...*
(d) *Traen de allá cosas muy* **ricas_aq0fp0**
(e) *El* **pasado_ncms000 pasado_aq0msp** *está y de su consolidación progresiva se sirve la vida restante* (a26).

Si estas palabras aparecen sin determinante como atributos en oraciones copulativas las hemos considerado adjetivos (ejemplo 3.26 (b)); si llevan determinación (indefinida o de otra clase), nombres (ejemplo 3.26 (a)):

(3.26) (a) *Rosa_Regás es casi una* **principiante_nccs000** (c2).
(b) *Qué famosos son* **gays_aq0cp** *o* **lesbianas_aq0fp0** (c2).

Por otra parte, en los casos en que aparecen dos nombres juntos y podría considerarse que el segundo de ellos está adjetivado (**verde esmeralda**), la etiquetación adoptada ha sido *verde_ncms000 esmeralda_ncfs000* (puesto que *verde* es ambigua, pero es el núcleo y *esmeralda* sólo puede ser sustantivo).

6. La ambigüedad Adjetivo – Adverbio

Sólo son quince las palabras que presentan, en el corpus, la ambigüedad adjetivo calificativo–adverbio: *antiguo, aparte, bueno, claro, conforme, harto, igual, justo, medio, mejor, peor, pronto, seguro, solo, temprano.* Pero también es posible la ambigüedad adjetivo ordinal–adverbio, que afecta, por ejemplo, a la palabra *primero.*

El criterio de discriminación es la variación morfológica: si la forma tiene variación de número (y/o género) recibe la etiqueta de adjetivo (ejemplo 3.27 (a)); si ha perdido la capacidad de flexión, la de adverbio (ejemplo 3.27 (b)):

(3.27) (a) *El duelo como reparación de ofensas no fue una práctica habitual en el mundo* **antiguo_aq0ms00** (dc10).
(b) *La medicina nuclear, es_decir, el uso de radiaciones ionizantes, como los rayos X o el radio, como medio de diagnóstico y tratamiento médico de ciertas enfermedades es conocida desde* **antiguo_rg** (dc1).

7. La ambigüedad Adjetivo calificativo – Determinante

Las palabras que pueden pertenecer a ambas categorías son un total de 32: *cierta, ciertas, cierto, ciertos, diferentes, distintas, distintos, diversas, diversos, doble, escasa, escasas, escaso, escasos, escasísimas, este, media, medias, medio, medios, mismísima, mismísimo, propia, propias, propio, propios, semejante, semejantes, triple, triples, varias, varios.*

Si aparecen antepuestas al nombre las consideramos determinantes (ejemplo 3.28 (a)), y si aparecen pospuestas al nombre o en cualquier otro contexto, adjetivos calificativos (ejemplo 3.28 (b)). Por lo general, un cambio en la posición de estas palabras respecto del nombre implica un cambio de significado: interpretación cuantificada del nombre si aparece pospuesto a estas palabras o interpretación no cuantificada en caso contrario.

(3.28) (a) **Cierto_di0ms0** *apoyo bien pensante a la persecución de la droga, por_ejemplo, tiene aquí sus raíces* (a23).
(b) *Pero lo* **cierto_aq0ms0** *es que mamíferos, aves, reptiles, ...* (dc1).

8. La ambigüedad Nombre común – Adverbio

Aunque poco frecuente (afecta sólo a 39 palabras), esta clase de ambigüedad también aparece en el corpus, y afecta fundamentalmente a palabras del tipo *hoy, mañana, ayer; bien mal*, etc. Si van precedidas de determinante han sido etiquetadas como nombre (ejemplo 3.29 (a)); y si no como adverbios (ejemplo 3.29 (b)).

(3.29) (a) *... rotas las cadenas que durante tanto tiempo me mantuvieron escayolado al* **bien_ncms000** *común* (c4).
(b) *Ya está* **bien_rg**, *hombre, coño, joder, cagonlaleche, es que no puede ser* (c4).

9. Las categorías morfológicamente invariables.

La ambigüedad entre las palabras invariables es muy reducida. En concreto no hay ambigüedad entre conjunción y preposición; pero sí la hay entre conjunción y adverbio y entre preposición y adverbio. En el primer caso las palabras afectadas son *aun, entonces, incluso, luego, mientras* y *ya*. En el segundo, *aparte, hasta*. Los criterios de desambiguación son los siguientes: las conjunciones y las preposiciones son elementos *transitivos*, es decir, necesitan un término, bien sea en forma proposicional para la conjunción, bien en forma generalmente de sintagma nominal para la preposición, mientras que los adverbios son elementos intransitivos, es decir, pueden aparecer sin ningún otro elemento en la oración. Presentamos aquí los casos concretos de *entonces* y *hasta*. *Entonces* ha sido considerado adverbio si tenía un valor temporal y conjunción si tenía un valor consecutivo. *Hasta* se ha etiquetado como adverbio si era el equivalente semántico de *incluso* (ejemplo 3.30 (a)) y como preposición en el resto de los casos (ejemplo 3.30 (b)).

(3.30) (a) *Ahora ya sé que Mario_Conde es un masón dormido y* **hasta_rg** *es posible que más de uno esté temblando sólo de pensar que el martes se despierte;* (c5).
(b) *Profundizar en la_nuestra pero jamás* **hasta_sps00** *la cursilería, la demagogia o la succión* (c5).

3.3.5. Clases de ambigüedad: ambigüedad intracategorial

La ambigüedad en los rasgos morfológicos afecta fundamentalmente a los verbos y a los nombres. En los verbos, la principal ambigüedad se da en los tiempos imperfectos así como

en el presente de subjuntivo, y en el condicional, entre las personas primera y tercera. Otras ambigüedades verbales se dan entre el imperativo y el subjuntivo (o incluso el indicativo), como en los ejemplos siguientes:

presenta	VMM02S0	VMIP3S0
imaginemos	VMM01P0	VMSP1P0
permitásenos	VMM03S0	VMSP1S0
pongamos	VMM01P0	VMSP1P0

Un tercer caso de ambigüedad en el verbo aparece en los lemas, dado que algunas formas de palabra pueden tener un lema distinto, como las siguientes:

fue	ser	VSIS3S0	ir	VMIS3S0
sé	ser	VSM02S0	saber	VMIP1S0

En todos los casos la única forma de deshacer esta ambigüedad es recorriendo al contexto, dado que sólo la semántica permite establecer el valor concreto de los atributos ambiguos.

En los nombres, la principal fuente de ambigüedad se da en el valor del atributo de género. Podemos distinguir dos casos: en el primero, hay alternancia masculino–femenino; en el segundo alternan una interpretación de género común con otra masculina o femenina. Por lo general, los determinantes o los adjetivos que complementan al nombre suelen indicar el valor adecuado de este atributo. A continuación se muestran algunos ejemplos de estas clases de ambigüedad, junto con el número de veces que aparecen en el corpus.

15	orden	NCFS000	NCMS000
11	final	NCFS000	NCMS000
9	modelo	NCFS000	NCMS000
6	tema	NCFS000	NCMS000
4	cura	NCFS000	NCMS000
3	papa	NCFS000	NCMS000
4	defensas	NCCP000	NCFP000
11	defensa	NCCS000	NCFS000
5	agentes	NCCP000	NCMP000
4	estrella	NCCS000	NCFS000

3.3.6. Palabras más ambiguas

Algunas palabras que presentan poca ambigüedad han resultado especialmente problemáticas para la fase manual de validación. El motivo es que las dos o tres posibilidades de etiquetado que admiten comparten los mismos contextos formales. Los principales casos han sido los de la palabra *se* con tres etiquetas morfológicas y *que*, con dos. Además de estos dos casos, comentamos también los de otras palabras especialmente conflictivas.

3.3.6.1. SE

Tal como ya se comentó en la página 91 del capítulo 2, hay tres posibilidades para la etiquetación de esta palabra:
(a) se él pp3cn000, para sus usos como pronombre;
(b) se él p0300000, para sus usos como marca de verbo pronominal;
(c) se se p0000000, para sus usos como marca de oración impersonal o pasiva refleja.

Las tres posibilidades se contemplan en el analizador morfológico, y los criterios que se han seguido son los que se detallan a continuación.

En primer lugar, se comprueba si es un pronombre personal variante contextual de *le* o con valor reflexivo o recíproco:

SE sustituto de LE. En este caso debe coaparecer obligatoriamente con un pronombre personal en función de complemento directo y su función sintáctica debe ser la de complemento indirecto[36].

(3.31) *Por todo esto le dieron el premio Príncipe_de_Asturias_de_la_Paz de 1987: porque logró ser una persona, aunque nada en su entorno* **se_pp3cn000** *lo permitiese* (a14).

SE pronombre reflexivo. En este caso, la oración admite la aparición del incremento reflexivo *a sí mismo* en masculino, femenino, singular o plural.

(3.32) *– ¡ Pero fíjate qué pintas ! –* **se_pp3cn000** *dijo a sí misma, porque acostumbraba hablar a_solas* (t5).
La verdadera heroicidad no es un acto único: el soldado que **se_pp3cn000** *inmola para salvar a sus compañeros, ...* (a14).

SE pronombre recíproco. En este caso, la oración admite la aparición de la secuencia *unos a otros, entre sí* o del adverbio *mutuamente*.

(3.33) *Sé que mientras católicos y protestantes* **se_pp3cn000** *matan en Irlanda_del_Norte, el Vaticano, ...* (c5).

Si la forma *se* no cumple ninguno de los anteriores requisitos, entonces se comprueba si la oración es impersonal o pasiva refleja:

Las oraciones impersonales aparecen con el verbo en tercera persona del singular y en ellas la partícula *se* bloquea la aparición del sujeto.

(3.34) *Si* **se_p0000000** *trata de "inéditos", de papeles privados que nunca pensaron en publicar, tanto mejor* (a10).

En las oraciones pasivas, el verbo puede aparecer en tercera persona del singular o del plural y hay un sujeto explícito, normalmente (aunque no siempre) pospuesto al verbo con el cual concuerda. Además, el verbo de estas construcciones debe ser transitivo.

[36] No en aquellos casos en que /se/ proviene de un verbo pronominal transitivo: *encontrarse algo*, puesto que en estos casos /se/ no es ningún argumento o complemento verbal: *se lo encontró en el parque* $\Longrightarrow$ se_P0300000 y lema = él.

(3.35) *Por_el_contrario, la verdadera heroicidad* **se_p0000000** *construye calladamente, día a día, sobreponiéndose una y otra vez a circunstancias dolorosas y extremas* (a14).

Finalmente, si ninguno de los criterios anteriores es aplicable, se ha asignado la etiqueta **p0300000**, que marca que el verbo es pronominal o está siendo utilizado como tal. En este caso, *se* no es argumento del verbo ni es indicador de impersonalidad ni de oración pasiva.

(3.36) *... arrabales de miseria en donde* **se** *hacinan cientos de miles de personas* (a14).
el tercer mal del Madrid **se** *llama Benito_Floro* (d1).
... y **se** *atrevía a llevar vestidos chillones en los tiempos en que sus primas y amigas iban de luto tibio de posguerra* (a20).

Se hallan en situación similar otras formas correspondientes a la primera y segunda personas (*me, te, nos, os*), dado que también pueden ser pronombres o bien ser marcas de verbo pronominal, aunque aquí ya no hay tres posibilidades sino sólo dos. En el primer caso deben ser argumento verbal (ejemplos 3.37 (a-b)), o bien aparecer en oraciones recíprocas (las formas plurales) o reflexivas (todas las formas, como en los ejemplos 3.37 (c-d)). Si no son pronombres, reciben las siguientes etiquetas: *me_p010s000., nos_p010p000., te_p020s000., os_p020p000* (ejemplos 3.37 (e-f)).

(3.37) (a) *Y la Susi* **me** *dice: Una guardería a todo plan, naturalmente de niñas* (c1).
(b) *Un soldador que no deje de la Susi ni una pata de gallo y así podré pedirle a doña Carmen_Alborch que* **me** *financie una reconstrucción completa, rápida y divina* (c1).
(c) *El héroe, ungido y tonsurado como tal por los conejos que saca de su brumosa chistera, sólo* **se** *salva a sí mismo y no apaga otro fuego que el de su propia ambición* (c4)
(d) **Nos** *pedíamos consejo unos a otros* (d2)
(e) *Y habrá que reconocer que, si* **nos** *dejamos de hipocresías , ésas son las dos cualidades de un buen regalo: que cueste un dineral y que te ponga mucha alegría en el cuerpo* (c1).
(f) *Pero los boricuas no tragan y piensan hacerse los fuertes: "No* **nos** *fiamos de las promesas"* (n1).

3.3.6.2. QUE

Esta palabra puede pertenecer a dos categorías: conjunción (**CS**) o pronombre relativo (**PR3CN000**). Junto con la anteriormente comentada, es una de las más difíciles de desambiguar, aun admitiendo un conjunto de etiquetas muy reducido. El motivo es que los contextos locales en que ambas etiquetas son posibles son muy similares; por ello, la desambiguación de esta palabra también se lleva a cabo fundamentalmente en la fase manual.

Que los contextos locales son muy similares lo demuestran los siguientes grupos de oraciones (en todos los ejemplos (a) la categoría de *que* es la de conjunción, mientras que en los casos de (b) se trata siempre de un relativo):

1. <adjetivo + *que*>

   (3.38) (a) *Es* **probable que** *sea más* **fácil que** *la segunda célula se dispare* (dc1).
   (b) *El escándalo* **público que** *provocó aquella decisión...* (t1)

2. <preposición + *que*>

   (3.39) (a) *Se ha deslizado en la mente de los españoles la convicción* **de que** *no somos "refinados"* (a10).
   (b) *... como no sea el compromiso común con la democracia y esa dignidad en el comportamiento público* **de que** *hacéis gala* (a4).

3. <nombre + *que*>

   (3.40) (a) *Se dio* **cuenta que**, *ante tanta erudición, hacía falta mucho valor para escribir sobre Elena_de_Troya* (sic.) (a1).
   Druso ordenó a la **tropa que** *plantara en la cima una bandera* (a1).
   (b) *Por muy intensa que sea la* **escena que** *se represente...* (a28).

4. <verbo + *que*>

   (3.41) (a) *Y yo queriendo hacer ver* **que** *no podrían notarme nada* (a26).
   (b) *Las veredas sin urbanizar* **que** *habían quedado abiertas entre las chozas* (a14).

5. <adverbio + *que*>

   (3.42) (a) *Es todavía hoy un aviso* **más que** *una constatación* (d1).
   (b) *Acertó en su diagnóstico de la mentalidad que se preparaba en nombre de la razón social, siempre reacia a asumir en serio la libertad individual, pero en cambio generosa a la hora de brindar soluciones globales impuestas desde* **arriba que** *la hagan superflua o dañina* (a23).

Los criterios que se han utilizado para la desambiguación de esta palabra han sido los siguientes:

a) Recibe la etiqueta de pronombre relativo (**pr0cn000**) cuando es un elemento anafórico que tiene una función sintáctica de constituyente en la oración subordinada que encabeza. Cuando tiene un antecedente explícito puede sustituirse por otro relativo *el cual* (con sus variaciones de género y número) (ejemplo 3.43 (a)). La subordinada introducida por el relativo es adjetiva, y como tal puede sustantivarse con un artículo definido, por lo que en las secuencias de [$(lo, la, los, las) + que$], recibe la etiqueta de relativo (ejemplo 3.43 (b)).

b) Recibe la etiqueta de conjunción subordinante (**cs**) cuando no es anafórico: cuando introduce una subordinada sustantiva (ejemplo 3.43 (c)) o adverbial (ejemplo 3.43 (d)).

(3.43) (a) *En una zona en la* **que** *aún no encuentra clasificación* (a26).
(b) *Lo* **que** *está sucediendo y va a suceder e incluso lo* **que** *ha sucedido* (a26).
(c) *Supongo* **que** *quise decir: ¿cómo han podido creer* **que** *a mí me diera un infarto?* (a26).
(d) *La gran mayoría de los españolitos de entonces, estábamos más preocupados por los estudios [...]* **que** *por lo que se estaba cociendo al otro lado de los Pirineos* (a27).

3.3.6.3. NADA

Esta palabra puede tener cinco etiquetas distintas: dos como forma verbal, una como nombre, otra como adverbio y finalmente otra como pronombre indefinido. Las dos primeras corresponden a formas verbales del verbo *nadar*: el presente de indicativo (*él nada*) y al imperativo (*nada (tú)*). La tercera es la del sustantivo que aparece en oraciones como *el ser o la nada*, siempre precedido del artículo definido en su forma femenina.

Las dos últimas son las que provocan más errores en el analizador, y, por tanto, han sido las que se han corregido manualmente. Tanto si es adverbio como si es pronombre indefinido es una palabra morfológicamente invariable, por lo que para su distinción hemos recurrido a la semántica. Hemos considerado que *nada* es adverbio si equivale a 'en absoluto' (ejemplo 3.44 (a)); mientras que la hemos etiquetado como pronombre indefinido si equivale a 'ninguna cosa' (ejemplo 3.44 (b)).[37].

En cualquier caso, para saber cuál es la interpretación correcta hay que recurrir siempre al contexto.

(3.44) (a) *Una sensibilidad* **nada_rg** *común en su medio* (t1).
(b) *Fermina no tenía* **nada_pi0cs000** *que ver con las decisiones del concurso* (t1).

3.3.6.4. ALGO

Esta palabra puede etiquetarse como pronombre indefinido o como adverbio. Lo hemos considerado pronombre si forma un sintagma nominal y equivale a 'alguna cosa' (ejemplo 3.45 (a)) y como adverbio si cuantifica a una expresión y equivale a 'un poco' (ejemplo 3.45 (b)).

(3.45) (a) *Él solía disimular cuando alguien le preguntaba* **algo_pi0cn000** (t2).
(b) *Ya putearán ellas con los demás cuando pase* **algo_rg** *de tiempo* (t2).

3.3.6.5. MIENTRAS

La palabra **mientras** puede recibir la etiqueta de adverbio o la de conjunción subordinante. En la mayoría de sus apariciones en el corpus es conjunción, dado que introduce una subordinada (ejemplo 3.46 (a)). Cuando es adverbio puede sustituirse por otra expresión temporal (ejemplo 3.46 (b)).

[37] A pesar de ello puede darse ambigüedad a nivel oracional, como en
No me gusta nada
que puede interpretarse de dos modos distintos: 'no me gusta en absoluto' (y entonces *nada* será **rg**) o bien 'no me gusta ninguna cosa' (y entonces *nada* será **pi0cs000**).

(3.46) (a) **Mientras_cs** *sus compañeros de "quinta"... resistían sin mayores agobios en su club, él empezaba a ser una estrella intermitente* (d1).
(b) **Mientras_rg**, *Roldán sigue huido; Interior, desbaratado; el Congreso, convertido en asamblea; la prensa, con el escándalo diario, y la Administración, paralizada* (r2).

3.3.6.6. HASTA

Esta palabra puede pertenecer a dos categorías: adverbio o preposición. La hemos considerado adverbio si equivale a 'incluso' (ejemplo 3.47 (a)), y preposición en el resto de los casos (ejemplo 3.47 (b)).

(3.47) (a) *Aquí se nos ha metido la recesión* **hasta_rg** *por los cojones* (a18).
(b) *Turbado* **hasta_sps00** *los tuétanos, no encontró una réplica oportuna para la inclemencia de Sara_Noriega* (t1).

3.3.6.7. LO MISMO

Esta secuencia de palabras puede analizarse bien como locución adverbial, bien como artículo seguido de determinante indefinido. Si forma un sintagma nominal y equivale a 'las mismas cosas' se ha etiquetado por separado como artículo e indefinido (ejemplo 3.48 (a)). En los otros casos, como adverbio (ejemplo 3.48 (b)).

(3.48) (a) *¿No se podría hacer* **lo_da0ns0 mismo_di0ms0** *con las ciudades españolas?* (a10).
(b) *afecta a todos los relojes y a su funcionamiento,* **lo_mismo_rg** *si son mecánicos, biológicos o atómicos* (dc2).

3.3.6.8. POCO

La palabra *poco* puede presentar tres etiquetas: determinante indefinido, pronombre indefinido o adverbio. Como adverbio es un palabra invariable (ejemplo 3.49 (a)); como pronombre es núcleo de un sintagma nominal (ejemplo 3.49 (b)) y como determinante precede a un sustantivo al que cuantifica (ejemplo 3.49 (c)).

(3.49) (a) *Los que vayan a salir del trágico caos yugoslavo importan* **poco_rg** (a21).
(b) *También ahora estas páginas son el hombre y de él tenemos no* **poco_pi0ms000** *que decir* (a1).
(c) *... con escasísimas creencias morales y desarbolado en lo* **poco_di0ms0** *que le quedaba en pie* (a21).

3.3.6.9. MEDIO

Esta palabra puede recibir cinco etiquetas: nombre común, determinante cardinal, pronombre cardinal, adverbio y adjetivo. Lo hemos tratado como nombre cuando aparece

(generalmente) determinado y admite plural (ejemplo 3.50 (a)); como determinante si cuantifica a un nombre al que precede (ejemplo 3.50 (b)); como pronombre si es núcleo de un sintagma nominal (ejemplo 3.50 (c)); como adverbio cuando es invariable (ejemplo 3.50 (d)); por último, como adjetivo (calificativo) si aparece pospuesto al nombre (ejemplo 3.50 (e)).

(3.50) (a) *Los individuos de una especie pueden extinguirse o evolucionar a una estirpe mejor adaptada al* **medio_ncms000** (dc1).
(b) *Pero de tanto en tanto había para* **medio_dn0ms0** *pollo o para un salchichón* (a21).
(c) *Se pasa medio libro discutiendo por teléfono y el otro* **medio_pn0ms000** *hinchándose de drogas hasta que se le salen por las orejas* (a12).
(d) *Si no nació en un bombardeo, sí creció en el de Londres, donde su papá era* **medio_rg** *espía inglés* (a20).
(e) *indignaciones de clase* **media_aq0fs0** *que sólo se acuerda de lo que sufren los pobres* (a23).

3.3.6.10. MUCHO

Esta forma puede etiquetarse como determinante o pronombre indefinido o como adverbio. La hemos etiquetado como adverbio cuando no presenta variación morfológica (ejemplo 3.51 (a)); como pronombre cuando es núcleo de un sintagma nominal (ejemplo b); y como determinante cuando acompaña a un sustantivo en el grupo nominal (ejemplo c).

(3.51) (a) *La salud es algo* **mucho_rg** *menos sujeto a discusión* (a23).
(b) *En_realidad, no sé* **mucho_pi0ms000** *sobre abuelos* (a11).
(c) *Así le quita solemnidad a la tarea que desempeñó durante* **mucho_di0ms0** *tiempo como 'el cronista' que suscribía con el seudónimo de G* (a19).

3.3.6.11. SEMEJANTE

Esta forma de palabra puede recibir las etiquetas de determinante demostrativo, adjetivo calificativo o nombre común. En el primer caso equivale a *tal* y aparece antepuesto al nombre (ejemplo 3.52 (a)); en el segundo aparece pospuesto al sustantivo y significa 'parecido' (ejemplo 3.52 (b)); en el tercero suele aparecer en plural para referirse a 'los demás de la especie' (ejemplo 3.52 (c)).

(3.52) (a) *Por aquí y por allá se suceden los hechos consuetudinarios que requieren* **semejante_dd0cs0** *medida* (a25).
(b) *Un nombre* **semejante_aq0cs0** *se considera perjudicial para la criatura* (c1).
(c) *...que usted es una persona poseída por el divino don de la caridad y quiere ayudar a sus* **semejantes_nccp000** (a22).

3.3.6.12. TAL

Esta palabra puede recibir tres etiquetas distintas: determinante demostrativo, pronombre demostrativo o adverbio. En el primer caso acompaña a un nombre al que se antepone

(ejemplo 3.53 (a)); en el segundo es núcleo de un sintagma nominal (ejemplo 3.53 (b)); como adverbio es morfológicamente invariable (ejemplo 3.53 (c)).

(3.53) (a) **Tales_dd0cp0** *errores son sólo fruto de mi devoción* (a13).
(b) *El héroe, ungido como* **tal_pd0cs000** *por los conejos que saca de su brumosa chistera* (c4).
(c) *Permítasenos señalar que,* **tal_rg** *como consta en la Carta_de_Situación de 24 de febrero...* (a21).

3.3.7. Elementos que sólo aparecen en el corpus

Determinado tipo de información como las palabras mencionadas, los extranjerismos, los títulos de obras, las listas o enumeraciones de series o las transcripciones de fenómenos específicos del habla deben ser tratadas en el análisis de corpus pero no tienen un estatus en los diccionarios de formas.

A continuación presentamos la solución adoptada para la anotación de cada uno de ellos en el marco del corpus **CLiC-TALP**.

3.3.7.1. Palabras mencionadas (usos metalingüísticos)

Aquellas palabras que en el texto aparecen mencionadas, las hemos etiquetado como nombres comunes, pero sin añadir ninguna otra clase de información, es decir, llevan la etiqueta **nc00000**, como las que aparecen en los ejemplos 3.54.

(3.54) *Habría que sustituir* **muerto_nc00000** *por* **destronado_nc00000**, *porque Butragueño, ya se ve, no está muerto* (d1).
El animal dotado de la capacidad de decir **no_nc00000**
Se parece a la palabra francesa **monde_nc00000**, *a la esperantista* **mond_nc00000**... (a15).

Los títulos de obras en general, ya sean de la literatura, del arte, de películas, etc. se han tratado como una sola unidad léxica y etiquetado como nombres propios:

(3.55) "**El_rey_y_el_país_con_granos_np00000**" *es un puro esperpento* (a1).
*A veces me viene a las mientes la historia de "***Le_roi_Bambace_np00000**" (a1).

3.3.7.2. Extranjerismos

Hemos tratado los extranjerismos como palabras españolas. Si el contexto permitía deducir sus rasgos morfológicos, hemos incorporado esta información a la etiqueta, tal como aparece en los ejemplos 3.56.

(3.56) *Una servidora está* **tresanchantada_aq0fs0** *de darle a la* **pen_ncfs000**, *o_sea, a la pluma* (c1).
No lo tengo sencillo. **Enigüei_rg**, *reconozco que lo tengo más sencillo que ayer* (c1).
El Mendicuti ya está mucho mejor, [...], razonablemente **japi_aq0ms0** *por el comportamiento de público y crítica con su último* **buk_ncms000** *[...] y* **veri_rg** *contento con el hijo* (c1).

Este **niuspéiper_ncms000** *ponía ayer un ejemplo* (c1).
... de la temporada normal y de los **playoffs_ncmp000** (d1).
*No me hace ninguna ilusión que los EEUU asuman el papel [...] de "*__sheriff_ncms000__*" del planeta* (c5).

3.3.7.3. Listas

En algunos textos, sobre todo de carácter científico, aparecen enumeraciones de elementos marcadas como listas. Las etiquetas que hemos dado a estos elementos son *Fz* y los paréntesis o marcas que aparecen van unidos al numerador:

(3.57) *... los tres tipos fundamentales que actualmente se conocen:* **a)_Fz** *Polímeros orgánicos;* **b_Fz)** *Metalomacrociclos poliméricos;* **c)_Fz** *Complejos de transferencia de carga...* (dc3).

3.3.7.4. Transcripciones de usos particulares de la lengua

Un último fenómeno que debemos comentar es el que ejemplifica la siguiente oración, en la que la ortografía reproduce una particular forma de pronunciación.

(3.58) (g) *Llegalán los bulldozers pa tilal la "casita", y nosotros delante, gritando no nos movelán* (n1).

La opción que hemos tomado ha sido la de etiquetar estas palabras como lo haríamos si su forma gráfica fuera la canónica.

```
" " Fe
Llegalán llegar VMIF3P0
los el DA0MP0
bulldozers bulldozer NCMP000
pa para SPS00
tilal tirar VMN0000
la el DA0FS0
" " Fe
casita casita NCFS000
" " Fe
, , Fc
y y CC
nosotros yo PP1MP000
delante delante RG
, , Fc
gritando gritar VMG0000
no no RN
nos yo PP1CP000
movelán mover VMIF3P0
. . Fp
```

3.3.8. Cambios de etiquetas respecto del analizador

Como se indicó al inicio de este capítulo, en la fase de validación manual se han mantenido las etiquetas proporcionadas por el analizador morfológico. Sin embargo, en algunos casos, en lugar de introducir más ambigüedad en el analizador (para contemplar nuevas etiquetas para algunas palabras), las hemos modificado. Estos casos afectan a formas femeninas de los adjetivos calificativos y al pronombre *le*.

En el primer caso, se trata de las secuencias coordinadas de adverbios en *-mente*. El primer elemento de la secuencia pierde el sufijo, por lo que la forma que queda es la misma que la del adjetivo femenino singular. En lugar de aumentar la ambigüedad en el analizador dando a todas estas formas la etiqueta de adverbio, con las dificultades que ello hubiera implicado para los desambiguadores automáticos, hemos optado por anotarlas sólo en el corpus, tal como muestran los ejemplos siguientes:

(3.59) *y es el impulso de castigar al hombre con una superior capacidad de felicidad, rebajándole al miserable nivel de las buenas personas, es_decir, de las personas* **estúpida_rg**, **cobarde_rg** *y crónicamente desdichadas* (a23).
El afortunado, el potente, el dominante, el agraciado **física_rg** *o intelectualmente, todos los dichosos han sabido en cualquier época y cultura que sus ventajas tienen un irremediable contrapeso* (a23).

El segundo caso de modificación de la etiqueta afecta al pronombre *le* en los casos en que se produce el fenómeno del *leísmo*. Este uso no es normativo, aunque sí se acepta. En principio esta palabra siempre viene etiquetada como *pp3csd00* (es decir, como pronombre personal en caso dativo). Sin embargo, se ha cambiado por **le_pp3csa00** (con caso acusativo) cuando tiene la función sintáctica de complemento directo, como en los siguientes ejemplos.

(3.60) *Como el Mendicuti, superciclotímico, anda samzin deprimido desde que* **le_pp3csa00** *regañaron por su columna del sándei ...* (c1).
Antonia **le_pp3csa00** *invitó a un café* (t5).

En el corpus no ha aparecido ningún caso ni de loísmo ni de laísmo, por lo que las etiquetas para las formas pronominales *lo, la* se han mantenido con el valor acusativo para el atributo de caso.

Hasta aquí hemos comentado los criterios seguidos para la validación manual de la desambiguación automática del corpus **CLiC-TALP**. El resultado de esta validación constituye un corpus de aprendizaje para desambiguadores automáticos. En la sección siguiente detallamos cómo ha sido posible mejorar los resultados de este desambiguador mediante la introducción de restricciones manuales.

3.4. Introducción manual de restricciones

RELAX admite la introducción manual de restricciones. La diferencia fundamental entre éstas y las inferidas de modo automático, es que las reglas manuales raramente son

puros bigramas, sino que intervienen en ellas más elementos. Además, en las restricciones manuales se ha trabajado también con etiquetas completas, con lemas y con palabras, tanto en la parte inicial de la regla (la que denota la clase de ambigüedad), como en la expresión del contexto.

Por otro lado, en las restricciones manuales aparecen más elementos de los comentados a propósito de las restricciones adquiridas de modo automático. En la figura 3.5 puede observase en primer lugar la declaración de la clase de ambigüedad, cuando se especifica la categoría sobre la que se realiza la restricción (NCF) y se marca la etiqueta que puede tener esa misma palabra, es decir, la palabra que está en la posición 0, que en este caso es NCM. En la expresión del contexto de aplicación de la restricción, aparece la disyunción, que afecta a la palabra anterior, que puede ser, en este caso, un determinante numeral (DN), o un determinante demostrativo (DD).

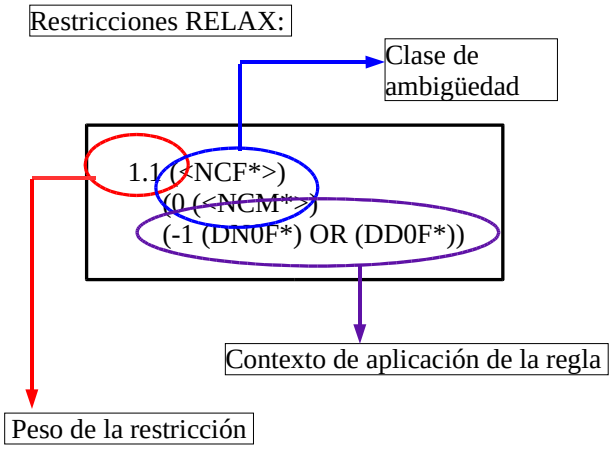

Figura 3.5: Restricciones (2)

Otro elemento que aparece en las restricciones manuales es la negación, siempre en el contexto de aplicación, como en la restricción siguiente:

```
10.1 (<RG>)
          (0 (<PP3*>))
          (NOT -1 (<SPS00>));
```

que debe leerse como *dar prioridad a la etiqueta* **RG** *sobre la que empieza por* **PP3** *si la palabra anterior* **no** *tiene la etiqueta* **SPS00**.

Para establecer las restricciones, se estudiaron los principales errores cometidos por el desambiguador automático una vez que había inferido las restricciones a partir de la validación manual de la desambiguación y comparándolo con el propio corpus **CLiC-TALP**. Los errores más frecuentes (diez o más casos) aparecen en el cuadro 3.10, en el que se señala el número de errores (**casos**) para cada forma, así como la palabra afectada, la etiqueta propuesta por el *tagger* y la que aparece en el corpus. Por otro lado, a nivel más general, se han introducido restricciones teniendo en cuenta fenómenos más amplios de los que aparecían en este corpus. Las reglas que afectan a los lemas, por ejemplo, no tienen

efectos de mejora sobre el resultado del desambiguador automático aplicado a este corpus pero sí pueden tenerlo sobre corpus más amplios.

casos	palabra	tagger	corpus	casos	palabra	tagger	corpus
411	se	P0000000	P0300000	15	orden	NCFS000	NCMS000
315	que	PR0CN000	CS	14	habría	VAIC1S0	VAIC3S0
152	lo	PP3CNA00	PP3MSA00	13	haya	VASP1S0	VASP3S0
127	se	P0000000	PP3CN000	13	me	PP1CS000	P010S000
115	era	VSII1S0	VSII3S0	13	mucho	RG	DI0MS0
80	que	CS	PR0CN000	13	parecía	VMII1S0	VMII3S0
78	había	VAII1S0	VAII3S0	13	sí	PP3CNO00	RG
39	tenía	VMII1S0	VMII3S0	12	llamaba	VMII1S0	VMII3S0
34	estaba	VMII1S0	VMII3S0	12	poco	RG	PI0MS000
30	le	PP3CSD00	PP3CSA00	12	una	DI0FS0	DN0FS0
29	Se	P0000000	P0300000	11	daba	VMII1S0	VMII3S0
26	la	DA0FS0	PP3FSA00	11	defensa	NCCS000	NCFS000
24	sea	VSSP1S0	VSSP3S0	11	final	NCFS000	NCMS000
24	sería	VSIC1S0	VSIC3S0	10	bueno	AQ0MS0	I
22	sí	RG	PP3CNO00	10	fuera	VSSI1S0	VSSI3S0
21	podía	VMII1S0	VMII3S0	10	los	DA0MP0	PP3MPA00
20	podría	VMIC1S0	VMIC3S0	10	nada	PI0CS000	RG
19	hacía	VMII1S0	VMII3S0	10	quería	VMII1S0	VMII3S0
19	hubiera	VASI1S0	VASI3S0	10	sabía	VMII1S0	VMII3S0
16	decía	VMII1S0	VMII3S0	10	un	DI0MS0	DN0MS0
15	Era	VSII1S0	VSII3S0	10	una	DI0FS0	PI0FS000
15	mismo	RG	DI0MS0				

Cuadro 3.10: Principales errores tras la desambiguación automática

Como se ha descrito anteriormente, las restricciones aprendidas de modo automático afectan a la etiqueta corta (**EC**). Las restricciones que se han introducido manualmente afectan tanto a la etiqueta corta, como a la larga (**EL**), pero también al lema de la palabra.

3.4.1. Restricciones sobre lemas

Estas restricciones afectan sólo a las formas verbales. La regla general ha sido dar prioridad a lo irregular sobre lo regular. Se ha dado prioridad a la tercera y a la segunda conjugación sobre la primera atendiendo a la idea de que lo que es irregular persiste si es frecuente y desaparece si no lo es. En el diccionario morfológico de MACO hay 11386 lemas verbales de la primera conjugación, 609 de la segunda y 660 de la tercera. Algunas formas flexionadas de los verbos pueden pertenecer a dos conjugaciones distintas, como por ejemplo la palabra *tejo*, que puede ser del paradigma de *tejer* o del de *tejar*, o la forma *salgo*, que tanto puede ser del verbo *salir* como de *salgar*. Una vez vistos los pares de lemas que aparecían para algunas de las formas verbales, se establecieron las siguientes restricciones, en las que se da prioridad a las conjugaciones segunda y tercera frente a la primera. Si la doble posibilidad de lema afectaba a la segunda y la tercera conjugaciones no se ha introducido ninguna restricción. Éste sería el caso del pretérito indefinido de los verbos *ser* e *ir*, por ejemplo.

Las restricciones son:

```
5.0 ("salir")                5.0 ("tañer")
        (0 ("salgar"));              (0 ("tañar"));
5.0 ("vivir")                5.0 ("ser")
        (0 ("vivar"));               (0 ("erar"))
```

```
5.0 ("tejer")              5.0 ("doler")
        (0 ("tejar"));             (0 ("dolar"));
5.0 ("rugir")              5.0 ("creer")
        (0 ("rujar"));             (0 ("crear"));
5.0 ("morir")              5.0 ("asentir")
         (0 ("murar"));            (0 ("asentar"));
```

Estas restricciones no tienen ninguna influencia sobre la etiquetación del **corpus CLiC-TALP**, pero se han introducido porque esta clase de errores se observaban en el análisis de otros textos.

3.4.2. Restricciones sobre la etiqueta corta (EC)

Las restricciones presentadas en este apartado afectan a las anteriormente mencionadas ambigüedades intercategoriales.

1. La ambigüedad Pronombre – Determinante

   Un primer conjunto de restricciones introducidas de modo manual afecta a las palabras que pueden ser determinantes o bien pronombres. Si estas formas aparecen ante formas verbales, entonces deben recibir la etiqueta de pronombres, lo que se expresa con la siguiente restricción:

   ```
   15.1 (<P*>)
             (0 (<D*>))
             (1 (<VMI*>) OR (<VMS*>) OR (<VAI*>) OR
                       (<VAS*>) OR (<VSI*>) OR (<VSS*>));
   ```

   Estas restricciones no afectan a esta clase de ambigüedad ante las formas infinitivas del verbo, dado que éstos admiten la anteposición de artículos y otros determinantes. Los resultados de unir esta restricción a las aprendidas de modo automático son los que aparecen en el cuadro 3.11. Como puede apreciarse, si se compara con el cuadro 3.3, que sólo contempla los resultados de las restricciones aprendidas de modo automático, se produce una mejora del 0.05 % en lo referente a la etiqueta corta y a la larga, (con o sin lema).

EC	97.34 %
EC+L	96.58 %
EL	94.53 %
EL+L	94.41 %

Cuadro 3.11: Resultados tras la desambiguación P / D (1)

En la oración siguiente aparece dos veces la palabra *la*, la primera como determinante y la segunda como pronombre:

(3.61) *Entonces decidió cambiar de perspectiva: desprestigió a* **la** *heroína devolviéndola a su hogar y* **la** *convirtió en una mujer común, salvo en su soberana belleza* (a1).

2. La ambigüedad Pronombre – Adverbio

Aunque esta clase de ambigüedad afecta a un número importante de palabras, no es posible establecer desde un punto de vista puramente formal los contextos necesarios para desambiguarla excepto en un caso. Como se vio en el cuadro 3.10, la palabra **sí** se etiqueta en 22 ocasiones como adverbio cuando debería recibir la etiqueta de pronombre personal. Revisados los casos en que este error se producía se introdujeron las restricciones siguientes:

```
10.1 (<PP3*>)
          (0 (<RG>))
          (-1 (<SPS00>));
10.1 (<RG>)
          (0 (<PP3*>))
          (NOT -1 (<SPS00>));
```

que indican que si la palabra anterior es una preposición, entonces la palabra *sí* debe recibir la etiqueta de pronombre personal; en este mismo contexto debe etiquetarse como adverbio si la palabra que precede no es una preposición.

Los resultados tras la introducción de esta restricción junto con las automáticas son los que se muestran en el cuadro 3.12. Al igual que ocurría con la restricción anterior, si se comparan estos resultados con los del cuadro 3.3, se produce una mejora del 0.05 % en lo referente a la etiqueta corta y a la larga, (con o sin lema).

EC	97.34 %
EC+L	96.58 %
EL	94.53 %
EL+L	94.41 %

Cuadro 3.12: Resultados para la desambiguación de SÍ

Estas restricciones son de aplicación en los casos siguientes:

(3.62) *Pero* **sí** *creo que la vida de Moyano, su entereza hasta el final y su coraje, forma parte del legado de los humanos, del inconsciente colectivo, de la sustancia común que todos somos* (a14).

Si no nació en un bombardeo, **sí** *creció en el de Londres, ...* (a20).

Transcurre ésta dentro de un cuadro que contiene, dentro de **sí**, *las formas sin fondo de otros dos* (a24).

Por eso, la espiral hacia **sí** *misma que describe Magüi_Mira nos sigue persiguiendo* (a24).

3. La ambigüedad Pronombre relativo – Conjunción

   Esta clase de ambigüedad afecta a la palabra *que*. A pesar de que, como se comentó anteriormente (sección 3.3.6.2), los contextos en que esta palabra puede ser pronombre relativo o conjunción son muy similares y de que para su desambiguación se necesita un contexto muy amplio y también conocimiento externo, hemos establecido una restricción para los casos en que la palabra anterior es la preposición *de* y dos palabras antes aparece un nombre:

```
5.0 (<CS>)
          (0 (<PR*>))
          (-1 ("de"))
          (-2 (<NC*>));
```

   Esta restricción es de aplicación en los casos de *la idea de que, la convicción de que, el hecho de que*, etc. que son secuencias mucho más frecuentes que la de <nombre + preposición *de* + relativo>.

   Los resultados tras la aplicación de esta restricción junto con las automáticas son los que aparecen en el cuadro 3.13 y corrigen 12 casos de error. En este caso, la mejora producida es inferior que en los casos anteriores, ya que sólo se mejora el resultado en un 0.01 % en la etiqueta corta, con y sin lema.

EC	97.30 %
EC+L	96.54 %
EL	94.50 %
EL+L	94.37 %

Cuadro 3.13: Resultados tras la restricción CS / PR

   Ejemplos de aplicación de esta restricción son:

(3.63) *Hasta hace relativamente pocos años, no era imaginable el* **hecho de que** *un compuesto orgánico mostrase propiedades conductoras como un metal* (dc3).
Le viene la **idea de que** *será atacado cuando baje la guardia* (d2)

4. La ambigüedad Nombre – Adverbio

   Esta ambigüedad afecta a palabras como *bien, mal, ayer, hoy* y *mañana*. Como ya se comentó anteriormente, si van precedidas de determinante deben etiquetarse como nombre, lo que se indica con la siguiente restricción:

```
5.0 (<NC*>)
          (0 (<RG>))
          (-1 (<D*>));
```

Los resultados se muestran en el cuadro 3.14. Dado que esta restricción se aplica a un número reducido de palabras, la mejora que aporta al sistema es muy baja. Si se compara con el cuadro 3.3, se produce una mejora del 0.01 % en lo referente sólo a la etiqueta larga.

EC	97.29 %
EC+L	96.53 %
EL	94.49 %
EL+L	94.36 %

Cuadro 3.14: Resultados tras la restricción NC / RG

La variación que produce la introducción de esta regla es muy baja, dado que en el corpus hay pocos casos afectados por esta clase de ambigüedad. Casos en que se aplica esta restricción son:

(3.64) *Entre* **el ayer_ncms000** *y* **el hoy_ncms000** *se comprueba la falta de secuencia en la vida* (a26).
Este niuspéiper ponía **ayer_rg** *un ejemplo: ...* (c1).

3.4.3. Restricciones sobre la etiqueta larga (EL)

Presentamos aquí las restricciones que afectan a las características morfológicas de las palabras. Son, por tanto, restricciones que actúan para la desambiguación intracategorial. Afectan a la persona verbal, al género de los nombres, al pronombre *lo* y a la palabra *se*.

1. Restricciones sobre la persona verbal.

   Los tiempos imperfectos de la conjugación, así como el condicional y el presente de subjuntivo presentan ambigüedad en el atributo de persona, ya que tanto pueden expresar la primera como la tercera persona. La única forma de conocer cuál es el valor de este atributo es a través de la información del contexto y, como en español el sujeto no tiene una presencia obligatoria, en ocasiones hay que recurrir a un contexto de líneas, incluso párrafos, anteriores. Por ello, se ha introducido una interpretación *por defecto*: la tercera persona, ya que estadísticamente, su aparición es más frecuente en el corpus que la de la primera.

   Las restricciones son las siguientes:

```
0.5 (<VMSP3S0>>)                0.1 (<VMII3S0>)
      (0 (<VMSP1S0>));                    (0 (<VMII1S0>));
0.1 (<VMSI3S0>)                 0.1 (<VMIC3S0>)
      (0 (<VMSI1S0>));                    (0 (<VMIC1S0>));
0.1 (<VAII3S0>)                 0.1 (<VSII3S0>)
      (0 (<VAII1S0>));                    (0 (<VSII1S0>));
```

```
0.1 (<VASI3S0>)                0.1 (<VSSI3S0>)
      (0 (<VASI1S0>));                 (0 (<VSSI1S0>));
0.1 (<VSSP3S0>)                0.1 (<VASP3S0>)
      (0 (<VSSP1S0>));                 (0 (<VASP1S0>));
0.1 (<VAIC3S0>)                0.1 (<VSIC3S0>)
      (0 (<VAIC1S0>));                 (0 (<VSIC1S0>));
```

Las mejoras que aportan estas restricciones respecto de las automáticas son las que aparecen en el cuadro 3.15. Esta restricción supone una mejora considerable en los resultados de la etiqueta larga. Si se compara con el cuadro 3.3, puede apreciarse que se pasa del 94.48 % al 95.89 % si no se tiene en cuenta el lema, y del 94.36 % al 95.76 % si éste se incluye en el cómputo.

EC	97.29 %
EC+L	96.53 %
EL	95.89 %
EL+L	95.76 %

Cuadro 3.15: Resultados tras la restricción sobre la persona verbal

2. Restricciones sobre el género de los nombres

   Muchos nombres admiten la interpretación como masculinos y femeninos. Para estos casos se ha utilizado la información del determinante, el adjetivo o de la preposición contraída (en el caso de los masculinos) para desambiguarlos. En las siguientes restricciones puede observarse que la interpretación masculina o femenina de los nombres que presentan esta doble posibilidad morfológica se selecciona en función de los atributos que presenta la palabra anterior. La disyunción de la segunda parte de la restricción incluye la lista de los determinantes y adjetivos que pueden determinar el género del nombre.

```
1.1 (<NCM*>)
  (0 (<NCF*>))
  (-1 (<DN0M*>) OR(<DD0M*>) OR (<DI0M*>) OR
                  (<DP3M*>) OR (<DP2M*>) OR (<DP1M*>)
                  OR (<DA0M*>) OR (<AQ0M*>) OR (<SPCMS>));

1.1 (<NCF*>)
  (0 (<NCM*>))
  (-1 (<DN0F*>) OR(<DD0F*>) OR (<DI0F*>) OR
                  (<DP3F*>) OR (<DP2F*>) OR (<DP1F*>)
                  OR (<DA0F*>) OR (<AQ0F*>));
```

El cuadro 3.16 muestra los resultados tras la incorporación de estas restricciones a

las aprendidas automáticamente por el desambiguador. La mejora que aportan estas restricciones es de un 0.08 % respecto de la desambiguación de la etiqueta larga.

EC	97.29 %
EC+L	96.53 %
EL	94.56 %
EL+L	94.44 %

Cuadro 3.16: Resultados tras la restricción sobre el género de los nombres

Las frases siguientes muestran la palabra *corte* en el primer caso como nombre masculino y en el segundo como femenino:

(3.65) *También dice que, ante la creciente marea de violencia criminal, el problema de la prostitución infantil es como un* **corte** *en un dedo frente_ a un destripamiento a puñaladas* (a21).
La hembra es, por_ lo_ tanto, la que hace la **corte**, *al_ contrario_ de lo que sucede habitualmente* (dc1)

Esta restricción se aplica a palabras como *orden, final* del cuadro 3.10 y otras como *cólera, cometa,* etc. Sin embargo, lo que no es posible es establecer, mediante una restricción puramente formal, la desambiguación entre nombres de género común y nombres masculinos y femeninos, como en el caso de *defensa*, del mismo cuadro. Para poder desambiguar estos casos se requiere de información semántica.

3. Restricciones sobre la palabra **LO**

La palabra *lo*, como pronombre, admite dos interpretaciones posibles: masculina-singular o común-invariable. En el establecimiento de estas etiquetas (cf. cap 2), la primera interpretación correspondía a su función sintáctica de complemento directo de verbos transitivos, mientras que la segunda se estableció para su función como atributo de verbos copulativos.

Para proceder a la desambiguación de esta palabra se han establecido restricciones que tienen en cuenta la forma verbal que aparece tras el pronombre. La primera de ellas discrimina los valores de género y número según si la forma siguiente es del verbo *ser* o no.

```
1.5 (<PP3CNA00>)              1.0 (<PP3MSA00>)
(0 (<PP3MSA00>))              (0 (<PP3CNA00>))
(1 ("ser"));                  (NOT 1 ("ser"));
```

Análogamente, el siguiente par de restricciones discrimina los valores de *lo* según si el verbo *ser* aparece en un tiempo compuesto o no.

```
0.5 (<PP3CNA00>)            0.5 (<PP3MSA00>)
        (0 (<PP3MSA00>))            (0 (<PP3CNA00>))
        (1 (<VA*>))                 (NOT 1 (<VA*>))
        (2 ("ser"));                (NOT 2 ("ser"));
```

Las mismas restricciones son aplicables cuando la palabra siguiente es el verbo *parecer*, también copulativo, en los tiempos simples o compuestos.

```
0.5 (<PP3MSA00>)             0.5 (<PP3CNA00>)
        0 (<PP3CNA00>))              (0 (<PP3MSA00>))
        (NOT 1 ("parecer"));         (1 ("parecer"));

0.5 (<PP3MSA00>)             0.5 (<PP3CNA00>)
        (0 (<PP3CNA00>))             (0 (<PP3MSA00>))
        (1 (<VA*>))                  (1 (<VA*>))
        (NOT 2 ("parecer"));         (2 ("parecer"));
```

Los casos en que este pronombre aparece precediendo al verbo *estar* son algo más complejos. Si *estar* es la única forma verbal tras el pronombre, entonces *lo* debe tener los valores común–invariable; pero si *estar* aparece seguido de gerundio entonces el pronombre ya no funciona como atributo sino como complemento directo, por lo que los valores de los atributos de género y número deben ser masculino–singular. Todo esto se expresa con las restricciones siguientes.

```
0.5 (<PP3MSA00>)           0.5 (<PP3CNA00>)
        (0 (<PP3CNA00>))           (0 (<PP3MSA00>))
        (1 ("estar"))              (1 ("estar"))
        (2 (<VMG*>));              (NOT 2 (<VMG*>));


0.5 (<PP3MSA00>)           0.5 (<PP3CNA00>)
        (0 (<PP3CNA00>))           (0 (<PP3MSA00>))
        (1 (<VA*>))                (1 (VA*))
        (2 ("estar"))              (2 ("estar"))
        (3 (<VMG*>));              (NOT 3 (<VMG*>));

0.5 (<PP3CNA00>)           0.5 (<PP3MSA00>)
        (0 (<PP3MSA00>))           (0 (<PP3CNA00>))
        (1 ("estar"))              (1 ("estar"))
        (2 (<VSG*>))               (2 (<VSG*>))
        (NOT 3 (<VM*>));           (3 (<VM*>));
```

Estas restricciones son complementarias con la anteriormente comentada sobre la ambigüedad pronombre–determinante, por lo que se han provado conjuntamente.

Los resultados son los que aparecen en el cuadro 3.17. La mejora que aportan, con independencia de las restricciones para desambiguar determinante y pronombre, son del 0.26 %.

EC	97.34 %
EC+L	96.58 %
EL	94.74 %
EL+L	94.62 %

Cuadro 3.17: Resultados tras la restricción sobre la forma *lo*

La frase siguiente muestra el pronombre *lo* en un contexto atributivo primero (género neutro) y en un predicativo después (género masculino):

(3.66) *No quiero decir que* **lo_pp3cna00** *sea, cínico o divertido, sino_que ante un mazo de hojas grabadas coloca un cristal bien tallado y* **lo_pp3msa00** *hace girar para_que el sol rompa contra él sus rayos* (a1).

4. Restricciones sobre **SE**

La forma *se* presenta tres posibles etiquetas. Para llevar a cabo una correcta desambiguación de esta palabra es necesario disponer de información sobre la construcción verbal y sobre el contexto amplio de la oración en que aparece. Sin embargo, hay verbos que se utilizan exclusiva o principalmente como pronominales. Para estos casos es posible, pues, establecer una restricción.

```
5.0 (<P03*>)                    5.0 (<P03*>)
        (1 ("dar"))                     (1 ("dar"))
        (2 ("cuenta"));                 (2 ("a"))
                                        (3 ("conocer"));
1.0 (<P03*>)
        (1 ("atreverse") OR ("abrumarse") OR ("adjudicarse") OR
             ("adormilarse") OR ("apostillarse") OR ("arracimarse")
             OR ("arrepentirse") OR ("atenerse") OR ("autodefinirse")
             OR ("autodestruirse") OR ("contorsionarse") OR
             ("desdibujarse") OR ("desgañitarse") OR ("ensimismarse")
             OR ("escabullirse") OR ("fugarse") OR ("resentirse")
             OR ("suicidarse"));
```

Dado que los verbos pronominales aparecen no sólo con la forma *se* sino también con las formas *me, te, nos, os*, el mismo tipo de restricciones se ha establecido para estas cuatro palabras, tal como aparece a continuación.

```
5.0 (<P0*>)                         5.0 (<P0*>)
    (0 ("-me-") OR ("-te-")             (0 ("-me-") OR ("-te-")
       OR ("-nos-") OR ("-os-"))           OR ("-nos-") OR ("-os-"))
```

```
        (1 ("dar"))                    (1 ("dar"))
        (2 ("cuenta"));                (2 ("a"))
                                       (3 ("conocer"));

1.0 (<P0*>)
    (0 ("-me-") OR ("-te-") OR ("-nos-") OR ("-os-"))
        (1 ("atreverse") OR ("abrumarse") OR ("adjudicarse") OR
             ("adormilarse") OR ("apostillarse") OR ("arracimarse")
             OR ("arrepentirse") OR ("atenerse") OR ("autodefinirse")
             OR ("autodestruirse") OR ("contorsionarse") OR
             ("desdibujarse") OR ("desgañitarse") OR ("ensimismarse")
             OR ("escabullirse") OR ("fugarse") OR ("resentirse")
             OR ("suicidarse"));
```

Los efectos de estas restricciones sobre el corpus no son perceptibles. Sin embargo, al igual que en el caso de los lemas, hemos considerado que si se trabaja con grandes cantidades de texto será posible desambiguar correctamente algunos de los casos de aparición de estas palabras.

3.4.4. Resultados

A continuación presentamos los resultados de todas y cada una de las restricciones introducidas manualmente así como los resultados globales, si todas actúan conjuntamente. La restricción sobre la palabra *lo* depende de la restricción que desambigua pronombres y determinantes, por lo que al evaluar la primera se ha tenido también en cuenta la segunda. Como podrá observarse, la restricción que más mejoras aporta al desambiguador automático es la que afecta a la persona verbal de los tiempos imperfectos y del condicional, dado que son formas con una frecuencia de aparición muy alta (especialmente la forma *había* que interviene en los tiempos compuestos.

Restricción	**EC**	**EC+L**	**EL**	**EL+L**
Bigramas	97.29 %	96.53 %	94.48 %	94.36 %
PP/RG	**97.34 %**	**96.58 %**	**94.53 %**	**94.41 %**
CC/PR	97.30 %	96.54 %	94.50 %	94.37 %
NC/RG	97.29 %	96.53 %	94.49 %	94.36 %
P/D	**97.34 %**	**96.58 %**	**94.53 %**	**94.41 %**
lo	(97.34 %)	(96.58 %)	94.74 %	94.62 %
pers. verbal	97.29 %	96.53 %	**95.89 %**	**95.76 %**
gén. - núm. N	97.29 %	96.53 %	94.56 %	94.44 %
TOTAL	97.40 %	96.66 %	**96.28 %**	**96.18 %**

Cuadro 3.18: Resultados globales

Tras la introducción de las restricciones manuales las palabras con diez errores o más

del corpus de entrenamiento han resultado reducirse de 43 a 17, y son las que aparecen en el cuadro 3.19.

casos	palabra	tagger	corpus	casos	palabra	tagger	corpus
411	se	P0000000	P0300000	13	mucho	RG	DI0MS0
313	que	PR0CN000	CS	12	poco	RG	PI0MS000
127	se	P0000000	PP3CN000	12	una	DI0FS0	DN0FS0
82	que	CS	PR0CN000	11	defensa	NCCS000	NCFS000
30	le	PP3CSD00	PP3CSA00	10	bueno	AQ0MS0	I
29	Se	P0000000	P0300000	10	nada	PI0CS000	RG
19	lo	PP3MSA00	PP3CNA00	10	un	DI0MS0	DN0MS0
15	mismo	RG	DI0MS0	10	una	DI0FS0	PI0FS000
13	me	PP1CS000	P010S000				

Cuadro 3.19: Errores tras la introducción de reglas lingüísticas

Lo que se desprende de estos resultados es que las restricciones introducidas de modo manual y basadas en conocimiento lingüístico mejoran significativamente los resultados del desambiguador puramente estadístico, especialmente en lo que respecta a la etiqueta larga. Además, con un pequeño número de restricciones, la mejora de la precisión es considerable. Lo que obtenemos tras la introducción de estas restricciones es un sistema híbrido de desambiguación, probabilístico por una parte, y basado en conocimiento lingüístico por otra, que proporciona unos resultados muy destacables para esta tarea.

Para resolver los casos restantes se precisa de información sintáctica y semántica que no está disponible en este nivel de procesamiento.

3.5. Conclusión

Las aportaciones más destacables de este apartado son, por una parte, el hecho de haber mejorado los resultados de un desambiguador automático en casi un 2 % gracias a la introducción manual de reglas basadas en conocimiento lingüístico. Por otra, el hecho de que como resultado del trabajo se dispone de un corpus del español de 100000 palabras (corpus **CLiC-TALP**) desambiguado y validado manualmente que se ha utilizado para la inferencia de restricciones para un desambiguador automático (RELAX) y que constituye un punto de referencia para el estudio lingüístico.

Capítulo 4

Análisis sintáctico del español: GramEsp

Dedicamos este capítulo al análisis sintáctico superficial del español. El objetivo básico del trabajo que aquí se presenta es poder llevar a cabo un análisis sintáctico correcto desde el punto de vista lingüístico y robusto desde el punto de vista computacional. Tomando como input el resultado del análisis y la desambiguación morfológicos, presentamos una gramática (**GramEsp**) que lleva a cabo un análisis superficial (*chunking*) del español. La complejidad del análisis sintáctico, teniendo en cuenta las características de la lengua española, nos ha llevado a adoptar una estrategia en la que se antepone la amplia cobertura y la robustez del proceso a la profundidad del análisis. Además, el hecho de que nuestro objetivo sea poder tratar el análisis textual sin restricciones hace que aparezcan limitaciones respecto de los fenómenos que podemos tratar de forma coherente y correcta. Por ello, el resultado del análisis es parcial pero robusto. Partiendo de las unidades léxicas se construyen *chunks* que son unidades iguales o más pequeñas que los sintagmas. El sistema de representación que hemos adoptado es el de constituyentes por las propias características del español y de la secuencia de procesamiento de lenguaje natural anterior. Una vez realizado el análisis sintáctico a nivel superficial, no sólo se puede extraer información que guíe un análisis sintáctico de mayor profundidad sino que también se ha enriquecido el texto con información que permite llevar a cabo este proceso con mayor precisión.

La figura 4.1 sitúa este trabajo en el marco de los procesos de análisis del lenguaje de CLiC-TALP.

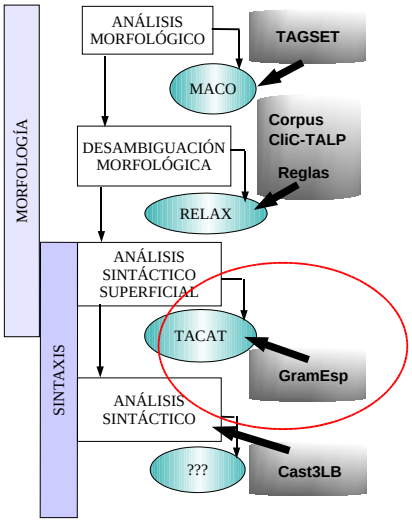

Figura 4.1: Procesos de análisis (3): GramEsp

4.1. Introducción

Habiéndose dado a conocer, aunque de un modo general,
los varios elementos de que se compone la oración,
es ya tiempo de manifestar el orden y dependencia en que los colocamos,
que es lo que se llama sintaxis
(Bello (1847) § 477.)

Los objetivos de un sistema de análisis sintáctico son fundamentalmente dos: agrupar las palabras en unidades de un nivel superior (sintagmas y cláusulas) que identifiquen los constituyentes principales de una oración y etiquetar estos constituyentes. Si el análisis morfológico del texto no está disponible, como paso previo deben identificarse las unidades léxicas a las que se debe asignar una descripción sintáctica adecuada para que el análisis sintáctico propiamente dicho pueda llevarse a cabo.

Para cumplir estos objetivos se necesita un alto grado de precisión y una gran robustez de los sistemas. Algunos sistemas llegan a obtener resultados entre el 80 % y el 90 %[1], aunque no se alcanzan los resultados obtenidos en tareas de POS-tagging, por ejemplo.

Tal como se comenta en Leech, Barnett, y Kahrel (1996):

> *In the case of morphosyntactic annotation, the units to be labelled are to a large extent defined in advance (text word being orthographically indicated). In syntactic annotation, not only do we have to determine which labels to apply to segments of the text, but the segments to which they apply have to be chosen*

[1] Véase, por ejemplo, Bod (2003).

from among many possibilities. [...] On the other hand, there is less consensus about how syntactic segments should be defined.

En este tipo de proceso es necesario establecer una distinción entre el analizador, la técnica de análisis y la gramática. La gramática es una especificación formal de las estructuras permitidas en el lenguaje; la técnica de análisis es el método de análisis de una oración para determinar su estructura de acuerdo con la gramática utilizada[2]. El analizador será el programa que, aplicando una determinada técnica y basándose en los datos de la gramática, efectúe el análisis sintáctico.

En este apartado nos centramos en el tercero de estos componentes: la gramática.

4.1.1. Tipos de gramáticas según su origen

Las gramáticas que se utilizan en los sistemas de procesamiento de lenguaje natural pueden ser escritas de modo manual por expertos lingüistas o bien pueden ser gramáticas inferidas a partir de corpus anotados a nivel sintáctico. McEnery y Wilson (1996a) presentan algunas de las gramáticas existentes en el siguiente cuadro[3]:

	Human rule creation	**No human rule creation**
quantitative motivated	Black et al.'93 **(A)**	Magerman'94, Bod'93 **(C)**
non-quantitative motivation	AI, Language Engineering **(B)**	Brill'92 **(D)**

(A) Estos sistemas utilizan una aproximación híbrida entre las reglas escritas manualmente y la estadística. Por lo general, el sistema utiliza la estadística (basada en corpus) para seleccionar el análisis adecuado de entre los producidos por una gramática escrita a mano. Estos sistemas proporcionan buenos resultados en dominios restringidos.

(B) Se trata de gramáticas escritas manualmente para codificar el conocimiento humano necesario para el análisis. En la práctica, dada la gran cantidad de conocimiento requerido, resulta imposible la construcción de tales gramáticas; las que hay, sólo funcionan para dominios restringidos, y sin embargo, este tipo de gramáticas fue en las que primero se trabajó. Su principal problema es el *knowledge acquisition*

[2] Allen (1995).

[3] Las referencias que los autores mencionan son:

Black, E., Garside, R. y Leech, G (Eds) (1993) *Statistically Driven Computer Grammars of English: the IBM / Lancaster Approach*, Amsterdam: Rodopi.

Bod, R. (1993) 'Using an annotated corpus as a Stochastic grammar', en *Proceedings of EACL'93*, Utrecht.

Brill, E. (1992) 'A Simple rule-based part-of-speech tagger', en *Proceedings of the Third Conference on Applied Natural Language processing (ANLP'92)*, Trento, Italia.

Magerman, D. (1994) 'Natural Language as statistical pattern matching', tesis Doctoral, Standford University.

Y un ejemplo de **(B)** lo proporciona el trabajo de Fordham, A. y Croker, M (1994) 'A Stochastic government and binding parser' aparecido en *Proceedings of the International Conference on new methods in Language Processing, CCL*, UMIST, pp. 190-7.

bottleneck[4], problema por otra parte común a la Inteligencia Artificial: *how do we sit down and write a huge set of rules which describe how to perform some task accurately and consistently?*. Otra de las limitaciones de este enfoque es la dificultad de la tarea de enumerar el conjunto de reglas de la gramática de una lengua natural.

(C) Los sistemas que responden a este enfoque renuncian a cualquier conocimiento lingüístico expresado en forma de reglas escritas a mano. Las gramáticas que producen no serían reconocidas por ningún lingüista y quedan muy lejos de la plausibilidad cognitiva. Los objetivos de estos sistemas son la utilización de técnicas de modelación estadística abstracta para descubrir la estructura interna del lenguaje a partir de textos anotados previamente que sirven para entrenar estos analizadores estadísticos.

(D) Se trata de gramáticas inducidas directamente a partir de corpus sin que en ellas intervenga para nada ningún tipo de conocimiento humano. Por otra parte, tampoco utilizan aprendizaje automático. El uso que hacen del corpus es simplemente el de inducir el conjunto de reglas, es decir, no lo utilizan como fuente de datos cuantitativos.

4.1.2. Tipos de análisis sintáctico según el nivel de profundidad

Por lo general se establecen dos grandes tipos de análisis sintáctico, el total y el parcial. Las técnicas de análisis sintáctico total tienen como objetivo la construcción de árboles de análisis que representen toda la estructura sintáctica de la oración, de la forma más detallada posible: *Full parsing aims to provide as detailed as possible an analysis of the sentence structure*[5]. La realidad de estos sistemas, sin embargo, es que tienen muy poca precisión, fundamentalmente porque para realizar el análisis sintáctico completo se requiere información tanto sintáctica como semántica y son sistemas computacionalmente muy lentos. A pesar de ello, funcionan bien para dominios específicos, donde el contexto tanto semántico como léxico puede ser más local.

El análisis sintáctico, igual que el morfológico, suele ser un estadio intermedio en el procesamiento de información. Las estructuras que construye se utilizan posteriormente con otros fines. Para la mayoría de ellos, sin embargo, el análisis sintáctico total no es estrictamente necesario. Por ello, en los últimos años, se ha tendido a desarrollar sistemas de análisis que proporcionaran información suficiente para las tareas que lo requerían y que a la vez no presentaran los problemas del análisis total. Una de las vías es la del llamado *análisis parcial* y también *skeleton parsing*. Éste consiste en asignar a una oración una estructura de análisis menos detallada.

Según Abney (1996a)

> *Partial parsing techniques aim to recover syntactic information efficiently and reliably from unrestricted text, by sacrificing completeness and depth of analysis.*

La idea principal es reconocer piezas o fragmentos a partir de información puramente (morfo)sintáctica, y dejar para una fase posterior el análisis completo, el que requiere

[4] McEnery y Wilson (1996a): p. 147.
[5] McEnery y Wilson (1996a).

información léxica. Si las oraciones se reducen a segmentos (tradicionalmente llamados *chunks*), hay menos unidades sobre las que considerar posibles asociaciones, de modo que la ambigüedad se reduce de forma importante. Por otra parte, las unidades resultantes, los *chunks* son de interés sintáctico, dado que no son grupos aleatorios. Los *chunks* constituyen, pues, una representación intermedia muy útil y que según Abney (1991) y Abney (1995) tienen un papel importante en el procesamiento del lenguaje por parte de los humanos.

Una primera aproximación al concepto de *chunk* es la siguiente: *The typical chunk consists of a single content word surrounded by a constellation of function words, matching a fixed template*[6].

El propio autor precisa más adelante esta definición:

> *I define chunks in terms of* major heads. *Major heads are all content words except those that appear between a function word* f *and the content words* f *selects OR a pronoun selected by a preposition*[7].

Los *chunks* son pues sintagmas no recursivos, entendiendo por ello sintagmas que no incluyen otros sintagmas, sean o no iguales a sí mismos.

Según (Màrquez, Padró, y Rodríguez, 2001) *la forma más habitual de construir las gramáticas de chunks es de modo manual. Una ventaja de este proceder es que las reglas incorporan conocimiento lingüístico y son directamente interpretables y manipulables por el lingüista.*

Adoptamos por tanto esta estrategia como paso siguiente al análisis morfosintáctico[8]. En esta fase se tratarán aspectos de la morfología no resueltos en la fase anterior y se llevará a cabo la primera fase del análisis sintáctico.

4.2. El sistema utilizado: TACAT

TACAT es un analizador sintáctico basado en *charts* que construye los árboles de análisis *bottom-up* y *left-right* con una gramática libre de contexto. Para la obtención del mejor árbol, sin embargo, opera *top-down* y elige como mejor opción la estructura más amplia y la más profunda (para más detalles sobre este analizador, puede consultarse Atserias y Rodríguez (1998)).

Este analizador sintáctico está integrado en la cadena de procesadores de CLiC-TALP y utiliza como *input* el *output* del desambiguador morfológico, aunque también puede operar con texto analizado sin desambiguar.

4.2.1. Utilidades de TACAT respecto de la gramática

Presentamos en esta sección las características del analizador respecto de la gramática.

[6] Abney (1991).

[7] Abney (1991): p. 2. Como ejemplo propone: proud *is a major head in* a man proud of his son, *but* proud *is not a major head in* the proud man, *because it appears between the function word* the *and the content word* man *selected by* the.

[8] Evidentemente, esta estrategia no es la única posible. En Santalla (2000) se presenta una gramática del español capaz de tratar estructuras oracionales complejas y que tiene en cuenta un lexicón verbal muy rico con información extraída de la *Base de Datos del Español Actual*.

1. Formalismo de las reglas.

   Tacat permite el uso de una gramática libre de contexto (CFG). La forma de las reglas de esta gramática es la siguiente: un elemento (a la izquierda de la regla) se reescribe como cero, uno o más elementos (a la derecha de la regla). La siguiente regla, por ejemplo, es la que reescribe un sintagma nominal como un especificador masculino singular (*espec-ms*) seguido de un grupo nominal masculino singular (*grup-nom-ms*):

   $$sn \rightarrow espec\text{-}ms,\ grup\text{-}nom\text{-}ms.$$

   La posibilidad de que un elemento se reescriba con un elemento vacío no ha sido utilizada en **GramEsp**. Este tipo de reglas suele utilizarse para marcar la opcionalidad de los elementos, pero su coste computacional en términos de tiempo es muy alto y además se crea con ello mucha ambigüedad en la gramática.

2. Elementos *literales*.

   En los elementos a la derecha de las reglas puede especificarse que esa regla se aplique sólo si el elemento contiene una determinada palabra. Por ejemplo, la regla

   $$sn \rightarrow vmip3s0(hace),\ sn(meses).$$

   especifica que el sintagma nominal (*sn*) se construye si y sólo si la primera palabra es la forma verbal *hace* seguida de un sintagma nominal que contiene la palabra *meses*, de modo que el sintagma nominal se construirá en los casos siguientes: *hace meses, hace varios meses, hace cuatro meses* pero no en los casos *hacía mucho frío*.

   Esta particularidad del formalismo permite escribir reglas muy específicas que no podrían construirse de otro modo. En cierta forma, este hecho permite contextualizar algunas reglas de la gramática.

3. Control de aplicación de las reglas en la salida.

   En ocasiones dos o más reglas pueden aplicarse a una misma secuencia de palabras. El analizador permite establecer prioridades para la elección del mejor árbol de los creados por el analizador para la salida del análisis. Ello se consigue declarando las reglas por orden de prioridad tras el elemento **@PRIOR**.

   En el caso de **GramEsp** los elementos de PRIOR son:

   @PRIOR grup-verb
   @PRIOR s-a-ms s-a-mp s-a-fs s-a-fp
   @PRIOR sn
   @PRIOR verb vaux vser
   @PRIOR grup-nom-ms grup-nom-fs grup-nom-mp grup-nom-fp
   @PRIOR sadv
   @PRIOR espec-ms espec-fs espec-mp espec-fp

Estas prioridades se han establecido para tres casos. Por un lado hay nodos máximos de la gramática que en ocasiones se reescriben como un único elemento. Incluyéndolos en PRIOR se asegura su presencia en el árbol de análisis de salida. Se hallan en este caso *grup-verb, sn* y *sadv*.

El segundo caso de uso de PRIOR es el que afecta a las formas verbales (*@PRIOR verb vaux vser*). Estableciendo esta prioridad se asegura que la selección del mejor árbol se hará en función de que las formas se *ser* o *haber* tengan la etiqueta de verbo principal (y no de (semi)auxiliar). Este criterio es de aplicación cuando estas formas verbales se utilizan como núcleos oracionales, en oraciones copulativas en el caso de *ser*, o en oraciones impersonales en el caso de *haber*. Así en una frase como

había mucha gente

se asegura que se seleccione el árbol con etiqueta *verb* y no *vaux*.

Por último, los casos restantes se han utilizado para dar prioridad a las formas masculinas singulares sobre las demás en el caso de los especificadores, los grupos nominales y los sintagmas adjetivos. La forma masculina singular es la no marcada en español, por lo que en caso de que puedan aplicarse ésta y otras reglas (lo que ocurre por ejemplo en el caso de los nombres propios) se seleccione ésta.

4. Control sobre la salida del analizador.

   Finalmente, es de destacar el control que desde la gramática puede realizarse sobre la salida del analizador. Este control se hace mediante otras listas que se declaran también en la gramática. Si se desea poder ver todos y cada uno de los nodos del árbol de análisis, no se declara ninguna categoría en esas listas, pero si la salida que se quiere es un formato reducido del árbol, pueden declararse etiquetas de la gramática en las siguientes listas (en el apéndice B sección B.1 aparecen estas listas con los elementos que contienen en la salida estándar de la gramática):

   a) @HIDEN. Los elementos en esta lista no aparecen en el árbol de salida. Los elementos pseudo-terminales de la gramática (4.3.1.1) están todos incluidos en la misma.

   b) @GROUP. Los elementos de esta lista sólo aparecen si son nodos máximos de análisis, es decir, si no tienen otro nodo que los incluya.

   c) @NOTOP. Es la lista opuesta a la anterior. Los elementos en ella contenidos no aparecen en la salida si son los nodos más altos del árbol.

   d) @FLAT. Esta lista actúa sobre las reglas de la gramática que se incluyen a sí mismas (que son recursivas) y hace que sólo aparezcan una vez en la salida.

   A continuación aparece una frase analizada, en primer lugar, sin que se haya incluido ninguna etiqueta en estas listas y, en segundo, con la salida estándar de la gramática:

```
S_[
  patons_[ paton-s_[ Me_pp1cs000 ] ]
  grup-verb_[ verb_[ gusta_vmip3s0 ] ]
  sn_[
     espec-fs_[ j-fs_[ la_da0fs0 ] ]
     grup-nom-fs_[ n-fs_[ cultura_ncfs000 ]
        sp-de_[ prepc-ms_[ del_spcms ]
        grup-nom-ms_[ n-ms_[ pelotazo_ncms000 ] ] ] ] ]
  conj-subord_[ porque_cs ]
  grup-verb_[ verb_[ sacrifica_vmip3s0 ] ]
  sn_[
     espec-fs_[ j-fs_[ la_da0fs0 ] ]
     grup-nom-fs_[ n-fs_[ búsqueda_ncfs000 ]
        sp-de_[ prep_[ de_sps00 ]
        sn_[ j-ms_[ lo_da0ns0 ]
             s-a-ms_[ a-ms_[ útil_aq0cs0 ] ] ] ] ] ]
  grup-sp_[
     prep_[ en_sps00 ]
     sn_[
        grup-nom-ms_[ n-ms_[ favor_ncms000 ]
        sp-de_[ prepc-ms_[ del_spcms ]
                grup-nom-ms_[ n-ms_[ cultivo_ncms000 ]
                   sp-de_[ prep_[ de_sps00 ]
                     sn_[ j-ms_[ lo_da0ns0 ]
                          s-a-ms_[ a-ms_[ admirable_aq0cs0 ]]]]]]]]]
  ._Fp ]

S_[
  patons_[ Me_pp1cs000 ]
  grup-verb_[ gusta_vmip3s0 ]
  sn_[ espec-fs_[ la_da0fs0 ]
       grup-nom-fs_[ cultura_ncfs000
                     sp-de_[ del_spcms
                             grup-nom-ms_[ pelotazo_ncms000 ]]]]
  conj-subord_[ porque_cs ]
  grup-verb_[ sacrifica_vmip3s0 ]
  sn_[ espec-fs_[ la_da0fs0 ]
       grup-nom-fs_[ búsqueda_ncfs000
                     sp-de_[ prep_[ de_sps00 ]
                             sn_[ lo_da0ns0 s-a-ms_[ útil_aq0cs0 ]]]]]
  grup-sp_[
     prep_[ en_sps00 ]
     sn_[
        grup-nom-ms_[ favor_ncms000
```

```
        sp-de_[ del_spcms
                grup-nom-ms_[ cultivo_ncms000
                   sp-de_[ prep_[ de_sps00 ]
                     sn_[ lo_da0ns0 s-a-ms_[ admirable_aq0cs0 ]]]]]]]]
._Fp ]
```

4.3. GramEsp

La gramática creada para el análisis de textos en español (**GramEsp**) es una gramática libre de contexto basada en *chunks*. La motivación para la elección de este tipo de análisis puede resumirse en los siguientes puntos:

1. la oportunidad de proceder a un análisis por etapas o niveles que permite adquirir conocimiento y preparar el siguiente nivel de análisis;

2. la necesidad de definir un nivel intermedio de análisis sintáctico que proporcione el mejor análisis posible con el conocimiento de que se dispone a través del *output* del análisis morfosintáctico;

3. la resolución del *pp-attachment* no es posible si la única información de que se dispone es la morfosintáctica, por lo que, en la estrategia de análisis adoptada, este problema queda pospuesto;

4. la dificultad de tratar de modo automático los constituyentes tal y como se definen en la lingüística teórica, dado que es muy difícil determinar su alcance. Dos casos concretos de este hecho los constituyen, en primer, lugar el tratamiento del sintagma verbal, dado que, desde un punto de vista teórico, incluye tanto el verbo como sus argumentos; pero la distinción entre argumentos y adjuntos resulta muy difícil de establecer en la práctica, incluso para expertos anotadores. En este sentido se expresan Marcus et al. (1994):

   > *It would also seem desirable to distinguish between the* arguments *of a predicate, and* adjuncts *of the predication. Unfortunately, while it is easy to distinguish arguments and adjuncts in simple cases, it turns out to be very difficult to consistently distinguish these two categories for many verbs in actual contexts. [...] After many attempts to find a reliable test to distinguish between arguments and adjuncts, we have abandoned structurally marking this difference.*

   Además, dadas las características del español, este constituyente puede ser discontinuo. El segundo ejemplo lo aportan las estructuras subordinadas: establecer de modo automático los límites de tales elementos es muy difícil, dado que en español es posible saber dónde se inician, pero no es posible determinar dónde finalizan, si no se dispone de información sobre la estructura argumental del verbo o de información semántica.

5. el procesamiento en la *pipeline* en la que se enmarca el análisis sintáctico;

Siguiendo la idea de Abney de definir *chunks* como *islands of certainty* (Abney, 1996b) hemos optado por desarrollar una gramática que agrupa elementos con el máximo de fiabilidad posible, de modo que el resultado del análisis de texto con esta gramática sea una base correcta (aunque incompleta) que sirva como punto de partida para realizar otros procesos y que a la vez se pueda utilizar para diversas aplicaciones.

La aplicación de la definición de *chunk* al español no es directa, puesto que, por ejemplo, no existe la distinción entre los usos de los adjetivos antepuestos y pospuestos que existe en inglés. Por ello hemos realizado una revisión de la definición de *chunk* propuesta para el inglés. En nuestra propuesta, entendemos por *chunks* los segmentos nominales, los segmentos preposicionales, los segmentos adjetivos y los segmentos verbales. Dada la concordancia que los adjetivos manifiestan en español en género y número con el nombre, es posible incluirlos en el sintagma nominal si presentan con el nombre una relación de adyacencia. De modo que el ejemplo de Abney (1991) *a man proud of his son*, que el autor analiza como

```
[np [ a man ] ap [ proud ] pp [ of his son] ]
```

quedaría, tratado según nuestra propuesta de *chunk* como sigue:

```
[sn [ un hombre orgulloso ] sp [ de su hijo] ]
```

Creemos, además, que la definición de *chunk* es y debe ser dependiente de la lengua. En este mismo sentido se manifiestan Kermes y Evert (2003) que proponen una definición de *chunk* específica para el alemán con dos extensiones de la definición de Abney: por un lado incluyen la incrustación recursiva de elementos prenucleares y, por otro, la incrustación parcial de elementos posnucleares. En concreto la definición que *chunk* que proponen es:

> *A chunk is a continuous part of intra-clausal constituent including recursion and pre-head as well as post-head modifiers but not pp-attachment, or sentential elements.*

En la misma línea, podemos proponer una definición de *chunk* para el español que incluya además de elementos prenominales (como en la definición de Abney) elementos posnominales que concuerden con el núcleo en el caso del sintagma nominal, y algunos casos de pp-attachment (véase la sección 4.3.3 para más detalles sobre este tema).

4.3.1. Descripción de la gramática

En esta sección presentamos **GramEsp**, una gramática para el análisis superficial del español. En primer lugar aparecen las reglas que hemos denominado *pseudo-terminales*, que reescriben las categorías morfológicas; en segundo lugar, presentamos el tratamiento que hemos dado los sintagmas nominal, adjetivo, adverbial y preposicional. Seguidamente presentamos el tratamiento que hemos dado al grupo verbal, con una atención especial al tratamiento de las perífrasis. Por último, presentamos cómo se tratan en **GramEsp** el resto de elementos sintácticos: las formas no personales del verbo y los elementos de relación (conjunciones y relativos).

4.3.1.1. Pseudo-terminales

El punto de partida de las reglas de la gramática son las etiquetas morfosintácticas asignadas tras el proceso de desambiguación morfológica. Todas ellas se han reescrito en etiquetas, que denominamos pseudo-terminales, para poder tratar las mismas categorías de modo agrupado, de forma que sea posible generalizar en las reglas de nivel más alto de la gramática. Estas reglas son reglas operativas que permiten no utilizar las categorías morfosintácticas. Todas tienen un único elemento a la derecha. En el caso de las categorías morfológicas con información de género y número, esta información se incorpora a las reglas y se va propagando hacia los nodos altos de la gramática.

Las palabras con valor neutro para el atributo de género (artículo y demostrativos) se han reescrito como masculinos singulares, dado que ésta es la concordancia que imponen, tal como manifiestan Alcina y Blecua (1989): *Como todos los neutros, [lo] impone concordancia masculina:* Lo *imposible es ambicionado por todos* (p.569); Martínez (1999): *[Cuando las palabras determinadas por* lo*] poseen variación de género o número, el artículo neutro* lo *les impone su neutralización, pues se presentan con la terminación del masculino singular, miembros extensos o no marcados de ambas categorías.*

En estas reglas se producen dos fenómenos. Por una parte, se introduce ambigüedad, dado que aquellos elementos que tienen *común* como valor del atributo de género e *invariable* como valor del de número se reescriben dos veces: en el primer caso como masculinos y femeninos, y en el segundo como singulares y plurales. El caso extremo se produce, por ejemplo, con la etiqueta *aq0cn0* que se reescribe cuatro veces:

$$a\text{-}ms \rightarrow aq0cn0.$$
$$a\text{-}fs \rightarrow aq0cn0.$$
$$a\text{-}mp \rightarrow aq0cn0.$$
$$a\text{-}fp \rightarrow aq0cn0.$$

Por otra parte, tiene lugar una reducción de elementos. Así por ejemplo, las 16 etiquetas morfosintácticas[9] utilizadas para los adjetivos quedan reducidas a 4 pseudo-terminales: *a-ms, a-fs, a-mp, a-fp.*

La reescritura de terminales (etiquetas morfosintácticas) en pseudo-terminales se ha llevado a cabo de dos modos distintos. En un caso hay un único nivel de reescritura: adjetivos, nombres, adverbios, preposiciones, conjunciones y las formas de los verbos[10]:

[9]Dado que TACAT y, por tanto, la gramática se integra en un entorno de procesamiento de lenguaje muy determinado, se han tenido que considerar algunas etiquetas que no responden estrictamente al análisis morfológico sino que son el resultado de la hipotetización que hace el desambiguador cuando se encuentra ante una palabra desconocida. Tal como se comentó en el capítulo 3, el desambiguador sólo trabaja con los primeros dígitos de las etiquetas, los que se corresponden con las categorías y las clases, de modo que al hipotetizar, trabaja del mismo modo y las etiquetas que propone son en cierto modo incompletas. Así, por ejemplo, para un adjetivo propone *aq0000*, para un nombre *nc00000*, etc. Por consiguiente, estas etiquetas también aparecen en la gramática.

[10]Todas las reglas para los pseudo-terminales aparecen en el apéndice B sección B.2.

a-fs → *aq0fs0.*
n-ms → *ncms000.*
adv → *rg.*
prep → *sps00.*
conj-subord → *cs.*
verb → *vmip1s0.*
vaux → *vami1s0.*
vser → *vsmi1s0.*

En otros casos, sin embargo, hemos optado por dos niveles de reescritura: determinantes, pronombres y verbos auxiliares y semiauxiliares. En un primer nivel, las etiquetas morfosintácticas se reescriben con una etiqueta que recoge la clase, tal como puede observarse en los siguientes ejemplos con los determinantes demostrativos, donde de seis etiquetas (*dd0ms0, dd0fs0, dd0mp0, dd0fp0, dd0cs0, dd0cp0*) se pasa a cuatro (*dem-ms, dem-fs, dem-mp, dem-fp*):

dem-ms → *dd0ms0.*
dem-fs → *dd0fs0.*
dem-mp → *dd0mp0.*
dem-fp → *dd0fp0.*

dem-ms → *dd0cs0.*
dem-fs → *dd0cs0.*
dem-mp → *dd0cp0.*
dem-fp → *dd0cp0.*

En un segundo nivel, todos los determinantes se agrupan bajo una única etiqueta *espec-*, de modo que con sólo cuatro etiquetas es posible operar con todos los determinantes:

espec-ms → *cuantif.*
espec-ms → *num-ms.*
espec-ms → *dem-ms.*
espec-ms → *int-ms.*
espec-ms → *exc-ms.*
espec-ms → *indef-ms.*
espec-ms → *j-ms.*
espec-ms → *grup-complex-spec-ms.*

Esta doble etiquetación de los determinantes se justifica, por un lado, porque con la etiqueta *espec* es posible tratarlos conjuntamente cuando, por ejemplo, se define la estructura del sintagma nominal (cf. sección 4.3.3); pero para las combinaciones de determinantes que pueden preceder al nombre, resulta más interesante poder tratarlos por tipos, de modo que aunque se aumente el número de reglas, el resultado aporta un mayor control sobre las estructuras permitidas. El mismo principio se aplica a los pronombres. Por un lado, todos quedan agrupados bajo *pron-*:

pron-fs → *pinterrog-s.*
pron-fs → *pinterrog.*
pron-fs → *psubj-fs.*
pron-fs → *pdem-fs.*
pron-fs → *pinterrog-fs.*
pron-fs → *pposs-fs.*
pron-fs → *pindef-fs.*

de modo que pueden tratarse unitariamente; pero se mantiene el pseudo-terminal con la clase para poder controlar desde la gramática sus apariciones en combinaciones de pronombres[11]. La única excepción a este doble tratamiento de los pronombres son los relativos, que se reescriben en un único nivel. Se han tratado separadamente porque dado su particular comportamiento sintáctico no aparecen en los mismos contextos que el resto de formas pronominales. A continuación aparecen ejemplos de las reglas pseudo-terminales para los relativos. Dado que algunos de ellos comparten las etiquetas morfológicas, pero sus contextos de aparición son distintos, se han tratado como literales en la gramática.

cuyo-ms → *pr0ms000(cuyo).*
cual-s → *pr0cs000(cual).*
quien-s → *pr0cs000(quien).*
prel → *pr0cn000.*
prel-ms → *pr0ms000(cuanto).*
prel-adv → *pr000000.*

En ningún caso los pseudo-terminales (tanto los de primer nivel como los de segundo nivel) aparecen en la salida estándar de la gramática. Como se ha comentado anteriormente (sección 4.3.1.1), son simples unidades operativas que permiten controlar secuencias de elementos permitidas o no permitidas por la gramática. Otra ventaja de los pseudo-terminales es que pueden utilizarse para simplificar el análisis morfológico, por la reducción en el número de elementos que implican respecto de las categorías morfosintácticas. Además, si se elimina el sufijo de estas etiquetas (es decir, la marca de la concordancia), la simplificación es mayor y puede resultar interesante para determinados tipos de análisis sintáctico en que no se requiera esta información.

4.3.2. *Chunks* considerados

En esta sección presentamos los *chunks* considerados por la gramática. Las estructuras complejas que **GramEsp** construye son sintagmas nominales (descritos en la sección 4.3.3); sintagmas adjetivos (cuya descripción se realiza en la sección 4.3.4); sintagmas adverbiales (comentados en la sección 4.3.5); grupos preposicionales (que presentamos en la sección 4.3.6); y grupos verbales (comentados en la sección 4.3.7). Además de estas agrupaciones, también se lleva a cabo el análisis de otros elementos que no quedan incluidos en ninguno de los *chunks* anteriores, como son los relativos, las formas no personales del verbo, y las conjunciones; a ellos dedicaremos la sección 4.3.8.

[11] Las reglas para los pronombres aparecen en el apéndice B sección B.2.8.

Los principios generales que han guiado la construcción de todos estos nodos han sido, en primer lugar, el hecho de realizar un análisis lingüísticamente correcto teniendo en cuenta que sólo disponemos de la información morfosintáctica: se trata de construir *chunks* que se aproximen lo más posible a lo que en lingüística se entiende por sintagmas o constituyentes. Cuando ello no ha sido posible, las construcciones resultantes son incompletas pero correctas, intentando proporcionar el máximo de cobertura.

El análisis resultante es apto para aplicaciones lingüísticas que no requieren un análisis en profundidad del español, y esta gramática se ha utilizado en diversos proyectos y trabajos del grupo de investigación[12], y se utiliza como punto de partida para la anotación sintáctica de corpus (cf. capítulo 5).

4.3.3. El sintagma nominal

A continuación se presentan los elementos que pueden aparecer como constituyentes del sintagma nominal. Sin embargo, y para una mayor claridad en la exposición, el sintagma adjetivo se presenta separadamente, en la sección 4.3.4. En primer lugar se presenta la estructura del sintagma nominal y, en segundo lugar, se detalla cada uno de los elementos de esta estructura: el núcleo, los modificadores y los especificadores.

Por lo general, entendemos por *chunk* nominal, llamado en la gramática *sintagma nominal*, la secuencia formada por *especificador* seguido de un *grupo nominal*; y, por *grupo nominal*, la secuencia formada por un núcleo nominal y algunos de sus modificadores. La representación de esta estructura aparece en la figura 4.2.

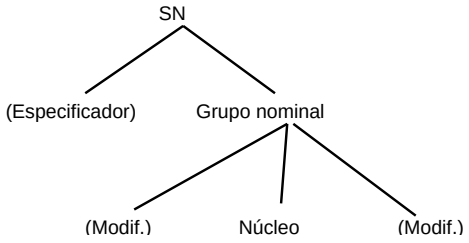

Figura 4.2: Estructura del sintagma nominal

Tanto los especificadores como los modificadores son elementos opcionales[13].

4.3.3.1. La concordancia en el seno del sintagma nominal

Antes de analizar en detalle las reglas del sintagma nominal, trataremos brevemente cómo se ha resuelto la concordancia. Dado que el análisis sintáctico que se lleva a cabo es parcial, la única concordancia considerada ha sido la que se da en el seno del sintagma nominal. Los nodos internos a este sintagma propagan las características morfológicas de

[12]Proyectos: *Spontaneous-Speech Dialogue System In Limited Domains CICYT (TIC98-423-C06-06)*, *MEANING Developing Multilingual Web-scale Language Technologies IST-2001-34460*;
Trabajos: Arévalo (2001), Arévalo et al. (2002), Turmo (2002), Ageno (2003).

[13]En la gramática no se ha considerado ninguna categoría vacía, tal como se ha comentado en la página 184.

género y número hacia los nodos altos del árbol, de modo que se asegura la buena formación interna de estos sintagmas. Sin embargo, el constituyente *sn* no las incluye, puesto que la concordancia de alcance oracional no ha sido tratada aquí.

Según Lehmann (1982)[14] la concordancia puede definirse del modo siguiente:

> *El constituyente B concuerda con el constituyente A (de la categoría C) ssi:*
> *(a)Existe una relación sintagmática gramatical entre A y B;*
> *(b)Existe una categoría gramatical C con un paradigma de formas;*
> *(c) A pertenece a la forma F de C y este hecho es independiente de B;*
> *(d) F es expresado en B y forma un constituyente con él ya sea sintagmático o morfológico*
> *A modo de ejemplo vamos a aplicar esta definición al sintagma español 'perros cazadores'. B será 'cazadores' y A, 'perros'. En primer lugar, sí se da una relación sintáctica entre 'perros' y 'cazadores'; se trata de la modificación. En segundo lugar, existe la categoría de nombre sustantivo en la que se distingue una forma F1 de singular frente a otra F2 de plural. En tercer lugar, 'perros' pertenece a la forma F2 de ese sustantivo y esto es independiente de 'cazadores'. Por último F2 se expresa en 'cazadores' formando una unidad morfológica. De esto se deduce que 'cazadores' concuerda con 'perros'.*

La concordancia entre los elementos del discurso puede afectar a muy diversos morfemas (género, número, caso, persona) y categorías gramaticales. Siguiendo a Martínez (1999), podemos afirmar que *las concordancias propiamente dichas del español se basan exclusivamente en la repetición del género, del número y de la persona, y en consecuencia, y en rigor, sólo pueden concordar entre sí las palabras o clases de palabras que poseen o incorporan alguno de estos signos morfológicos*[15].

Los casos concretos de concordancia que se han tenido en cuenta en **GramEsp** son los que según Martínez (1999) quedan estrictamente en el ámbito del sintagma nominal[16]. En este ámbito, la concordancia de persona no se da, pues afecta al verbo y al sujeto y a la pronominalización de algunos elementos. Por ello, en todos los pseudo-terminales de la gramática que reescriben categorías morfosintácticas que expresan género y/o número, esta información se ha trasladado a la parte izquierda de la regla, como ya se habrá observado en los ejemplos anteriores.

Los casos en que hay concordancia en el seno del sintagma nominal en español pueden esquematizarse como se muestra en el cuadro 4.1[17]:

[14]Citado por (Cabrera, 1987): p. 103-104.

[15]Martínez (1999): p. 2705.

[16]Dado que la gramática sólo analiza *chunks*, no hemos tenido en cuenta los fenómenos de concordancia que tiene un alcance intersintagmático y que son los siguientes: concordancia en género y número de adjetivos y participios a través del verbo con el sujeto o el complemento directo; concordancia al menos en género de calificativos y determinativos utilizados como pronombres con el elemento al que refieren anafóricamente; concordancia en género, número y persona, del pronombre personal con el sustantivo al que refiere, y de los pronombres átonos con los tónicos correspondientes; concordancia en persona y número del sujeto de la oración con el verbo (Martínez, 1999).

[17]Adaptación de las propuestas de Martínez (1999).

Elemento concordante	**Elemento rector**
Artículo	sustantivo adjetivo infinitivo subordinadas sustantivas relativos adverbios
Determinativo	sustantivo infinitivos pronombres
Adjetivo	sustantivo infinitivo determinativos
Relativo	antecedente
Aposición	sustantivo

Cuadro 4.1: Relaciones de concordancia nominal

Así, el artículo concuerda con casi todas las clases de palabras. Además de con el nombre (cf. ejemplo 4.1 (a)) y el adjetivo calificativo (cf. ejemplo 4.1 (b)), concuerda con el infinitivo nominal[18] al que debe preceder obligatoriamente aunque puede ser sustituido por otro determinativo (cf. ejemplo 4.1 (c)) y con el verbal[19] (aunque en este caso su presencia es meramente opcional) (cf. 4.1 (d)). En los dos últimos casos aparece la forma no marcada del artículo.

(4.1) (a) *el bebé, los cuartos, la persona, las habitaciones*
(b) *Lo breve de la visita agradó a todos.*
(c) *Me despertó el alborotado piar de los gorriones.*
(d) *La fatigaba (el) subir por las escaleras diariamente.*[20]

De modo parecido, las oraciones subordinadas sustantivas pueden tomar también opcionalmente el artículo, con valor enfático:

(4.2) (a) *Me molesta (el) que sospechen.*
(b) *Nos sorprendió (el) cómo lo habían hecho.*[21]

[18] El infinitivo nominal es aquél que *se comporta como sustantivo tanto en relación con las unidades a las que se subordina como con aquellas que se le subordinan (adjetivo, complemento determinativo).* Martínez (1999): p 2718.

[19] El infinitivo verbal es el que *aun funcionando como sustantivo respecto de otras palabras a las que se subordina, o formando con ellas -según algunos- oraciones subordinadas sustantivas sin flexión verbal, conserva la capacidad de llevar subordinados o complementos propios del verbo (como el directo o el indirecto).* Martínez (1999): p 2718.

[20] Los ejemplos están tomados de Martínez (1999): p 2718.

[21] Ejemplos tomados de Martínez (1999): p. 2719.

El artículo también puede anteponerse a los relativos así como a las oraciones subordinadas relativas. Como los relativos concuerdan con su antecedente, el artículo toma también esas mismas marcas morfológicas.

(4.3) *No llegó a tiempo el tren en el que venían*[22]

Finalmente, sobre la concordancia del artículo con algunos adverbios, el autor señala que sólo la forma neutra del artículo puede aparecer con adverbios: *los adverbios –salvo algunos que aceptan el neutro lo– no se combinan con el artículo, pues los que lo toman son, en realidad, sustantivos: el sí ['aceptación'], el no ['negativa'], el todo ['totalidad'], la nada ['inexistencia'], el ayer ['el pasado'], el mañana ['el futuro'], ...*[23] El artículo indefinido presenta una distribución más restringida que la del definido, pues sólo puede acompañar al sustantivo y al adjetivo:

(4.4) *Necesitaría un lápiz y una pluma.*
Tu amigo es un tonto.

Por su parte, los determinativos[24] concuerdan con el nombre (4.5 (a-b)), el infinitivo nominal (4.5 (c)) y algunos determinativos (4.5 (d)):

(4.5) (a) *No cayó sin_duda en la cuenta de que se trataba del* **mismo** *Juan_Fernández que hizo a pulso el OAR_Ferrol* (d2).
(b) *Y ya sé que, con* **esos** *nombres,* **todos nuestros** *niños parecerán medio comanches* (c1).
(c) **Ese** *continuo contraponer la trivialidad a la dignidad es la mejor eficacia del relato* (a1).
(d) *Como* **estos** *dos, pívot y base, siguen juntos esta temporada, es lógico que se cuente con ellos al final* (d1).

El adjetivo calificativo, así como el participio, puede concordar también con el nombre (4.6 (a-b)) o el infinitivo nominal (4.6 (c)), además de hacerlo con algunos determinativos utilizados como pronominales (4.6 (d-e)):

(4.6) (a) *El gato negro.*
(b) *El famoso Ramón.*
(c) *Se oyó un arrastrar de cadenas estruendoso.*
(d) *Acudieron dos centenares largos de personas.*
(e) *Se ganó el triple prometido.*[25]

Todavía en el ámbito nominal es preciso comentar la concordancia que se manifiesta en los relativos, en las aposiciones y en las expresiones parentéticas. Los relativos concuerdan con su antecedente en número (*quien*) como puede verse en los ejemplos 4.7 (a-b), o en género y número (*cual, que* precedidos de artículo) como se muestra en los ejemplos 4.7 (c-d)[26]:

[22] Ejemplo tomado de Martínez (1999): p. 2741.
[23] Martínez (1999): p 2713.
[24] Que para el autor son *todos los adjetivos diferentes a los calificativos* (Martínez (1999): p. 2724).
[25] Ejemplos tomados del autor.
[26] Martínez (1999): pp. 2739-40.

(4.7) (a) *No conozco al chico con quien veníais.*
(b) *Las personas a quienes se lo dije no acudieron.*
(c) *El cuchillo con el cual lo hizo estaba en su mano.*
(d) *Eso lo sabrás tú, al que se lo cuentan todo.*

Los relativos *cuyo, cuanto* no concuerdan con su antecedente sino con el sustantivo que les sigue[27]:

(4.8) *Buscan al chico a cuyas hermanas te presenté.*
Cuanta gente lo conoce termina odiándolo

Por lo que respecta a la concordancia en las construcciones apositivas, hay que señalar que la aposición es una *construcción en que entran dos o más sustantivos que, siendo lingüísticamente distintos, confluyen sin embargo en una única referencia extralingüística*[28]. Según Martínez (1999), *las unidades correferentes concordarán entre sí en género y número cuando una de ellas, o las dos, presenten algún tipo de variación morfológica*[29]:

(4.9) *Ella, Matilde, estaba sentada ahí.*
Las madres, las más listas, procuraban vivir del estraperlo

Pero en muchos casos, tal concordancia, especialmente la de género, no se presenta[30]:

(4.10) *El animal clonado, la oveja, salió en todos los periódicos.*

La concordancia de número es más frecuente, aunque también hay casos en que no está presente[31]:

(4.11) *El rebaño, cien ovejas y veinte cabras, pacía tranquilamente.*

Estas divergencias en la concordancia se dan sólo en las aposiciones explicativas, pero no en las especificativas, donde la concordancia se sigue de modo estricto (evidentemente, si por lo menos uno de los sustantivos presenta variación)[32]:

(4.12) *El rey profeta.*
El río Ebro.
Madrid ciudad

Las frases nominales parentéticas o incidentales se comportan como las aposiciones explicativas[33]:

(4.13) *Satélite de la Tierra, la luna gira alrededor.*

[27] Ejemplos tomados de Martínez (1999): p. 2741.
[28] Martínez (1999): pp. 2741-42.
[29] Ejemplos tomados de Martínez (1999): p. 2742.
[30] Ejemplo tomado de Martínez (1999): p. 2742.
[31] Ejemplo tomado de Martínez (1999): p. 2742.
[32] Ejemplos tomados de Martínez (1999): p. 2743.
[33] Ejemplos tomados de Martínez (1999): p. 2743.

En estas construcciones no entran sólo nombres sino también adjetivos y participios, y entonces se mantienen todas las concordancias (incluso con el nombre elidido)[34]:

(4.14) *Pálidos de miedo, los viajeros se alejaron de la explosión.*
Tienen que dejar el trabajo, obligados por la enfermedad.

Una estructura paralela a la de las aposiciones especificativas se da en casos como[35]:

(4.15) *el primer bebé probeta.*
los primeros bebés probeta.

No son aposiciones (no hay identidad referencial) y además la concordancia no se mantiene. El segundo sustantivo parece haberse fijado respecto de la flexión.

Todas estas concordancias internas al sintagma nominal se tratan en **GramEsp** excepto tres: las que afectan a los relativos[36], las de las construcciones parentéticas y las aposiciones explicativas, dado que son relaciones de concordancia que tienen un alcance que va más allá del *chunk* nominal.

A continuación aparecen ejemplos de análisis de *chunks* nominales complejos.

```
El bebé probeta
sn_[ espec-ms_[ el_da0ms0 ]
     grup-nom-ms_[ bebé_ncms000 probeta_ncfs000 ] ]

La Madre Coraje peruana
sn_[ espec-fs_[ la_da0fs0 ]
     grup-nom-fs_[ Madre_Coraje_np00000
        s-a-fs_[ peruana_aq0fs0 ] ] ]

El Madrid oficial
sn_[ espec-ms_[ El_da0ms0 ]
     grup-nom-ms_[ Madrid_np00000
        s-a-ms_[ oficial_aq0cs0 ] ] ]

Un buque estadounidense repleto de proyectiles de gas mostaza
sn_[ espec-ms_[ Un_di0ms0 ]
     grup-nom-ms_[ buque_ncms000
        s-a-ms_[ estadounidense_aq0cs0
           s-a-ms_[ repleto_aq0ms0 ] ] ] ]
grup-sp_[ prep_[ de_sps00 ]
     sn_[ grup-nom-mp_[ proyectiles_ncmp000
          sp-de_[ prep_[ de_sps00 ]
                  sn_[ grup-nom-ms_[ gas_ncms000 mostaza_ncfs000 ]]]]]]
```

[34] Ejemplos tomados de Martínez (1999): p. 2744.
[35] Ejemplos tomados de Martínez (1999): p. 2744.
[36] Aunque sí se tienen en cuenta las de los relativos determinantes, como se comentará en la sección 4.3.8.2.

El hecho que queremos destacar aquí es que la fase de desambiguación morfológica acaba con la gramática, que es la que asegura las concordancias internas de los *chunks* nominales. En todos los casos de reescritura de terminales en pseudo-terminales (excepto en el caso del relativo *que* y de algunos pronombres personales), las etiquetas con valor *C* (común) para el atributo de género se han reescrito dos veces: una como masculinos y otra como femeninos. Lo mismo ha sucedido con los elementos con valor *N* (invariable) para el atributo de número: se han reescrito como singulares y plurales. Las formas de género neutro se han reescrito como masculinos, dado que imponen concordancia en masculino singular. Si los atributos de género y número no aparecen especificados (caso de los nombres propios o de palabras que el analizador no reconoce) las etiquetas se han reescrito cuatro veces para que la desambiguación puede hacerse por el contexto.

4.3.3.2. Núcleo del sn

Los elementos que pueden aparecer como núcleo del sintagma nominal son los nombres (comunes y propios), un grupo de nombres coordinados y los pronombres. Sin embargo, los relativos y los pronombres personales átonos se etiquetan por separado dado que son elementos que nunca pueden recibir complementos.

Los nombres comunes aparecen como núcleos en las reglas cuya parte izquierda es *grup-nom*, mientras que los pronombres se reescriben directamente desde *sn*, dado que no reciben complementos[37]. Sobre la aparición del nombre propio con complementación, hay que señalar que si bien es cierto que la mayoría de los nombres propios en español no precisa de determinante por su naturaleza semántica definida, es decir, porque se comportan como sintagmas nominales definidos, no lo es menos que en algunos casos estos nombres se acompañan de un artículo definido (*La Coruña, Los Alpes*). Este artículo tiene un valor expletivo, es decir, está vacío de contenido[38], lo mismo que ocurre cuando la partícula aparece ante el nombre propio en contextos coloquiales (*La María*). Además del artículo, pueden preceder al nombre propio los demostrativos o los posesivos (*Este Alexis; Mi David*). En estos casos lo que se produce es una recategorización del nombre propio como nombre común, que adopta, por este hecho, una interpretación contrastiva. Sin embargo, hay otra interpretación posible, la afectiva[39]. Por estos motivos, hemos decidido tratar el nombre propio igual que el nombre común. Además, puede también aparecer un artículo si el nombre propio aparece con algún adjetivo o equivalente que lo modifique: *el antedicho Natalio*, *la sitiada Cuba*, *la imperial Todelo* o *el Toledo judío*[40].

Ejemplos de nombres propios con complementación y/o determinación los hallamos en las siguientes frases:

(4.16) (a) *¿Qué perdí en él, oh dioses? Soy un* **Teseo** *acojonado* (a13).
(b) *Cuando escribo esto la* **Madre_Coraje** *peruana acaba de ser reventada por los senderistas* (a14).
(c) *El* **Madrid** *oficial vivía bajo el terror y la paranoia* (a25).

[37] Sin embargo, véase la sección 4.3.3.6 para los casos en que los pronombres reciben especificadores.
[38] Rigau (1999): p. 320.
[39] Rigau (1999): p. 321.
[40] Ejemplos tomados de Martínez (1999): pp. 2717-2718.

En el análisis de la última oración puede observarse cómo se lleva a cabo la desambiguación del género y el número del nombre propio y la del género del adjetivo a través de los rasgos del determinante:

```
S_[
   sn_[ espec-ms_[ El_da0ms0 ]
        grup-nom-ms_[ Madrid_np00000
        s-a-ms_[ oficial_aq0cs0 ] ] ]
   grup-verb_[ vivía_vmii1s0 ]
   grup-sp_[ prep_[ bajo_sps00 ]
             sn_[ espec-ms_[ el_da0ms0 ] grup-nom-ms_[ terror_ncms000 ] ] ]
   coord_[ y_cc ]
   sn_[ espec-fs_[ la_da0fs0 ] grup-nom-fs_[ paranoia_ncfs000 ] ] ]
```

Las reglas para el grupo nominal son las siguientes[41]:

(a) *grup-nom-ms* → *n-ms.*
(b) *grup-nom-ms* → *n-ms, n-fs.*
(c) *grup-nom-ms* → *n-ms, sp-de.*
(d) *grup-nom-mp* → *w-mp.*
(e) *grup-nom-ms* → *w-ms, w-ms.*
(f) *grup-nom-ms* → *n-ms, w-ms.*
(g) *grup-nom-ms* → *n-ms, s-a-ms.*
(h) *grup-nom-ms* → *s-a-ms, grup-nom-ms.*
(i) *grup-nom-ms* → *w-ms, s-a-ms.*
(j) *grup-nom-fp* → *grup-c-nom-fp.*

Las reglas (a) y (d) son de aplicación en el caso de que el grupo nominal esté formado sólo por un nombre común o propio. En el resto de reglas aparecen nombres y complementos en forma de otro nombre (b, e, f), sintagma adjetivo (g, h, i) o sintagma preposicional (c).

Otra posibilidad es que aparezcan nombres coordinados (regla j). El alcance de la coordinación va desde lo que podemos llamar la coordinación léxica (coordinación entre palabras aisladas) hasta la coordinación de grandes estructuras oracionales. El tratamiento de la coordinación que hemos hecho en **GramEsp** queda restringido sólo a la coordinación léxica de nombres y adjetivos. En esta sección comentamos la coordinación referida a los nombres.

La casuística de la coordinación es muy amplia. A continuación presentamos algunos ejemplos de coordinación de elementos nominales (nombres o sintagmas nominales):

(4.17) (a) *es señal de desviación y salida* (a26).
(b) *de una riqueza y calidad excepcionales* (a10).

[41] La enumeración de las reglas no es exhaustiva. Teniendo en cuenta que la inmensa mayoría de ellas se ven afectadas por el fenómeno de la concordancia, presentamos aquí sólo una regla de cada tipo. Todas las reglas aparecen al final, en el apéndice B, sección B.3.1.

(c) *Reciclaje, ahorro, aprovechamiento de los residuos y lucha contra el despilfarro* (a15).
(d) *incendia sopranos y grandes duquesas, pavarotis y rasputines, liceos y palacios de invierno, leidicháterlis y susis* (c1).
(e) *el mundo que fueron mis amores y la memoria que me los restringe* (a13).

Sólo los ejemplos 4.17 (a-b) pueden considerarse coordinación léxica; en el primer caso ninguno de los nombres va acompañado ni de determinantes ni de adjetivos u otros complementos; en el ejemplo 4.17 (c) dos de los nombres coordinados reciben complementación (*aprovechamiento* y *lucha*); en 4.17 (d) hay coordinaciones léxicas a distintos niveles y en algunos casos los nombres reciben complementación; finalmente, en 4.17 (e) los dos sintagmas nominales coordinados presentan una estructura muy compleja.

Los únicos casos aquí tratados son los que se corresponden con estructuras del tipo 4.17 (a-b), es decir la coordinación léxica. Se ha considerado que el resultado de una coordinación es siempre plural, y se han tenido en cuenta las reglas de la concordancia (cf. Martínez (1999)) en lo referente al género y el número, lo que hace que el número de reglas sea elevado. Algunas de estas reglas aparecen a continuación[42]:

grup-c-nom-mp → *n-mp, coord, n-fs.*
grup-c-nom-mp → *n-ms, coord, n-ms.*
grup-c-nom-mp → *n-ms, coord, n-mp.*

En lo que respecta a la concordancia de los grupos nominales complejos se ha tenido en cuenta los rasgos concordantes con el adjetivo que sigue al grupo, y no los rasgos del determinante, que siempre concuerda con el primer elemento.

Las posibilidades de concordancia de los adjetivos con nombres coordinados es muy amplia[43], y aquí sólo hemos tenido en cuenta la que se da cuando en adjetivo se pospone al grupo coordinado.

Algunas frases en que estas reglas son de aplicación son las siguientes:

(4.18) *y con multitud de poetas, músicos, esclavos, soldados y bufones* (a1).
pues hay currinches y relapsos (a1).
Es curioso cómo los críticos y estudiosos de los últimos años (a10).
o donde padre y madre trabajan (a11).

A continuación aparece el análisis de la primera secuencia:

```
S_[ coord_[ y_cc ]
    grup-sp_[ prep_[ con_sps00 ]
              sn_[ grup-nom-fs_[ multitud_ncfs000
                   sp-de_[ prep_[ de_sps00 ]
                   sn_[ grup-nom-mp_[
                                poetas_ncmp000 ,_Fc
                                músicos_ncmp000 ,_Fc
```

[42] El resto aparece en el apéndice B sección B.3.1.1.
[43] Martínez (1999): pp. 2710-2712 y pp. 2738-39.

```
                                esclavos_ncmp000 ,_Fc
                                soldados_ncmp000
                                coord_[ y_cc ]
                                bufones_ncmp000 ] ] ] ] ] ] ]
```

Casos de coordinación que no se resuelven correctamente desde un punto de vista lingüístico son los que aparecen a continuación. En el primer caso las agrupaciones de nombres que lleva a cabo la gramática no se corresponden con las que marcaría un lingüista, que corresponderían a *[sopranos y grandes duquesas] , [pavarotis y rasputines], [liceos y palacios de invierno], [leidicháterlis y susis].* En el segundo caso de trata de la coordinación de dos sintagmas nominales complejos: *[el mundo que fueron mis amores] y [la memoria que me los restituye].*

```
incendia sopranos y grandes duquesas, pavarotis y rasputines, liceos y palacios
de invierno, leidicháterlis y susis (c1)
S_[ grup-verb_[ incendia_vmip3s0 ]
    sn_[ grup-nom-fp_[ sopranos_ncfp000 ] ]
    coord_[ y_cc ]
    sn_[ grup-nom-fp_[
         s-a-fp_[ grandes_aq0cp0 ]
         duquesas_ncfp000 ,_Fc
         pavarotis_nc00000
         coord_[ y_cc ]
         rasputines_nc00000 ] ]
    ,_Fc
    sn_[
    grup-nom-mp_[ liceos_ncmp000
                  coord_[ y_cc ]
                  palacios_ncmp000 ] ]
    grup-sp_[
              prep_[ de_sps00 ]
              sn_[ grup-nom-mp_[
                       invierno_ncms000 ,_Fc
                       leidicháterlis_nc00000
                       coord_[ y_cc ]
                       susis_nc00000 ] ] ] ]

el mundo que fueron mis amores y la memoria que me los restituye (a13)
S_[ sn_[ espec-ms_[ el_da0ms0 ] grup-nom-ms_[ mundo_ncms000 ] ]
    relatiu_[ que_pr0cn000 ]
    grup-verb_[ fueron_vsis3p0 ]
    sn_[ espec-mp_[ mis_dp1cps ]
         grup-nom-mp_[ amores_ncmp000 ] ]
    coord_[ y_cc ]
    sn_[ espec-fs_[ la_da0fs0 ]
```

```
        grup-nom-fs_[ memoria_ncfs000 ] ]
    relatiu_[ que_pr0cn000 ]
    patons_[ me_pp1cs000 ]
    patons_[ los_pp3mpa00 ]
    grup-verb_[ restituye_vmip3s0 ] ]
```

4.3.3.3. Modificadores del sintagma nominal

Como se comentó anteriormente, los complementos del núcleo del sintagma nominal pueden ser otro nombre, común o propio (reglas a,c,d), un sintagma adjetivo (reglas e,f,g) o un sintagma preposicional *sp-de* (regla b).

(a) *grup-nom-ms* → *n-ms, n-fs.*
(b) *grup-nom-ms* → *n-ms, sp-de.*
(c) *grup-nom-ms* → *w-ms, w-ms.*
(d) *grup-nom-ms* → *n-ms, w-ms.*
(e) *grup-nom-ms* → *n-ms, s-a-ms.*
(f) *grup-nom-ms* → *s-a-ms, grup-nom-ms.*
(g) *grup-nom-ms* → *w-ms, s-a-ms.*

Los casos de nombre complementando a otro nombre tratados en **GramEsp** no son las aposiciones explicativas sino la concatenación de nombres (aposiciones especificativas o secuencias de nombres no apositivas), como en los ejemplos 4.19.

(4.19) *La etapa reina* (d2).
Un equipo cantera (d2).
Otra carrera contrarreloj (d2).
Los rayos gamma (dc1).
Un buque estadounidense repleto de proyectiles de gas mostaza (dc2).

En estos casos el nombre que determina los rasgos de concordancia del grupo nominal es el primero, como puede observarse en los ejemplos anteriores.

```
sn_[
    espec-ms_[ un_di0fs0 ]
    grup-nom-ms_[ equipo_ncms000 cantera_ncfs000 ] ]


sn_[
    grup-nom-mp_[ proyectiles_ncmp000
                  sp-de_[ prep_[ de_sps00 ]
                  sn_[
                       grup-nom-ms_[ gas_ncms000 mostaza_ncfs000 ] ] ] ] ]
```

Las aposiciones explicativas no se han tratado debido a que a partir de rasgos puramente formales es imposible identificarlas: la concordancia no siempre se manifiesta. Por otra

parte, los signos de puntuación que parentizan la aposición pueden aparecer en estructuras no apositivas.

Los sintagmas adjetivos (cf. sección 4.3.4 para los elementos incluidos en esta estructura) pueden preceder o seguir al nombre núcleo del sintagma (tanto si es un nombre propio como común), o aparecer en ambas posiciones, lo que se analiza mediante la última regla de las anteriormente mencionadas, que es recursiva. Ejemplos de estas combinaciones aparecen en 4.20[44].

(4.20) (a) *Como hacía* **buen tiempo** *decidió salir* (a14).
(b) *La* **verdadera heroicidad** *no es un* **acto único** (a14).
(c) *en casos de* **buena capacidad mercantil**... (a21).

El análisis de la segunda de las frases anteriores es el siguiente:

```
S_[ sn_[
       espec-fs_[ la_da0fs0 ]
       grup-nom-fs_[
             s-a-fs_[ verdadera_aq0fs0 ]
             heroicidad_ncfs000 ] ]
    neg_[ no_rn ]
    grup-verb_[ es_vsip3s0 ]
    sn_[
       espec-ms_[ un_di0ms0 ]
       grup-nom-ms_[ acto_ncms000
                     s-a-ms_[ único_aq0ms0 ] ] ] ]
```

Como última posibilidad de complementación del nombre común se ha tenido en cuenta la aparición de un sintagma preposicional pospuesto al nombre. Se trata aquí de una ampliación de la noción de *chunk*, dado que bajo esta concepción del análisis sintáctico parcial, los sintagmas preposicionales deberían quedar como nodos independientes. Sin embargo, no todos los sintagmas preposicionales que siguen a un nombre se han incluido en el *chunk* nominal. Sólo se ha tenido en cuenta el sintagma preposicional encabezado por la preposición *de* y sólo en el caso de que haya una relación de inmediatez entre el sintagma y el nombre. Esta decisión se ha tomado después de realizar varios estudios sobre corpus, que se detallan a continuación:

1. En primer lugar, para verificar que este sintagma preposicional complementa realmente al nombre que le antecede de forma inmediata, se ha tomado un fragmento de 210 frases del corpus LexEsp, del que se han extraído las secuencias de [nombre_común + preposición_*de*]. Los resultados del análisis realizado han sido los siguientes:

[44] *En términos generales, los adjetivos modificadores pueden aparecer tanto antepuestos como pospuestos al nombre (*el extravagante bolso anaranjado / la agraciada señora veneciana*) dentro del sintagma nominal. Los adjetivos que admiten adverbios de grado, esto es, los calificativos y algunos de los intensionales (*posible *pero no* único*), pueden ir acompañados de estos intensificadores en todas sus posiciones sintácticas (*el muy estúpido profesor extraordinariamente gordo*)* Demonte (1999) p. 182.

a) Ejemplos recogidos: 237

b) Modificadores del nombre: 230 (97 %)

c) Ejemplos ambiguos: 3 (1,26 %)[45].

d) Modificadores de otro elemento: 4 (1,68 %)[46]

2. El estudio anterior se ha realizado sin tener en cuenta la presencia de algún otro elemento entre el nombre y la preposición; pero, evidentemente, esta posibilidad existe, tal como se muestra en los ejemplos siguientes:

(4.22) (a) *la cantidad total de sólidos en suspensión*
(b) *Son un mito básico de nuestra sociedad*
(c) *Las miradas irónicas de los provocadores*
(d) *los espadones goteantes de sangre*
(e) *son un pozo lleno de ecos*
(f) *con materiales susceptibles de ser erosionados*[47]

En este segundo caso, sin embargo, la relación entre los que dependen del nombre anterior o de otro elemento no es tan clara como lo era en el caso anterior en que había una relación de inmediatez entre el nombre y la preposición, por lo que se ha optado por no tener en cuenta estas estructuras complejas en **GramEsp**, donde estos sintagmas preposicionales aparecen como nodos independientes, tal como se muestra a continuación:

```
S_[ sn_[
        espec-fs_[ la_da0fs0 ]
        grup-nom-fs_[ cantidad_ncfs000
                      s-a-fs_[ total_aq0cs0 ] ] ]
    grup-sp_[ prep_[ de_sps00 ]
              sn_[ grup-nom-mp_[ sólidos_ncmp000 ] ] ]
    grup-sp_[ prep_[ en_sps00 ]
              sn_[ grup-nom-fs_[ suspensión_ncfs000 ] ] ] ]


S_[ grup-sp_[
        prep_[ con_sps00 ]
        sn_[ grup-nom-mp_[ materiales_ncmp000
```

[45]Se entiende por *casos ambiguos* aquellos en que el sintagma preposicional podía interpretarse como modificador del nombre precedente o de otro elemento anterior en el texto.

[46]Las frases que recogen estos casos son las siguientes (se marca el análisis incorrecto):

(4.21) (a) ** un día de [verano [de calor rabioso]]*
(b) ** me silban los [oídos [de la presión]]*
(c) ** las antiguas minas romanas de [oro [del norte de León]]*
(d) ** un pueblo de [chabolas [de varios cientos de miles de personas]].*

[47]Todos los ejemplos citados en este apartado provienen de LexEsp.

```
          s-a-mp_[ susceptibles_aq0cp0 ] ] ] ]
grup-sp_[ prep_[ de_sps00 ]
           infinitiu_[ ser_vsn0000 erosionados_vmp00pm ] ] ]
```

3. Por otra parte, también se ha procedido a un estudio similar con las demás preposiciones y en idénticas condiciones, a saber, contigüidad entre el nombre y la preposición. En esta ocasión, los datos demuestran que en la mayoría de los casos (aunque ciertamente no en todos) unir el sintagma preposicional al nombre produciría errores de análisis, por lo que se ha optado por mantener ambos elementos separados. A continuación se presentan ejemplos de sintagmas preposicionales con preposición distinta a *de* (con cada preposición, excepto *cabe* y *so*, que no aparecen en el corpus tras un sustantivo):

(4.23) (a) *De pronto tumbaron la puerta* **a** *golpes*
(b) *... por el juego que desarrolló su equipo* **ante** *el Salamanca*
(c) *Daba gusto pisar el suelo* **bajo** *la girola*
(d) *... confundir el puro* **con** *las témporas*
(e) *... jugar un partido* **contra** *un equipo iraquí*
(f) *Determinó la existencia de zonas de descarga* **desde** *el medio hiporreico hasta la superficie*
(g) *Para este encuentro, que comenzará a las 21.00 horas* **en** *el estadio Luis_Casanova...*
(h) *... intercambios de materia y energía* **entre** *las aguas superficiales y las subterráneas*
(i) *Mueven todo el cuerpo* **hacia** *un lado*
(j) *Se aprovecha su caudal* **hasta** *dejarlos casi exhaustos*
(k) *Nosotras somos un misterio* **para** *ellos*
(l) *Va en moto* **por** *una autopista*
(m) *La resistencia a ser degradada ordena los materiales* **según** *su potencialidad energética*
(n) *Pasar las horas del día* **sin** *hacer nada*
(o) *... participar en un debate televisivo de una hora de duración* **sobre** *la homosexualidad*
(p) *Otras tres se perdían de vista en aquellos momentos* **tras** *el altar*

Tras este estudio, como se ha comentado al principio de este apartado, el único sintagma preposicional que se incluye en el sintagma nominal es el que va introducido por la preposición *de*, siempre y cuando esta preposición vaya inmediatamente pospuesta al nombre. Las reglas que expresan esta relación son:

grup-nom-ms → n-ms, sp-de.
grup-nom-mp → n-mp, sp-de.
grup-nom-fs → n-fs, sp-de.
grup-nom-fp → n-fp, sp-de.
sp-de → prep(de), sn.
sp-de → prepc-ms(del), grup-nom-ms.

Su aplicación puede observarse en el siguiente sintagma nominal:

```
sn_[ espec-ms_[ el_da0ms0 ]
     grup-nom-ms_[
        número_ncms000
        sp-de_[ prep_[ de_sps00 ]
                sn_[ grup-nom-fp_[
                        posibilidades_ncfp000
                        sp-de_[ prep_[ de_sps00 ]
                               infinitiu_[ subir_vmn00000]
```

4.3.3.4. Especificadores del núcleo

La etiqueta *espec-* (especificador) agrupa, manteniendo la información morfológica sobre el género y el número, todos los adjetivos determinativos, sus posibles combinaciones y, además, ciertos usos de los adverbios *más, menos, casi*, que se han reescrito como *cuantif*[48].

Las agrupaciones posibles de determinantes se han expresado en función de los ejemplos tomados del Corpus CLiC-TALP así como en función de la teoría lingüística Rigau (1999) [49]:

espec-fp → grup-complex-spec-fp.
grup-complex-spec-fp → dem-fp, numer-fp.
grup-complex-spec-fp → indef-fp, coord, indef-fp.
grup-complex-spec-fp → indef-fp,indef-fp.

Algunos ejemplos de combinación de determinantes extraídos del corpus son los siguientes:

(4.24) (a) **el primer** *turista de nuestra cultura* (a13).
(b) *o a* **todos los** *voceros que han contaminado el aire de enero* (c1).
(c) *aparte_ de en* **el mismo** *equipo , en el mismo podio* (d2).
(d) *Hubo un grupo que formamos* **unos cuantos** *corredores* (d2).
(e) *existe* **toda una** *gama de virus que provoca este tipo de enfermedades* (dc10).
(f) **Las bastantes pocas** *aptitudes*[50].

[48] Tal como ya se comentó en la sección 4.3.1.1.
[49] Para una lista completa de estas reglas, puede consultarse el apéndice B sección B.3.1.2.
[50] Ejemplo tomado de Rigau (1999): p. 331.

4.3.3.5. La sustantivación sintáctica

El único caso de sustantivación que hemos considerado ha sido el del artículo definido antepuesto a sintagmas adjetivos y preposicionales. En este caso hemos utilizado el pseudo-terminal del artículo.

sn → *j-ms, s-a-ms.*
sn → *j-fs, s-a-fs.*
sn → *j-mp, s-a-mp.*
sn → *j-fp, s-a-fp.*
sn → *j-ms, grup-sp.*
sn → *j-fs, grup-sp.*
sn → *j-mp, grup-sp.*
sn → *j-fp, grup-sp.*

Casos de sustantivación no considerados han sido la sustantivación de las relativas sin antecedente y la sustantivación del adverbio. En el primer caso se hace necesario delimitar previamente la relativa:

```
S_[
   espec-mp_[ los_da0mp0 ]
   relatiu_[ que_pr0cn000 ]
   grup-verb_[ vayan_vmsp3p0 a_sps00 infinitiu_[ salir_vmn0000 ] ]
   grup-sp_[ del_spcms s-a-ms_[ trágico_aq0ms0 ] ]
   sn_[ grup-nom-mp_[ caos_ncmn000 ] ]
(a21)
```

En el segundo, una regla que sustantivara sintagmas adverbiales entraría en conflicto con casos como el que presentamos aquí, en que no es el adverbio el que está sustantivado sino el adjetivo al que el adverbio complementa:

```
sn_[ los_da0mp0
     s-a-mp_[ sadv_[ menos_rg ]
              pusilánimes_aq0cp0 ] ]
(a25)
```

Estos tipos de sustantivación no se han tratado porque formalmente no es posible distinguirlos.

4.3.3.6. Los pronombres

Los pronombres se han incluido directamente dentro del sintagma nominal, aunque también pueden aparecer con algunos especificadores. Como esta combinatoria está restringida a casos muy concretos tanto de pronombres como de especificadores, se ha optado por introducir esta casuística en el seno de "sn", ya que mantenerlos bajo la etiqueta *grup-nom* hubiera significado que habrían entrado en combinación con todos los determinantes y

con los modificadores del nombre; en cambio, de este modo, se controlan mucho mejor los contextos de aparición de estas formas.

Para los casos en que los pronombres se reescriben directamente desde el sintagma nominal, las reglas son:

(a) *sn* → *pron-ms.*
(b) *sn* → *pron.*
(c) *sn* → *psubj-fp, num-fp.*
(d) *sn* → *psubj-ms, indef-ms.*
(e) *sn* → *pdem-ms, indef-ms.*
(f) *sn* → *j-ms, pindef-ms.*
(g) *sn* → *pindef-ms, s-a-ms.*
(h) *sn* → *indef-ms, pnum-ms.*
(i) *sn* → *indef-ms, pindef-ms.*
(j) *sn* → *pindef-mp, prep(de), psubj-mp.*
(k) *sn* → *pindef-mp, sp-de.*

Las dos primeras reglas corresponden a los casos menos marcados, en que el pronombre es el único elemento del sintagma nominal. Las otras reglas incluyen determinantes y modificadores de los pronombres. Ejemplos de estos casos son:

(4.25) *Pero* **ella** *misma terminó declarándose editora a_secas* (c2). regla (d)
El **uno** *era blanco y el* **otro** *rojizo* (a22). regla (f)
No debía hablar así porque los hombres hacemos **algo** *parecido* (t4). regla (g)
El frío ambiental y la visita anterior a los barracones donde quizá murió **alguno** *de los_suyos, no daba para alegrías* (n1). regla (k)

Los pronombres átonos, los clíticos, no se incluyen en el sintagma nominal dado que en ningún caso pueden recibir ni complementación ni determinación. Constituyen, en **GramEsp**, nodos unarios e independientes. Por su parte, las formas de *se* marca de oración impersonal o pasiva refleja se etiquetan como *morfema-verbal*, mientras que las formas *me, te, se, nos, os* marca de verbo pronominal reciben la etiqueta *morf-pron*[51].

4.3.4. El sintagma adjetivo

El núcleo del sintagma adjetivo es el adjetivo ordinal o bien calificativo incluyendo el participio, esto es, todos los elementos que se reescriben como *a-* (regla (a)).

Hay casos en que estos elementos aparecen coordinados (regla (c)) o yuxtapuestos (regla (b)) o simplemente concatenados (regla (d)).

(a) *s-a-mp* → *a-mp.*
(b) *s-a-mp* → *s-a-mp, Fc, s-a-mp.*
(c) *s-a-mp* → *s-a-mp, coord, s-a-mp.*
(d) *s-a-ms* → *a-ms, s-a-ms.*
(e) *s-a-mp* → *sadv, a-mp.*

[51] En el apéndice B sección B.2.8 aparecen todas las etiquetas para los clíticos.

La coordinación de adjetivos se ha tratado al mismo nivel que la coordinación de nombres: a nivel exclusivamente léxico, como en el caso de la siguiente oración que aparece también analizada:

(4.26) *Toda la derecha* **política y económica** *tiene de la limpieza un concepto* **penitencial y detergente, revanchista y represivo** (c1).

```
S_[   sn_[
        espec-fs_[ toda_di0fs0 la_da0fs0 ]
        grup-nom-fs_[
           derecha_ncfs000
           s-a-fs_[ s-a-fs_[ política_aq0fs0 ]
             coord_[ y_cc ]
             s-a-fs_[ económica_aq0fs0 ] ] ] ]
      grup-verb_[ tiene_vmip3s0 ]
      grup-sp_[
        prep_[ de_sps00 ]
        sn_[ espec-fs_[ la_da0fs0 ]
             grup-nom-fs_[ limpieza_ncfs000 ] ] ]
      sn_[ espec-ms_[ un_di0ms0 ]
           grup-nom-ms_[ concepto_ncms000
             s-a-ms_[ s-a-ms_[ penitencial_aq0cs0 ]
             coord_[ y_cc ]
             s-a-ms_[ s-a-ms_[ detergente_aq0cs0 ]
             ,_Fc
             s-a-ms_[
                s-a-ms_[ revanchista_aq0cs0 ]
                coord_[ y_cc ]
                s-a-ms_[ represivo_aq0ms0 ] ] ] ] ] ] ]
```

Por último, los adjetivos pueden recibir complementos en forma de adverbios (regla (e))[52].

Finalmente, hay una regla que agrupa todos los sintagmas adjetivos con marcas de concordancia en un único nodo: *sa*. Esta etiqueta aparece en @GROUP, de modo que sólo forma parte de la salida del árbol en el caso de que no esté incluida en el sintagma nominal. De este modo, se asegura la concordancia en el seno del sintagma nominal.

Los otros tipos de complementación del adjetivo (adjetivo seguido de sintagma preposicional y construcciones comparativas) no se han considerado en esta gramática, dada la imposibilidad de definir criterios formales claros de agrupación.

4.3.5. El sintagma adverbial

El sintagma adverbial está formado fundamentalmente por adverbios, aunque se han introducido reglas específicas para algunas formas que siempre (o en la mayoría de las

[52] Todas las reglas de sintagma adjetivo aparecen en el apéndice B sección B.4.

ocasiones) aparecen seguidos de preposición (Kovacci (1999)). Estos adverbios son: *cerca, lejos, arriba, abajo, después, antes, fuera, dentro, delante, detrás, encima, debajo, más, menos, enfrente* seguidos de la preposición *de* y *frente* y *junto* seguidos de la preposición *a*[53]. El modo en que se han introducido ha sido declarando tanto el adverbio como la preposición que lo acompaña de modo literal, tal como aparece a continuación[54]:

sadv → *adv.*
sadv → *adv-interrog.*
sadv → *adv(cerca), prep(de), sn.*
sadv → *adv(lejos), prep(de), sn.*
sadv → *adv(frente), prep(a), sn.*
sadv → *adv(junto), prep(a), sn.*

No se ha considerado en la gramática ni la concatenación ni la coordinación de adverbios, por los errores en el análisis que ello podía implicar, ya que no siempre que dos adverbios aparecen seguidos forman un sintagma, tal como puede apreciarse en los ejemplos siguientes:

(4.27) *Aunque* **quizá ya** *no lograra reponerse* (a14).

4.3.6. El grupo preposicional

El sintagma preposicional está formado por una preposición seguida de otro sintagma nominal, adjetivo o adverbial, además de otros elementos como infinitivos, cifras o fechas[55].

1. El término de la preposición es un sintagma nominal. Dado que las contracciones se tratan como una sola palabra, hay que prever que los sintagmas nominales con un nombre (o un pronombre) masculino singular y las preposiciones *a, de* deben tratarse con una regla aparte.

   grup-sp → *prep, sn.*
   grup-sp → *prepc-ms, grup-nom-ms.*

   Los pronombres personales en caso oblicuo también pueden ir precedidos de preposición. Esto se representa de dos formas. La primera regla reescribe los pronombres tónicos *mí, ti, sí*, mientras que las siguientes son para las formas pronominales que ya incorporan la preposición en su forma.

   grup-sp → *prep, ptonic.*
   grup-sp → *pp1cso00(conmigo).*
   grup-sp → *pp2cso00(contigo).*
   grup-sp → *pp3cso00(consigo).*

[53] En el apéndice B sección B.5 aparecen todas estas reglas.
[54] El término de esta construcción puede ser también un sintagma adverbial.
[55] Todas las reglas para el sintagma preposicional aparecen en el apéndice B sección B.6.

Ejemplos de aplicación de las reglas anteriores son:

```
...
   conj-subord_[ que_cs ]
   grup-verb_[ colabore_vmsp1s0 ]
   grup-sp_[ conmigo_pp1cso00 ]
   grup-sp_[ prep_[ en_sps00 ]
      sn_[ espec-fs_[ la_da0fs0 ]
           grup-nom-fs_[ campaña_ncfs000
             s-a-fs_[ electoral_aq0cs0 s-a-fs_[ vasca_aq0fs0 ]]]]]
(r2)
```

2. Los adjetivos y los adverbios también pueden ser el término de una preposición:

   grup-sp → *prep, s-a-ms.*
   grup-sp → *prep, sadv.*

```
...
   grup-sp_[
     prep_[ sin_sps00 ]
     infinitiu_[ haber_van0000 visto_vmp00sm ] ]
   sn_[
     espec-ms_[ un_di0ms0 ]
     grup-nom-ms_[ bosnio_ncms000 ] ]
   grup-sp_[
     prep_[ de_sps00 ]
     sadv_[ cerca_rg ] ]
(a22)
```

3. Por último, otros elementos que pueden combinarse con la preposición son el infinitivo, las cifras y las fechas:

   grup-sp → *prep, infinitiu.*
   grup-sp → *prepc-ms, infinitiu.*
   grup-sp → *prep, numero.*
   grup-sp → *prep, data.*
   grup-sp → *prepc-ms, w.*

```
grup-sp_[
  prep_[ sin_sps00 ]
  infinitiu_[ haber_van0000
              llegado_vmp00sm
              a_sps00
              infinitiu_[ rectificar_vmn0000 ] ] ]
(t6)
```

Se ha utilizado la etiqueta *grupo preposicional* y no la de sintagma preposicional porque es una estructura no recursiva, es decir, no se incluye a sí misma; la única excepción a esto se dará en el caso de que el sintagma nominal término de la preposición incluya un *sp-de*. Pero en los otros casos no hay recursividad, de modo que si aparecen dos o más sintagmas preposicionales (que no contengan la preposición *de*) el analizador los tratará como unidades separadas.

4.3.7. El grupo verbal

En este apartado se comentan todas aquellas reglas de la gramática que afectan a las formas verbales finitas sin tener en cuenta la complementación o la modificación. Como ya se ha señalado al inicio del capítulo, **GramEsp** realiza únicamente agrupaciones sintagmáticas y, si bien es cierto que en el caso del sintagma nominal las agrupaciones que se tratan pueden llegar a ser bastante amplias, no lo es menos el hecho de que a nivel formal o superficial no puede realizarse la agrupación del sintagma verbal (entendido como predicado oracional). Para ello sería preciso poder determinar la naturaleza de los complementos verbales, esto es, clasificarlos como argumentos o como adjuntos, para lo que hay que disponer de información semántica tanto del propio verbo como de los complementos.

Por todo ello, aquí se propone el constituyente *grup-verb* que incluye exclusivamente formas verbales finitas (simples o complejas) como *Llegó* o *Vas a tener que volver a empezar a trabajar* que aparece analizada a continuación:

```
grup-verb_[ vas_vmip2s0 a_sps00
              infinitiu_[ tener_vmn0000 que_cs
                 infinitiu_[ volver_vmn0000 a_sps00
                    infinitiu_[ empezar_vmn0000 a_sps00
                       infinitiu_[ trabajar_vmn0000 ] ] ] ] ]
```

Las formas simples de los verbos, reescritas como pseudo-terminales, quedan incluidas bajo la etiqueta *grup-verb*:

$$grup\text{-}verb \rightarrow verb.$$

A continuación nos ocupamos de las formas finitas complejas: las que forman los tiempos compuestos de la conjugación, la llamada *voz pasiva* y las perífrasis verbales.

Los tiempos compuestos de la conjugación se forman con el verbo auxiliar *haber* seguido de un participio en la forma masculina singular[56]:

$$verb \rightarrow vaux,\ parti.$$

Del mismo modo, la voz pasiva se forma con el verbo semiauxiliar *ser* seguido de un participio en cualquiera de sus cuatro formas; y en el caso de las formas compuestas de la

[56] Recuérdese que *vaux* es la etiqueta para las formas del verbo haber; que *parti* reescribe todas las formas masculinas singulares de los participios, que son las que aparecen en los tiempos compuestos.

voz pasiva, con el auxiliar *haber*, seguido del participio del verbo *ser* en su forma masculina singular, seguido de un participio en cualquiera de sus formas flexivas[57]:

verb-pass → *vser, parti-flex.*
verb-pass → *vaux, parti-ser, parti-flex.*

El último tipo de forma verbal finita compleja lo constituyen las perífrasis verbales, a las que dedicamos la siguiente sección.

4.3.7.1. Las perífrasis verbales

Una perífrasis verbal es la unión de dos o más verbos que constituyen un solo 'núcleo' del predicado. El primer verbo, llamado 'auxiliar', comporta las informaciones morfológicas de número y persona, y se conjuga en todas (o en parte de) las formas o tiempos de la conjugación. El segundo verbo, llamado 'principal' o 'auxiliado', debe aparecer en infinitivo, gerundio o participio, es decir, en una forma no personal[58].

Lo importante aquí es el hecho de que ambas formas funcionen como núcleo del predicado (del mismo modo que en los tiempos compuestos de la conjugación en que *haber* aporta la información morfológica y siempre aparece seguido de un participio). De hecho, las formas compuestas de los tiempos verbales pueden considerarse como una perífrasis de participio en que éste último ha perdido la capacidad de flexión. Asimismo, pueden también considerarse perífrasis las formas *pasivas* de los tiempos verbales.

Como se verá más adelante, el fenómeno perifrástico es gradual y depende fundamentalmente del significado de las piezas léxicas y del contexto es que éstas aparecen[59].

Por otra parte, las perífrasis no siempre aparecen como secuencia única: *el hecho de que los verbos de una perífrasis constituyan un solo núcleo del predicado, no implica que entre ellos no puedan introducirse otros elementos. Así, por ejemplo, en la mayoría de los casos se pueden introducir adverbios, locuciones adverbiales o secuencias nominales de complemento circunstancial [...] o bien incisos varios. [...] o bien el sujeto [...]. Todo esto quiere decir que en la mayoría de las perífrasis el grado de conexión entre 'auxiliar' y 'auxiliado' no es tan fuerte como el que se da en los tiempos compuestos*[60]. Las frases siguientes ejemplifican este fenómeno.

(4.28) (a) *No podemos* **en absoluto** *establecer diferencias.*
(b) *Tuvimos* **el otro día** *que marcharnos con urgencia.*
(c) *Empezó* **de repente** *a llover.*
(d) *Debes,* **si puedes**, *intentarlo.*
(e) *Acaban,* **hace un momento**, *de llamar a la puerta.*
(f) *¿Puede* **alguien** *decirnos lo que pasó?*[61]

[57] Las etiquetas *vser* son las que reescriben las formas del verbo *ser*; *parti-flex* reescribe las cuatro formas de los participios de todos los verbos; *parti-ser* reescribe la forma *sido*.
[58] Gómez (1999): p. 3325.
[59] Yllera (1999): pp. 3400 y 3428-29; y Gómez (1999): pp. 3334-35.
[60] Gómez (1999): pp. 3325-26.
[61] Ejemplos tomados de Gómez (1999): p. 3326.

Dado el carácter lineal del analizador de *charts*, el análisis no podrá tener en cuenta estas escisiones en las perífrasis, por lo que estos casos no podrán quedar recogidos. Ejemplos de este fenómeno tomados del corpus son:

(4.29) *Tampoco ellos podrán jamás recuperar su antiguo ser* (a12).
¿Se tendrá Romario que ir a Río? (d2)

El análisis de **GramEsp** para estas estructuras es el siguiente:

```
S_[ ¿_Fia
    morf-pron_[ Se_p0300000 ]
    grup-verb_[ tendrá_vmif3s0 ]
    sn_[
       grup-nom-fp_[ Romario_np00000 ] ]
    conj-subord_[ que_cs ]
    infinitiu_[ ir_vmn0000 ]
    grup-sp_[ prep_[ a_sps00 ]
              sn_[ grup-nom-fp_[ Río_np00000 ] ] ]
    ?_Fit ]
```

Las perífrasis se han declarado en la gramática, pero no de forma general con reglas del tipo:

$$perifrasis \rightarrow verb,\ prep,\ inf. \qquad perifrasis \rightarrow verb,\ inf.$$

que hubiera producido análisis erróneos en algunas ocasiones (por considerar como perifrástica cualquier estructura expresable mediante estas reglas) sino que se ha utilizado la posibilidad de declarar elementos literales en las reglas para poder restringir las formas auxiliares.

Cabe mencionar que, en principio, hay ocho formas posibles para cada perífrasis, tal como se muestra a continuación:
Ej.: *tener* + inf:

tiene que cantar - vaux-simple + inf simple activo
tiene que ser cantado - vaux-simple + inf simple pasivo
ha tenido que cantar - vaux-comp + inf simple activo
ha tenido que ser cantado - vaux-comp + inf simple pasivo
tiene que haber cantado - vaux-simple + inf comp activo
ha tenido que haber cantado - vaux -comp + inf comp activo
tiene que haber sido cantado - vaux-simple + inf comp pasivo
ha tenido que haber sido cantado - vaux-comp + inf comp pasivo

Las formas simples de las perífrasis se declaran del siguiente modo:

verb → *vmsp1s0(deba), infinitiu.*
verb → *vmsp1s0(deba), sps00(de), infinitiu.*
verb → *vmis1p0(tuvimos), cs(que), infinitiu.*
verb → *vmis1s0(estuve), sps00(a_punto_de), infinitiu.*
verb → *vacp3s0(habría), cs00(que), inf.*

Para las compuestas el formalismo es el siguiente:

verb → *vaux, vmp00sm(debido), infinitiu.*
verb → *vaux, vmp00sm(debido), sps00(de), infinitiu.*

El elemento *infinitiu* de estas reglas se reescribe luego como simple o compuesto, activo o pasivo (cf. sección 4.3.8.1), de modo que con relativamente pocas reglas pueden analizarse formas perifrásticas muy complejas[62].

4.3.7.1.1. Perífrasis de infinitivo.

Criterios lingüísticos de reconocimiento.

En el caso concreto de las perífrasis de infinitivo, y siguiendo a Gómez (1999), los criterios de reconocimiento son[63]:

1. Imposibilidad de conmutación del infinitivo por una construcción nominal (nombre, pronombre u oración completiva) o por pronombre interrogativo *qué* (basta con que se dé una de estas posibilidades de sustitución para que haya que considerar que se está ante una construcción no perifrástica[64]);

   (4.31) (a) *Juan {tiene que / puede} presentar el carné.*
   (b) **Juan lo {tiene que / puede}*
   (c) **Juan {tiene que / puede} que se presente el carné*
   (d) **¿Qué {tiene que / puede} Juan?*

2. Capacidad selectiva del infinitivo, que es el único que selecciona el sujeto y los complementos de la perífrasis;

3. Otros criterios secundarios:

   3.1) en las pasivas perifrásticas sólo puede pasivizarse en infinitivo (si es transitivo);

[62] Las reglas para las perífrasis verbales aparecen en el apéndice B sección B.7, aunque dado su elevado número (54 por cada una) sólo se han reproducido, para los tiempos simples, las formas del presente de cada una de ellas. Las formas compuestas aparecen todas.

[63] Los ejemplos son también del autor.

[64] Por ejemplo:

(4.30) 1 *Dejé jugar a los niños.*
(a) *Dejé que los niños jugaran.*
(b) **Lo dejé a los niños.*
(c) **¿Qué dejé a los niños?*

Puesto que en este caso es posible la conmutación por una completiva, hay que concluir que la construcción *dejar + infinitivo* no constituye una perífrasis.

(4.32) (a) *Juan {tiene que / puede} leer la carta.*
(b) **Leer la carta es {tenido que / podido} por Juan.*[65]

3.2) la pasiva refleja afecta a todo el núcleo perifrástico;

(4.33) *Se {tienen que / pueden} celebrar las elecciones.*

3.3) no puede focalizarse el infinitivo de la perífrasis con una estructura enfática de relativo[66];

(4.34) (a) *Juan {tiene que / puede} leer mi libro.*
(b) **Lo que Juan {tiene que / puede} es leer mi libro.*

3.4) los clíticos o bien se anteponen al primer verbo o bien se posponen al infinitivo;

(4.35) *Te lo tengo que decir = Tengo que decírtelo.*

3.5) la marca /se/ de pasiva refleja puede anteponerse o posponerse a la perífrasis;

(4.36) (a) *Se {pueden / deben / tienen que ...} discutir los problemas.*
(b) *{Pueden / deben / tienen que ...} discutirse los problemas.*

3.6) la marca /se/ de impersonalidad ha de anteponerse a toda la perífrasis;

(4.37) (a) *Aquí se {puede / debe / tiene que ...} estar bien.*
(b) **Aquí {puede / debe / tiene que ...} estarse bien.*

3.7) el clítico de un verbo pronominal auxiliar no puede adherirse al auxiliado, mientras que el del auxiliado sí puede adherirse al auxiliar;

(4.38) (a) *Se puso a contar chistes.*
(b) **Puso a contarse chistes.*
(c) *Tuvo que marcharse.*
(d) *Se tuvo que marchar.*

Formas y ejemplos.

En **GramEsp** se han declarado las perífrasis de infinitivo que aparecen en el cuadro 4.2 (con las variaciones de tiempos compuestos y formas pasivas antes mencionadas).

Las construcciones [*haber de* + infinitivo], [*romper a* + infinitivo] y [*terminar de* + infinitivo] se han incluido a pesar de que no se hallaron ejemplos en la parte de LexEsp estudiada. Por otra parte, no se han considerado las estructuras [*querer* + infinitivo] ni [*venir a* + infinitivo], por los motivos que a continuación se comentan.

[65] En cambio es posible: *La carta tiene que / puede ser leída por Juan.*

[66] A pesar de que hay algunas zonas de inseguridad como por ejemplo en *Lo que tú debes es leer mi libro.* Por otra parte, determinadas construcciones de infinitivo no perifrásticas tampoco admiten esta posibilidad: *Hizo llorar a sus amigos ? *Llorar a sus amigos es lo que hizo.*

acabar de + infinitivo	*acertar a* + infinitivo
acostumbrar a + infinitivo	*alcanzar a* + infinitivo
comenzar a + infinitivo	*deber de* + infinitivo
deber + infinitivo	*dejar de* + infinitivo
echar a + infinitivo	*empezar a* + infinitivo
haber de + infinitivo	*ir a* + infinitivo
llegar a + infinitivo	*poder* + infinitivo
romper a + infinitivo	*soler* + infinitivo
tardar en + infinitivo	*tener que* + infinitivo
terminar de + infinitivo	*volver a* + infinitivo

Cuadro 4.2: Lista de perífrasis de infinitivo declaradas en GramEsp

[*Querer* + infinitivo] es una construcción no perifrástica cuando lleva sujeto de persona. Cuando el sujeto es de cosa o 'cero', el comportamiento sintáctico es el de una perífrasis verbal con un significado aspectual de 'estar a punto de' o modal, indicando disposición, e incluso, posibilidad[67].

(4.39) (a) *Dos lágrimas querían asomarse a sus ojos*
(b) *Hoy quiere llover*[68]

De los 22 ejemplos de esta estructura hallados en el corpus, 21 presentaban un sujeto [+humano] por lo que no pueden considerarse perífrasis. He aquí algunos ejemplos:

(4.40) (a) *Yo* **quería conocer** *aquello, hacer un reportaje, y él se ofreció a servirme de guía.*
(b) *Una no se* **quería marchar** *de allí.*
(c) **Quiero decir** *con esto que sé bien que aquello es un infierno.*
(d) *Porque ellos jamás dicen lo que nosotras* **queremos oír**, *y lo que nosotras decimos les abruma.*

En el otro caso, también con sujeto [+humano] tampoco puede hablarse de perífrasis verbal, sino más bien de una fórmula de cortesía:

(4.41) *Lisbeth, que patroneaba entonces el yate, se dirigió a Frans: - - Por_favor, ¿* **quieres determinar** *la posición exacta del barco ? Creo que estamos aproximándonos ya a la costa española.*

El caso de la construcción [*venir a* + infinitivo] es algo distinto. Se han hallado siete ejemplos de aparición:

(4.42) (a) *¡ Pues lo limpié, no como usted, cuando hizo la obra aquella que ni El_Escorial,* **venga a subir** *sacos de yeso por la escalera, que lo dejó todo guarrísimo ! - - ¡ Por_favor, doña Paquita, ¡ olvídese de aquello de una vez ! - - rogó el administrador - -.*

[67] Gómez (1999): pp. 3363-64.
[68] Los ejemplos son también del autor.

(b) **viene a decir** *Nabokov, puede ser un peligro.*
(c) *Por_eso se nos dice que, en el cielo, los santos y las almas buenas se dedican a la contemplación de Dios: lo cual* **viene a ser** *para mí algo así_como contemplar el tiempo remansado, detenido.*
(d) *Y lo más notable quizá es que este giro (equivalente en_principio al cambio del cambio que el Gobierno socialista había introducido en la política exterior de la España democrática)* **se ha venido a producir** *dentro_de un escenario de tensión en las relaciones Este-Oeste, cuyo precedente más próximo se remonta a la crisis suscitada por la instalación de los cohetes soviéticos en Cuba.*
(e) *pero, al propio tiempo,* **viene a revelar** *todo ello qué género de cuestión es el que se dilucida más allá del nivel de transparencia que se deben exigir a las cuentas de todo organismo público.*
(f) *Boruslawa_Smolka, la maestra, recuerda: "Las bandas de ucranianos* **venían a saquear, violar y matar**.
(g) *El "Everzwijn "entró en el puerto sorteando los fondos arenosos de la ría, y* **vino a atracar** *casi enfrente_de la vieja casilla de carabineros.*

En el primer caso (4.42 (a)), parece tratarse de una estructura enfatizadora. Los ejemplos 4.42 (b-e) podrían considerarse perífrasis, especialmente en los tres últimos casos en que el sujeto es [-animado]. Sin embargo, en los ejemplos 4.42 (f-g) no puede hablarse de perífrasis, puesto que aquí la forma de "venir" es la que selecciona el sujeto de las oraciones. Por estas razones no se han tratado como formas perifrásticas.

4.3.7.1.2. Perífrasis de gerundio.

Criterios lingüísticos de reconocimiento.

Según Yllera (1999),

> *Se consideran perifrásticos los complejos verbales que funcionan como una sola unidad verbal. El significado propio de la perífrasis surge de la conjunción del auxiliar y el gerundio y no se reduce a la suma del significado de sus dos componentes*[69]. Además, para que *exista una perífrasis de gerundio es [...] necesario que: a) el gerundio posea carácter verbal y no adverbial o adjetival, b) coincida el sujeto del gerundio con el sujeto del auxiliar y c) no existan complementos que modifiquen exclusivamente al auxiliar*[70].

Por otra parte, es cierto que muchas veces sólo el significado, el contexto o incluso la situación contribuyen claramente a la diferenciación de estructuras superficialmente idénticas. Tal como se indica en Yllera (1999) p. 3393: *Sólo la situación permite interpretar los ejemplos* [Van contando los números *o* Andan mirando a todo el mundo por encima del hombro] *como 'cuentan progresivamente los números', 'miran a todo el mundo por*

[69] Yllera (1999): p. 3393.
[70] Yllera (1999): p. 3393.

encima del hombro' (perífrasis) o como 'caminan y cuentan los números', 'miran a todo el mundo por encima del hombro cuando andan' (no perífrasis).

Los criterios que se establecen para la delimitación de las perífrasis han sido tanto semánticos como sintácticos, y, evidentemente, los resultados varían según se utilicen unos u otros[71]. En el primer caso aparecen como auxiliares los verbos *estar, andar, ir, venir, llevar*, mientras que en el segundo aparecen además *acabar, seguir, continuar, terminar empezar, comenzar*, pero quedan excluidos *quedarse, salir.*

Para poder hablar de perífrasis de gerundio deben darse las siguientes condiciones:

1. interrogación por medio de *qué* sustituyendo el gerundio por la proforma verbal *haciendo*;
2. focalización del auxiliar mediante el infinitivo de su base léxica (retomándolo por *haciendo*), aunque en algunos casos se admite también la focalización del gerundio como tal;
3. pasivización (tanto con *ser* como con *se*) de toda la secuencia verbal;
4. imposibilidad de conmutación del gerundio por un adverbio, un adjetivo o un circunstancial;
5. imposibilidad de escindir las dos formas verbales en dos oraciones con verbo finito;
6. si no hay elementos incrustados entre ambas formas verbales, los clíticos pueden anteponerse o posponerse al conjunto (siempre que la perífrasis aparezca en una forma finita).

Sin embargo, no debe olvidarse una *gradualidad en el comportamiento perifrástico*[72] de las construcciones así como tampoco el papel desempeñado por el contexto o la situación antes mencionado.

Por otra parte, las perífrasis con gerundio pueden ir precedidas de auxiliares modales o temporales de infinitivo, así como del auxiliar frecuentativo *soler*. Sin embargo no aceptan, los auxiliares que indican una fase del proceso (inceptiva, terminativa) o su reiteración, incompatibles con su visión de la acción en curso:

(a) *Deben de estar durmiendo.*
(b) *Van a ir llegando.*
(c) *A las tres empieza suele estar haciéndolo.*
(d) **Empieza a estar viéndolo.*
(e) **Terminó de estar haciéndolo.*

Sobre la perífrasis [*estar* + gerundio] es de notar que responde a la pregunta *¿Durante cuánto tiempo estuvo haciéndolo?* mientras que las estructuras no perifrásticas responden a *¿En cuánto tiempo lo hizo?*

[71] Véase Yllera (1999) pp: 3395-96.
[72] Yllera (1999): p. 3400.

Formas y ejemplos.

Las perífrasis de gerundio que se consideran en la gramática son las que aparecen en el cuadro 4.3.

acabar + gerundio	*andar* + gerundio
comenzar + gerundio	*continuar* + gerundio
empezar + gerundio	*estar* + gerundio
ir + gerundio	*llevar* + gerundio
seguir + gerundio	*terminar* + gerundio
venir + gerundio	

Cuadro 4.3: Lista de perífrasis de gerundio declaradas en GramEsp

4.3.7.1.3. Perífrasis de participio. No hemos considerado ninguna perífrasis de participio, excepción hecha de los tiempos compuestos de la conjugación, que se han comentado en la página 212.

4.3.8. Otros nodos en GramEsp

Comentamos en esta sección el resto de elementos que trata **GramEsp**: las formas no personales del verbo y las conjunciones y relativos. Todos ellos son nodos básicamente unarios (formados por un solo elemento) si exceptuamos las formas verbales compuestas.

4.3.8.1. Formas no personales del verbo

Al igual que en el caso de los verbos finitos, tampoco en el caso de las formas no personales del verbo se han tenido en cuenta los complementos que puedan tener.

Los participios entran en las formas compuestas de la conjugación, tanto activas como pasivas. En la reescritura de los terminales[73] ya se establece la distinción entre las formas masculinas singulares del participio (las formas no marcadas) del resto de formas flexivas. La primera es la que aparece en los tiempos compuestos; las otras son las que aparecen en la voz pasiva[74].

En lo que se refiere a las construcciones absolutas, el participio mantiene su etiqueta verbal, tal como muestra el siguiente ejemplo:

```
S_[ abrumado_vmp00sm
    sn_[ espec-ms_[ el_da0ms0 ]
         grup-nom-ms_[ lector_ncms000 ] ]
    ... ]
(a1)
```

Los gerundios, por su morfología, entran en distribución con los adverbios, pero también reciben complementos propiamente verbales. En la siguiente oración se puede observar un gerundio seguido de complemento directo:

[73] Cf. apéndice B, sección B.2.10.
[74] Cf. apéndice B, sección B.8.1.

(4.43) *El padre Martín, por el contrario, parecía un pachá recorriendo sus dominios*[75].

Además, pueden intervenir en construcciones absolutas (4.44 (a)) o pueden haberse fijado como adjetivos (4.44 (b)):

(4.44) (a) *Pero, andando los años, aquel río de oro asentado en las márgenes del Tejo fue aclarando su caudal.*[76]
(b) *Un carrito en la cuesta con cintas de colores colgando del techo.*[77]

Todos los usos anteriormente comentados del gerundio quedan agrupados bajo la misma etiqueta *gerundi*, tal como puede observase en el apéndice B, sección B.8.1.

El tratamiento de los infinitivos sigue un camino paralelo al de los gerundios. Todos sus usos se agrupan bajo una única etiqueta (*infinitiu*).

4.3.8.2. Elementos de enlace

Incluimos aquí las conjunciones y los relativos. Por lo general, tanto las conjunciones subordinantes como las coordinantes forman nodos independientes. No se han agrupado con ningún elemento a su derecha dado que es imposible saber, con criterios puramente formales, dónde finaliza el elemento afectado por estas palabras. La excepción la constituyen las conjunciones coordinantes en la coordinación léxica de nombres y adjetivos.

Algunos relativos[78] entran en agrupaciones con nombres: se trata de las formas de *cuanto* y *cuyo* que ocupan la posición de determinante del nombre:

relatiu → *prel-fs, grup-nom-fs.* %cuantos asistentes
relatiu → *cuyo-ms, grup-nom-ms.* %cuyo amigo

Las formas de *cual* aparecen siempre con artículo y a veces pueden ir precedidas de preposición:

relatiu → *j-ms, cual-s.* %el cual
relatiu → *prep, j-ms, cual-s.*

En todos estos casos se ha tenido en cuenta la concordancia interna al sintagma nominal que forman estos elementos. Si no se han etiquetado estas agrupaciones como tales, como sintagmas nominales, ha sido porque se ha considerado más relevante para el análisis sintáctico su etiquetación como elementos relacionantes que marcan el inicio de una subordinada

Los casos de <artículo + *que*> son ambiguos: si el relativo tiene antecedente el artículo forma un constituyente con el relativo; pero si no lo hay, el artículo actúa como sustantivador de la relativa. Dada esta ambigüedad, hemos optado por no introducir ninguna regla para este tipo de secuencias.

[75] Ejemplo tomado de Alcina y Blecua (1989): p. 751.
[76] Ejemplo tomado de Alcina y Blecua (1989): p. 753.
[77] Ejemplo tomado de Alcina y Blecua (1989): p. 751.
[78] Las reglas en que intervienen los relativos aparecen en el apéndice B, sección B.8.2.

Por lo demás, los relativos no entran en ninguna otra combinación de elementos. Al igual que ocurría en el caso de las conjunciones, es imposible determinar dónde acaban las relativas.

4.4. Conclusiones

En este capítulo se ha presentado una gramática para el análisis superficial del español (**GramEsp**), de amplia cobertura aunque trata los constituyentes a un nivel bajo de análisis. La ventaja de este tipo de proceso es que la tasa de error es mínima y se consiguen agrupaciones sintácticas que no sólo van a facilitar posteriores niveles de análisis, sino que por sí mismas permiten el desarrollo de sistemas de interés aplicativo.

Por otro lado, con esta gramática se resuelve, además, el tratamiento de algunas unidades que no se han podido tratar en el módulo de análisis morfológico. Se trata de unidades que se sitúan entre la morfología y la sintaxis, como las perífrasis verbales y los tiempos compuestos.

Finalmente, es de destacar que en esta fase de análisis finaliza la resolución de la desambiguación morfológica de los atributos de género y número.

Otra aportación que nos parece significativa es la de una primera aproximación a la definición de *chunk* para el español, dado que la propuesta por Abney no es universal.

Capítulo 5

Anotación sintáctica de corpus

Los diferentes niveles de análisis tratados hasta el momento corresponden a procesos totalmente automáticos que proporcionan resultados de alta calidad con una tasa baja de error. A partir de este nivel, la progresión del análisis lingüístico, cuando concierne al tratamiento de textos sin restricciones, precisa del desarrollo de recursos básicos como son los *treebanks*, que permiten captar información sobre el comportamiento sintáctico de las lenguas. Estos recursos son, por tanto, previos al desarrollo de una gramática que trate la lengua en profundidad y su desarrollo debe seguir una metodología estricta y basarse lo más posible en criterios lingüísticos. A partir de un banco de datos de árboles sintácticos se pueden inferir gramáticas o bien se puede observar el comportamiento sintáctico de la lengua sobre la base de datos reales. Se trata, por tanto, de un requisito imprescindible para abordar esta tarea.

En este apartado describimos la metodología que se está siguiendo para el desarrollo de **Cast3LB**, un banco de árboles sintácticos del castellano que constituye uno de los componentes del proyecto **3LB**[1].

La figura 5.1 sitúa este trabajo en el marco de los procesos de análisis del lenguaje de CLiC-TALP.

[1] Véase capítulo 1.

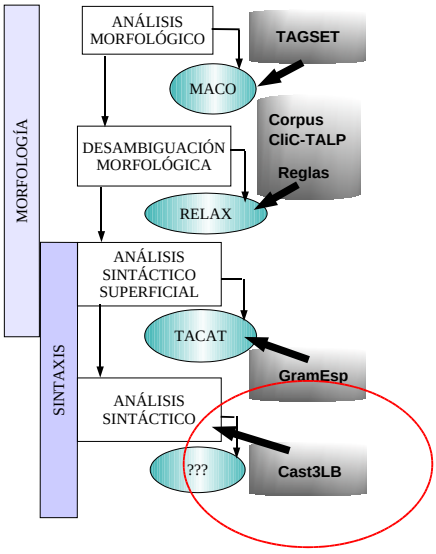

Figura 5.1: Procesos de análisis (4): anotación sintáctica de corpus

5.1. Introducción: estado de la cuestión

De acuerdo con Leech, Barnett, y Kahrel (1996) la anotación sintáctica de corpus es *the practice of adding syntactic information to a corpus by incorporating into the text indicators of syntactic structure.*

Suele utilizarse el término *esquema de anotación* para hablar de las especificaciones de las prácticas de anotación utilizadas para un corpus particular, para distinguirlo del método utilizado para aplicar los esquemas de anotación al corpus. Al igual que ocurría con la anotación morfológica, la anotación sintáctica puede realizarse de modo manual (Sampson, 1995), automático (Karlsson et al., 1995) o semiautomático (Böhmova y Hajicova, 1999)[2]. Las modalidades de anotación semiautomática pueden ir desde la corrección, por parte de expertos lingüistas, de los errores producidos por un analizador automático (Marcus, Santorini, y Marcinkiewicz, 1993), hasta la anotación interactiva (Brants y Plaehn, 2000) o la construcción manual de los árboles a partir de un análisis superficial previo, como en el caso del **Cast3lb** (Civit y Martí, 2002).

En la anotación sintáctica hay que determinar no sólo las etiquetas que se utilizarán, sino también los segmentos a que éstas se van a aplicar[3]. Por lo general, las oraciones son

[2]Recuérdese también a este respecto el cuadro 1.1 mencionado en el capítulo 1.

[3]En Sampson (1995) aparece la siguiente anécdota: *In 1991, at the annual conference of the Association for Computational Linguistics at Berkeley, California, a workshop was held in which NLP researchers from nine institutions were given a range of authentic English sentences and were asked to specify, by annotating them with labelled bracketings, what their respective research groups would regard as the ideal analyses for these sentences [...] The nine sets of annotations were then compared. One cannot expect separate research groups' grammatical category labels to coincide, because choice of linguistic terminology is open-ended and the same technical term can be abbreviated in different ways, but the brackets themselves are a different matter: a particular sequence of words either is or is not bracketed together as a grammatical constituent.*

las unidades máximas de análisis y las palabras las mínimas; y en principio se asume que tanto la palabra ortográfica como la oración ortográfica se corresponden con la palabra y la oración sintáctica.

Los formatos de presentación de los corpus anotados son generalmente dos: horizontal, como en el caso del corpus *Le Monde*[4], tal como aparece en el primero de los siguientes ejemplos, y vertical, como en el segundo ejemplo tomado del corpus *SUSANNE*[5]:

```
<NP> deux:D <COORD> ou:cc trois:D </COORD> enfants:nc </NP>
```

```
YB       <minbrk>        -          [Oh.Oh]
AT       The             the        [O[S[Nns:s.
NP1s     Fulton          Fulton     [Nns.
NNL1cb   County          county     .Nns]
JJ       Grand           grand      .
NN1c     Jury            jury       .Nns:s]
VVDv     said            say        [Vd.Vd]
NPD1     Friday          Friday     [Nns:t.Nns:t]
AT1      an              an         [Fn:o[Ns:s.
NN1n     investigation   investigation .
IO       of              of         [Po.
NP1t     Atlanta         Atlanta    [Ns[G[Nns.Nns]
GG       +<apos>s        -          .G]
JJ       recent          recent     .
JJ       primary         primary    .
NN1n     election        election.Ns]Po]Ns:s]
VVDv     produced        produce    [Vd.Vd]
YIL      <ldquo>         -          .
ATn      +no             no         [Ns:o.
NN1u     evidence        evidence.
YIR      +<rdquo>        -          .
CST      that            that       [Fn.
DDy      any             any        [Np:s.
NN2      irregularities  irregularity    .Np:s]
VVDv     took            take       [Vd.Vd]
NNL1c    place           place      [Ns:o.Ns:o]Fn]Ns:o]Fn:o]S]
YF       +.              -          .O]
```

En Leech, Barnett, y Kahrel (1996) se recomienda distinguir de un modo muy claro la forma de la función: *it is a common practice to annotate firstly at a* **form** *level, and at a functional level optionally afterwards. This can be advantageous since the initial level of annotation is one which provides basic descriptive information about the language [...], the further annotation at a functional level would provide a richer, deeper analysis.*

When one considers how long and fully the English language has been worked on, it might seem reasonable to expect different researchers' bracketings of sentences usually to coincide.

For three of the workshop examples, here are the total sets of bracketings that were identified by every workshop participant:

The famed Yankee Clipper, now retired, has been assisting [as [a batting coach]]
One of those capital-gains ventures, in fact, has saddled him [with [Gore Court]]
He said this constitued a [very serious] misuse [of the [Criminal court] processes].

[4]Abeillé, Clément, y Kinyon (2000).

[5]Sampson (1995).

Los niveles de anotación que se proponen desde EAGLES[6] son los siguientes:

1. parentización de segmentos: delimitación de constituyentes sintácticos (*nivel a*);
2. etiquetación de las categorías de los segmentos, con etiquetas como *sintagma nominal, sintagma verbal*, etc. (*nivel b*);
3. marcaje de las relaciones de dependencia entre los núcleos y los elementos dependientes de estos núcleos (*nivel c*);
4. marcaje de la función sintáctica del tipo *Sujeto, Objeto*, etc. (*nivel d*);
5. subclasificación de los segmentos sintácticos, con valores como *singular, pasado*, etc.(*nivel e*) ;
6. relaciones lógicas: correferencia, referencia cruzada, elipsis, estructuras de control, huellas o discontinuidad sintáctica (*nivel f*);
7. información sobre el rango de una unidad sintáctica, mediante indentación u otras marcas (*nivel g*);
8. información sobre fenómenos de disfluencia en lenguaje hablado (*nivel h*).

La mayoría de estos niveles de anotación presentan una relación jerárquica, porque unos se implican a otros.

El cuadro 5.1, tomado de estos autores[7], relaciona algunos de los corpus existentes con los distintos niveles de anotación:

	Niveles de anotación							
	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**
Lancaster	+	+	-	-	-	-	+	-
SUSANNE	+	+	+	+	+	+	+	+
PTB	+	+	+	+	-	+	+	-
IBM Paris	+	+	-	(+)	-	-	-	-

Cuadro 5.1: Corpus y niveles de anotación (1)

Estos criterios pueden utilizarse para caracterizar otros *treebanks*, como aparece en el cuadro 5.2[8]:

Las principales formas de representación de la estructura sintáctica son los *constituyentes* y las *dependencias*.

Los constituyentes representan la estructura sintagmática de las oraciones. Las dependencias representan las relaciones de núcleo–modificador entre los elementos terminales de las oraciones.

[6] Leech, Barnett, y Kahrel (1996).
[7] Leech, Barnett, y Kahrel (1996).
[8] Los paréntesis indican que la anotación correspondiente a este nivel no se hace de forma explícita pero puede inferirse del resultado de la anotación. Las referencias para los *Treebanks* que aparecen aquí mencionados son, por orden, Brants y Plaehn (2000), Hajic (1998), Abeillé, Clément, y Kinyon (2000), Bosco et al. (2000), Moreno y López (1999) y Montemagni et al. (2003).

	Niveles de anotación							
	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**
NEGRA	+	+	+	+	-	-	+	-
PDT	-	+	+	(+)	-	-	+	-
Le Monde	+	+	(+)	+	-	-	+	-
TUT	-	(+)	+	+	-	+	+	-
UAM	+	+	+	+	(+)	+	+	-
ISST	+	+	(+)	+	-	+	+	-

Cuadro 5.2: Corpus y niveles de anotación (2)

En las representaciones de constituyentes aparecen dos clases de nodos: los terminales (las hojas de los árboles, las palabras de las oraciones), y los nodos no-terminales, los constituyentes. Por su parte, en las dependencias todos los nodos de la representación arbórea son terminales, dado que las relaciones núcleo–modificador se establecen directamente entre las palabras.

Otra diferencia es que, en la representación de constituyentes, la relación núcleo--modificador no aparece de forma explícita, aunque no resulta muy difícil introducirla (bastaría con marcar cuál es el núcleo de cada sintagma). Sin embargo, establecer los núcleos no es una tarea trivial. Hay construcciones en las que se hace difícil, como por ejemplo, las estructuras coordinadas. En el corpus francés *Le Monde* (Abeillé, Toussenel, y Chéradame, 2002) se considera que el núcleo de las estructuras coordinadas es el elemento a la izquierda:

```
<NP> deux:D <COORD> ou:cc trois:D </COORD> enfants:nc </NP>
```

mientras que en el PennTreeBank (Bies et al., 1995), por ejemplo, los elementos coordinados se tratan como hermanas y, por tanto, no se establece ninguna relación jerárquica entre ellos:

```
S [ S [Maty likes Bach] and S [Susan Beethoven] ]
```

Por último, cabe señalar que en la representación de dependencias, la dirección de la ramificación indica el orden de los elementos en la oración: la ramificación a la izquierda indica precedencia y la ramificación a la derecha, la inversa.

El cuadro 5.3 presenta una caracterización de algunos de los *treebanks* existentes atendiendo a las siguientes características: en primer lugar a la lengua que representan; en segundo, al tamaño en términos de palabras (p) u oraciones (o) que contienen; en tercer lugar, la proceso de anotación que se ha seguido (**M** significa *manual*, **SA**, *semiautomático* y **A** *automático*); seguidamente se indica si se ha aplicado, para su construcción, alguna teoría lingüística concreta; y, por último se señala si se anotan constituyentes (**C**) o dependencias (**D**).

Treebank	**lengua**	**tamaño**	**proceso**	**teoría**	**C/D**
Susanne	inglés	120000p	M	N	C
PennTB	inglés	2Mp	SA	*X-bar*	C
ICE-GB	inglés	1Mp			
ISST	italiano	300000p	SA	N	C/D
PDT	checo	450000p	SA	N	D
UAM	español	1500o	SA	*X-bar*	C
Le Monde	francés	1Mp	SA	N	C
NEGRA	alemán	20000o	SA	N	C
TIGER	alemán		SA	N	C
Polaco	polaco			*HPSG*	C
Bultreebank	búlgaro		SA	*HPSG*	C
Turco	turco	10000o	M	N	D
TUT	italiano	1000o	SA	N	D
Hebreo	hebreo	500o	M	N	C
Sueco	sueco	1Mp	SA	N	
Floresta	portugués	1Mp	SA	N	D/C
Cast3LB	español	100000p	SA	N	C
Cat3LB	catalán	100000p	SA	N	C
Eus3LB	euskera	50000p	SA	*X-bar*	D

Cuadro 5.3: Características de los principales Treebanks existentes

5.2. Anotación sintáctica de Cast3LB: generalidades

En esta sección presentamos los aspectos generales del esquema de anotación de constituyentes adoptado para el corpus **Cast3LB**. Este corpus proviene en un 75 % del corpus CLiC-TALP y en un 25 % de un corpus de noticias cedido por la agencia EFE para investigación.

En la anotación sintáctica de **Cast3LB** se anotan los niveles *a, b, c, d, g* y parcialmente *f* de Leech[9] que se corresponden, respectivamente, con la parentización de constituyentes, su etiquetación, el marcaje de las relaciones de dependencia (que se hace de forma implícita), el marcaje de la función sintáctica, el marcaje del rango de las unidades mediante indentación y algunas relaciones lógicas como la elipsis de sujeto.

La anotación se lleva a cabo de modo manual, partiendo del análisis de los textos proporcionado por TACAT y GramEsp[10] mediante la interfaz AGTK (Cotton y Bird, 2000).

La figura 5.2 muestra el punto de partida del análisis de constituyentes, mientras que la figura 5.3 muestra el resultado final para el análisis de la oración *Hay quien intenta cumplir y llamar la atención regalando cosas completamente absurdas* siguiendo los criterios de **Cast3LB**.

[9] Leech, Barnett, y Kahrel (1996). Aspectos comentados en la página 226.
[10] Cf. capítulo 4.

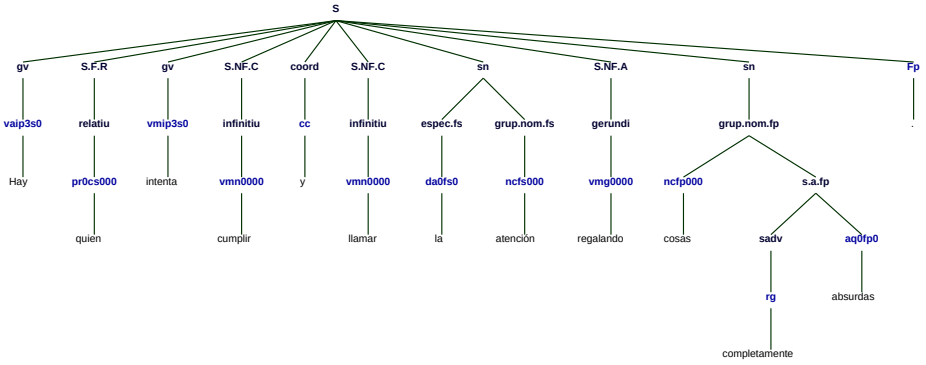

Figura 5.2: Input de la anotación manual

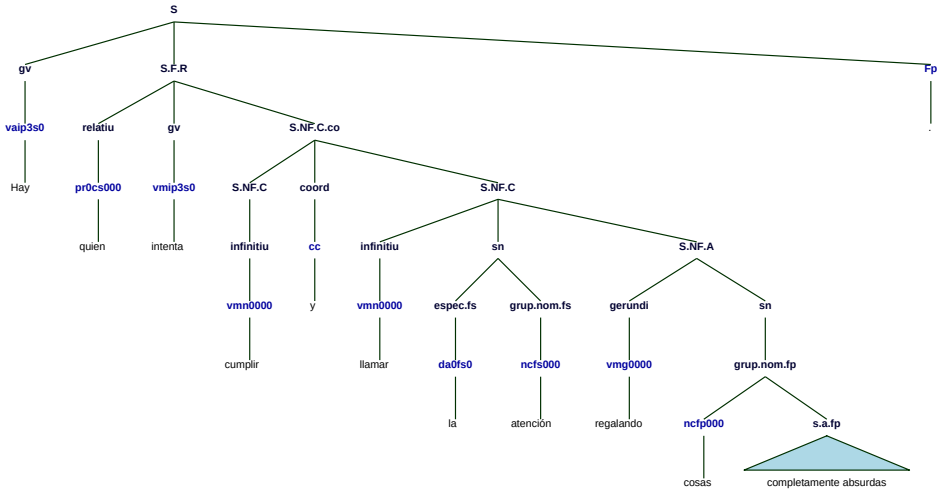

Figura 5.3: Output de la anotación manual

Las utilidades que permite esta interfaz son las siguientes: borrar y añadir nodos al árbol; moverlos (aunque sin permitir el cruce de ramas); añadir huellas (para los elementos elípticos); cambiar etiquetas (a nivel de categoría morfológica pero también sintáctica); unir o separar oraciones; unir o separar palabras y añadir hojas al árbol.

La figura 5.4 muestra parcialmente esta interfaz.

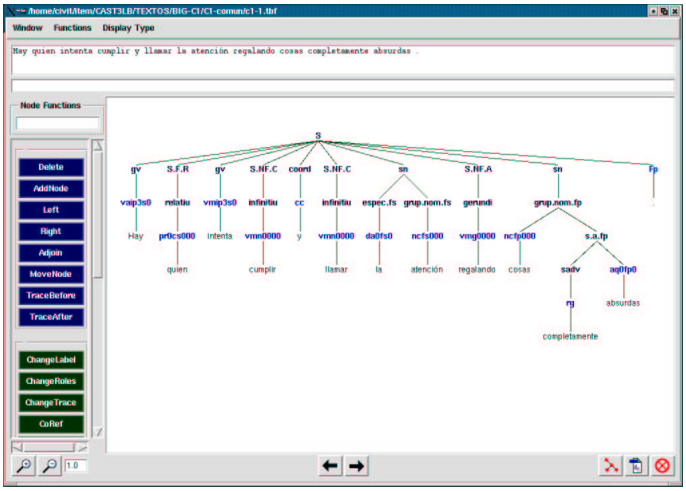

Figura 5.4: Interfaz AGTK

A continuación presentamos el marco de anotación del corpus **Cast3LB**, centrándonos en los aspectos generales, como el hecho de seguir o no una teoría lingüística, el formalismo adoptado, el tratamiento del orden de los elementos en la oración y de la elipsis. En la sección 5.3 comentamos los aspectos particulares de la anotación sintáctica de constituyentes.

5.2.1. Esquema de anotación

Una de las primeras decisiones que hay que tomar a la hora de iniciar la anotación sintáctica de corpus se refiere a la utilización o no de un marco teórico determinado que la guíe. La mayoría de treebanks existentes optan por trabajar en un marco teórico neutro. Este hecho debe interpretarse como la voluntad de realizar una anotación básica sobre la que diferentes personas puedan estar de acuerdo, que pueda ser útil para distintas aplicaciones y para diversos tipos de estudios teóricos sobre el lenguaje. El motivo fundamental de esta decisión es la *reutilización* que puede hacerse del corpus. Si el esquema de anotación es lo suficientemente general o neutro respecto de las teorías y formalismos lingüísticos, el corpus puede ser utilizado por lingüistas, informáticos, psicólogos, etc., tanto para llevar a cabo estudios sobre la lengua como para realizar aplicaciones de procesamiento automático del lenguaje, mientras que si el esquema que se sigue está muy determinado por la teoría, su utilización se restringe significativamente[11].

Algunos treebanks se han anotado siguiendo un marco teórico lingüístico determinado: la teoría de la *X-barra* en el PennTreeBank ((Marcus, Santorini, y Marcinkiewicz, 1993) y (Marcus et al., 1994)), en el UAMSpanishTreebank ((Moreno y López, 1999), (Moreno et al., 2000), (Moreno et al., 2003)) y en *Eus3LB* (Aduriz et al., 2002) y, la HPSG, en los treebanks del Búlgaro (Simov et al., 2002) o del polaco (Marciniak et al.,

[11] Abeillé, Toussenel, y Chéradame (2002).

2003). En el primer caso, la anotación se lleva a cabo siguiendo la teoría de la *X barra*, aunque sus principios no se aplican de modo íntegro, especialmente por la dificultad que implica, en la práctica, la distinción entre argumentos y adjuntos verbales tal como se manifiesta en (Marcus et al., 1994) y en (Taylor, Marcus, y Santorini, 2003). Los argumentos que se dan en Marciniak et al. (2003) para la elección de la teoría HPSG son que este procedimiento facilita la evaluación de una gramática HPSG; que proporciona una manera uniforme de representar diferentes tipos de información lingüística; y, finalmente, que este formalismo es ampliamente utilizado en lingüística computacional. Por su parte, en Simov et al. (2002) se sostiene que este formalismo permite representar simultáneamente los constituyentes y las dependencias; que esta teoría permite una descripción consistente de los hechos lingüísticos y, por último, que permite la traducción a otros formalismos.

En el caso de **Cast3LB** hemos optado por no seguir ningún marco teórico concreto, sino por proporcionar una anotación lo suficientemente neutra para que el resultado pueda ser utilizado por una amplia variedad de investigadores de todos los campos, tanto de la lingüística como de la informática, entre otros. Además creemos que, dado que es un primer paso en la anotación sintáctica, una base neutra respecto de la teoría nos permitirá cualquier tipo de posterior enriquecimiento, sea éste teóricamente determinado o no.

5.2.2. Anotación de constituyentes

Tal como se ha descrito en la sección 5.1, dos son las formas básicas de expresar las relaciones sintácticas entre los elementos de las oraciones: *constituyentes* y *dependencias*. La mayoría de treebanks existentes presentan una anotación de constituyentes. En algunos artículos se plantea el hecho de que las dependencias son mejores para las lenguas de orden libre de palabras (Boguslavsky et al. (2002), Brants, Skut, y Uszkoreit (2003), Oflazer et al. (2003)). Tal sistema de anotación es seguido para la anotación del *Prague dependency treebank*[12], de un treebank para el ruso (Boguslavsky et al., 2002) y de otro para el euskera (Aduriz et al., 2002). Los argumentos utilizados por los autores para justificar la adopción de este tipo de anotación van desde la tradición gramatical (en el caso del checo) hasta los condicionantes lingüísticos o las características de los procesos automáticos de análisis en el caso del euskera.

En el marco de **Cast3LB** hemos optado por la anotación en constituyentes por motivos tanto lingüísticos como técnicos. Por una parte, el español es una lengua en que los constituyentes presentan un orden bastante libre, aunque en el interior de cada constituyente la posición de las palabras está muy determinada. Desde un punto de vista técnico, dada la cadena de procesos de tratamiento automático, lo más acertado parecía utilizar el análisis parcial proporcionado por Tacat y GramEsp.

El marcaje que se hace de los constituyentes no es la simple parentización, sino que también se marca su estructura interna.

[12] Hajic (1998); Böhmova, Panevová, y Sgall (1999); Bemova et al. (1999); Böhmova y Hajicova (1999).

5.2.3. Mantenimiento del orden superficial de los elementos en la oración

Uno de los problemas de la anotación de corpus a nivel sintáctico es el de mantener el orden de palabras tal como aparecen en el texto o bien reubicar los elementos para recuperar el orden natural o no marcado. Dado que los movimientos de los constituyentes en las oraciones tienen efectos sobre el significado de las mismas, hemos optado por mantener el orden original de los elementos de la oración. Esta decisión ha influido también en el hecho (que se comenta en la sección 5.3.1) de no considerar como constituyente de las oraciones el sintagma verbal. En el caso del PennTreeBank, por ejemplo, también se mantiene el orden superficial de los elementos, pero se marcan las huellas, como puede apreciarse en la siguiente frase, tomada de Bies et al. (1995):

```
(SBARQ (WHNP-2 What
       (SQ did
           (NP-SUBJ Casey)
           (VP throw
                (NP *T*-2)))
       ?)
```

Esta decisión conlleva un problema: el tratamiento de los elementos discontinuos y de los *movimientos de Q* en las relativas y las interrogativas como en el caso de *qué te gustaría ser* (a21). En **Cast3LB** la resolución de estos problemas se producirá en la fase de asignación de funciones sintácticas, donde los complementos de un elemento incrustado recibirán una etiqueta especial.

5.2.4. Tratamiento de los elementos elípticos

La elipsis puede definirse según Brucart (1999) como:

> *un mecanismo de infraespecificación léxica mediante el cual se evita la realización fónica de alguno de los constituyentes necesarios para interpretar adecuadamente el enunciado. Tal omisión es posible gracias a que el contenido de la unidad elíptica es directamente accesible al oyente a través del contexto discursivo o situacional*[13].

El problema de tratar con los elementos elípticos es el alcance de este fenómeno. El español es una lengua *pro-drop*, por lo que el sujeto de las oraciones finitas suele estar elíptico. Pero además, casi cualquier elemento puede elidirse. Las frases siguientes muestran diversos casos de elipsis:

(5.1) (a) *Medardo_Fraile juega a un cinismo fácil y divertido* (a1).
(b) *No quiero decir que lo sea, cínico o divertido* (a1).
(c) *'El_rey_y_el_país_con_granos' es un puro esperpento, pero no porque el héroe se haya enfrentado a un espejo deformante, sino porque el escritor va haciendo guiar un berilo clarísimo sobre una página de la historia* (a1).
(d) *Zarrabeitia puso la rebeldía, y Delgado la gallardía* (d2).
(e) *Habría alcanzado a Camargo y obtenido un botín de...* (d2)

[13] Brucart (1999): p. 2789.

(f) *¿Por qué?* (t5)
(g) *Un tonto inmortal* (a26)

En el Treebank construido en la Universidad Autónoma de Madrid se trata la elipsis de sujeto y la que se produce en estructuras coordinadas; en el PennTreeBank se trata la elipsis de modo muy exhaustivo: se marcan las huellas de los movimientos de alfa, los PRO, los complementadores nulos y otros fenómenos como los expletivos, las estructuras de control etc.[14]; en la anotación del *Corpus Le Monde* no se trata ningún caso de este fenómeno[15].

En la fase actual de la anotación sintáctica de **Cast3LB** el único elemento elíptico que se incorporará a las oraciones es el sujeto, pero sólo para las oraciones finitas o con verbo conjugado[16] dado que es especialmente necesario para el marcaje de la anáfora y la correferencia. La elipsis verbal se marca en los nodos oracionales[17]. El resto de elementos elípticos no se tratará.

5.2.5. Resolución de la ambigüedad en la incrustación

Uno de los mayores problemas del análisis sintáctico es la resolución de la ambigüedad en el establecimiento de los constituyentes oracionales. Fundamentalmente, este fenómeno se presenta a nivel de sintagma preposicional y de coordinación (pero también en las relativas y en general en los complementos del nombre).

Un ejemplo de la ambigüedad que presentan los sintagmas preposicionales lo ofrece la siguiente oración: *La Iglesia habla del problema del Mal en el Mundo* (c2). Las estructuras que puede presentar esta oración son:

```
(i) [ La Iglesia habla [ del problema [ del Mal ] ] [ en el Mundo ]]
(ii)[ La Iglesia habla [ del problema [ del Mal [ en el Mundo ]]]]
```

Una u otra estructura dependerá de la interpretación semántica que se dé a los sintagmas preposicionales.

Es cierto que muchas veces, el conocimiento que tenemos como hablantes nos permite asignar la estructura correcta a la oración; es cierto que en el habla las estructuras no suelen ser ambiguas porque bien el contexto, bien nuestro conocimiento del mundo, nos guían en la interpretación. Pero cuando estas estructuras carecen de esta información complementaria, cosa que puede ocurrir cuando se anota un corpus, porque el contexto puede verse muy limitado, entonces hay que establecer, de modo arbitrario, un criterio por defecto, para que siempre se tomen las mismas decisiones.

En el caso del corpus **Cast3LB**, la decisión ha sido la de incrustar los nodos en el nivel más alto posible; esto siempre que la información contextual no sea suficiente para decidirse por una u otra opciones.

[14] Para el tratamiento detallado de todos estos fenómenos puede consultarse Bies et al. (1995), sección 4, pp.59–100.
[15] Recuérdese que ni el inglés ni el francés son lenguas *prodrop*.
[16] Sin embargo, no se descarta incorporar en una fase posterior de anotación otros elementos elípticos.
[17] Véase a este respecto la sección 5.3.3.

5.3. Anotación sintáctica de Cast3LB: particularidades

En esta sección comentamos cuestiones más concretas referidas al establecimiento y a la anotación de los constituyentes[18].

El punto de partida de la anotación de los constituyentes es el análisis sintáctico proporcionado por TACAT y GramEsp (cf. capítulo 4). El resultado de este análisis son constituyentes del tipo *chunks*, de modo que la tarea de anotación manual del corpus consiste en pasar de los chunks a los sintagmas y construir los sintagmas y las cláusulas a partir de los *chunks*. Por lo general, la tarea consiste en ampliar el margen derecho de las estructuras establecidas de modo automático, aunque en ocasiones también hay que crear estructuras nuevas (caso de la coordinación y de algunas subordinadas).

5.3.1. La estructura oracional

En la sección 5.2.3 ya se ha mencionado el hecho de que se mantiene el orden superficial de los elementos de las oraciones. Esta decisión tiene consecuencias sobre el establecimiento de algunas de las unidades que se van a considerar constituyentes de la oración, más en concreto, sobre si se considera o no que toda oración está formada por dos constituyentes: sujeto y predicado, siendo el primero un sintagma nominal y el segundo un sintagma verbal.

En el PennTreeBank se marca la estructura *Predicate* que se corresponde o bien con el sintagma verbal (VP) más a la derecha o con el constituyente etiquetado PRD, que corresponde a sintagmas verbales copulativos o cláusulas reducidas. Se considera que hay dos tipos de argumentos en el predicado: los externos (el sujeto gramatical) y los internos (objeto directo e indirecto). En el esquema de representación hay, por lo general, dos nodos hijas (además de los signos de puntuación) del nodo oración: NP-SUBJ (sujeto) y VP (predicado) que incluye el resto de elementos. Sin embargo, en ocasiones, otros elementos aparecen también como hijas de S, especialmente los premodificadores verbales:

```
S [ [Sandy] [often] [throws curves]]
```

En el caso del Corpus *Le Monde* no se considera la existencia de un sintagma verbal dado que *en français, la séquence postverbale inclut aussi bien des compléments que des circonstants ou des sujets inversés. Donc, soit VP englobe tout et il est inutile, soit il n'inclut que les compléments et il est discontinu*[19], de modo que la representación es mucho más plana que en el caso anterior. Sin embargo, el nodo VN (*noyau verbal*) incluye los pronombres sujeto y objeto (CL), el adverbio negativo *ne* y los auxiliares *avoir* y *être* seguidos de participio pasado:

```
Jean <VN> n':Adv en:CL veut </VN> plus.
<VN> Regarde moi:CL </VN>.
<VN> Elle:CL nous:CL verra </VN> bien.
<VN> Voici </VN> <NP> Pierre </NP>.
```

En nuestro caso hemos optado por no considerar el constituyente sintagma verbal como nodo que incluye el verbo y sus complementos. Dado el orden libre de constituyentes en

[18] Todos los detalles de la anotación sintáctica así como numerosos ejemplos de análisis pueden consultarse en Civit (2002).

[19] Abeillé, Toussenel, y Chéradame (2002).

español, para considerar SV o bien habría que alterar el orden superficial de las palabras o bien habría que considerar la discontinuidad de elementos. De modo que en **Cast3LB** los elementos principales de la oración, los que desempeñarán las funciones de sujeto, complemento directo, indirecto, ... son todos ellos hijas de S, y el nodo GV (grupo verbal) contiene únicamente las formas verbales. Los clíticos y demás partículas preverbales forman nodos independientes de GV.

5.3.2. La coordinación

Como se ha comentado anteriormente (página 227), la coordinación es un fenómeno que plantea problemas de representación. Las soluciones generalmente adoptadas son dos: o bien se toma de modo arbitrario uno de los elementos como núcleo (caso del corpus *Le Monde*) o bien se establece un nuevo nodo que incluye a los elementos coordinados. En este último caso, el nodo adicional puede incorporar una marca en la etiqueta (caso del *Prague Dependency treebank*) o no (caso del *PennTreeBank*). Por otra parte, la representación o la solución adoptadas pueden variar según el tipo de elementos coordinados.

En el *PennTreeBank*, por ejemplo, el principio general que se sigue para la coordinación es coordinar los elementos *at the lowest level possible. However, the addition of modifiers forces a higher level of coordination*[20]. Además, se establecen distintas soluciones en función de los elementos coordinados:

1. coordinación léxica: anotación totalmente plana:

```
(S (NP-SUBJ Baking and eating)
   (VP are
      (ADJP-PRD fun)))
```

2. coordinación de sintagmas del mismo tipo: el nodo madre recibe la misma etiqueta que los nodos hijas:

```
(NP (NP three cookies)
    and
    (NP a book))

(VP (VP fishing)
    and
    (VP tying
       (NP flies)))
```

3. coordinación de sintagmas de diferente tipo: el nodo madre recibe una etiqueta especial: UCP (*Unlike Coordinated Phrase*):

```
(S (NP-SBJ U.S. interest)
   (VP may
      (VP be
         (UCP-PRD (ADJP-PRD big)
                  and
                  (VP growing)))))
```

[20]Bies et al. (1995).

4. coordinación a nivel de oración: sólo se lleva a cabo si las oraciones tienen su propio sujeto.

Un fenómeno relacionado con la coordinación es la elipsis. En el *PennTreeBank* se marcan algunos de los elementos elididos con huellas o se identifican los elementos que realizan la misma función, como en el caso siguiente:

```
(S (S (NP-SBJ-1 Mary
      (VP likes
         (S (NP-SBJ *-1)
            (VP to
               (VP swim
                  (PP-TMP-2 on
                     (NP Tuesdays)))))))
   and
   (S (NP-SBJ=1 Bill)
      (PP-TMP=2 on
         (NP Wednesdays))))
```

Además, establecen una diferencia según si los elementos coordinantes están formados por una palabra, en cuyo caso no reciben ninguna etiqueta (como en el ejemplo anterior), o por más de una palabra, en cuyo caso quedan agrupados bajo el nodo CONJP con una estructura interna totalmente plana:

```
(S (NP-SBJ She)
   (VP valued
      (NP wisdom
      (CONJP as well as)
      knowledge)).)
```

Las conjunciones discontinuas se etiquetan también como CONJP:

```
(S (NP-SBJ Her actions)
   (VP were
      (ADJP (CONJP not only)
            (ADJP compassionate)
            (CONJP but also)
            (ADJP inspiring))))
```

En el caso de **Cast3LB** seguimos el principio básico de anotar al nivel más bajo posible, y lo aplicamos a todos los constituyentes. Los nodos que pueden coordinarse son los correspondientes a todas las estructuras oracionales, todos los tipos de sintagmas y los especificadores. En el ejemplo siguiente puede verse un caso de coordinación a nivel de grupo nominal (grup.nom.fp.co)[21]. Sin embargo, si hay complementos y determinantes que afectan de modo distinto a los nombres, entonces la coordinación se realiza a nivel de sintagma nominal. Estos principios se aplican a cualquier estructura coordinada[22].

[21] Una alternativa hubiera sido realizar la coordinación a nivel de sintagma nominal (sn) pero es un nivel más alto.

[22] Utilizamos los paréntesis cuadrados para mostrar análisis simplificados, mientras que utilizaremos los redondos para mostrar análisis reales, que incluyen la categoría gramatical de las palabras.

```
    (sps00 por))
                (sn
                  (espec.fp
                    (da0fp0 las))
                  (grup.nom.fp.co
                    (grup.nom.fp
                      (ncfp000 subidas))
                    (coord
                      (cc y))
                    (grup.nom.fp
                      (ncfp000 bajadas))))
(a1)
```

Un ejemplo de coordinación a nivel de sintagma nominal lo proporciona el siguiente fragmento:

```
      (sn.co
        (sn
          (espec.ms
            (di0ms0 un))
          (grup.nom.ms
            (nccs000 monarca)
            (S.NF.P
              (aq0msp degradado)
              (sp
                (prep
                  (sps00 a))
                (sn
                  (grup.nom.ms
                    (ncms000 gañán)))))))
        (coord
          (cc o))
        (sn
          (espec.fs
            (di0fs0 una))
          (grup.nom.fs
            (ncfs000 reina)
            (sp
              (prep
                (sps00 sin))
              (sn
                (sn
                  (espec.fp
                    (rg más))
                  (grup.nom.fp
                    (ncfp000 consideraciones)))
(a1)
```

Como habrá podido observarse en los ejemplos anteriores, la coordinación se marca añadiendo el sufijo **.co** a las etiquetas de los constituyentes madre. Si como en el caso anterior, los constituyentes hija pertenecen a la misma categoría gramatical, la etiqueta para el nodo madre es la misma. Si, al contrario, la categoría de las hijas es distinta, entonces el nodo madre recibe una de las etiquetas del nodo hija: la menos marcada respecto de la función sintáctica que los nodos coordinados desempeñan. A continuación presentamos algunos de estos casos:

1. En la frase *se felicitan con chocar de palmas y mucho alarde de festividad*, la etiqueta del nodo coordinado es *sn* porque hemos considerado que *sn* era menos marcado que la etiqueta correspondiente a la cláusula infinitiva.

2. Si un sintagma adverbial y uno preposicional se coordinan, la etiqueta del nodo madre es la de sintagma adverbial si los nodos dependen de un verbo, pero la de sintagma preposicional si dependen de un nombre.

3. Si la coordinación afecta a estructuras oracionales con y sin verbo (véase la sección 5.3.3 para el tratamiento de la elipsis verbal), la etiqueta del nodo madre corresponde a la etiqueta de la estructura con verbo.

```
S.co_[
    S_[ El libro es divertido, muy divertido ,]
    coord_[y]
    S*_[su estilo un auténtico regalo]
    .]
(a1)
```

4. Un caso extremo se produce en la siguiente oración: *Que la distribución de procesos, su heterogeneidad y repercusión no afectan a una base fundamentalmente andaluza y, sobre_todo, pensar que América es un inmenso continente donde caben las modalidades de Nuevo_México y de Chiloé* (a1), donde la coordinación se da entre nodos **S*** (*Que la ... andaluza*) y **S.NF.C**, que es la etiqueta correspondiente a las construcciones de infinitivo. Dado que la estructura es independiente, la etiqueta adoptada para el nodo madre es la de **S*.co**.

5. Si las estructuras coordinadas son una subordinada finita y una no finita, la etiqueta para el nodo madre será la de la subordinada finita.

6. Si lo que aparece coordinado es un sintagma preposicional (**sp**) y una construcción de gerundio (**S.NF.A**), como en la frase *a_partir_de ayer y coincidiendo con su cuarto triunfo parcial, ...* (d2), el nodo madre recibe la etiqueta de **S.NF.A**, dado que la función que realizan es la de modificador verbal.

En la coordinación de estructuras oracionales subordinadas es frecuente que el complementador aparezca sólo una vez, al principio. En estos casos, queda situado como nodo hija de la estructura coordinada y hermana de las subordinadas, como puede apreciarse en el ejemplo *la vacuidad de lo que se cree transcendente y no es ...* (a1) cuyo análisis es:

```
...
sn_[ la vacuidad
     sp_[ de
          sn_[ espec.ms lo
               S.F.R.co_[
                    relatiu_[que]
                    S.F.R_[ se cree trascendente]
                    coord_[y]
                    S.F.R_[no es...]
...
(a1)
```

También hemos de mencionar el hecho de que los elementos coordinantes, conjunciones o locuciones conjuntivas, reciben siempre la misma etiqueta *coord*. En el caso de las coordinaciones con doble nexo, como por ejemplo las distributivas y en otras estructuras que hemos tratado del mismo modo (*no sólo ... sino (también)*), los nexos son hermanas de los elementos coordinados.

Por último, sobre la elipsis que se produce en algunas coordinaciones y dado que no incorporamos ningún elemento elíptico (excepción hecha del sujeto), la solución que hemos adoptado ha sido la de mantener el principio general de anotar al nivel más bajo posible siempre y cuando las estructuras resultantes fueran estructuras previstas en el sistema. En el primero de los ejemplos siguientes aparece una coordinación de sintagmas nominales a nivel de sujeto, mientras que en el segundo, dado que hay elipsis verbal pero aparecen tanto el sujeto como el objeto directo del verbo elidido la coordinación se realiza a nivel oracional:

```
 (S
      (sn
        (grup.nom.p
          (pp1cp000 nos)))
      (gv
        (vmip3s0 queda))
      (sn.co
        (sn
          (espec.ms
            (di0ms0 un))
          (grup.nom.ms
            (nccs000 monarca)
            (S.NF.P
              (aq0msp degradado)
              (sp
                (prep
                  (sps00 a))
                (sn
                  (grup.nom.ms
                    (ncms000 gañán)))))))
        (coord
          (cc o))
        (sn
          (espec.fs
            (di0fs0 una))
          (grup.nom.fs
            (ncfs000 reina)
            (sp
              (prep
                (sps00 sin))
              (sn
                (sn
                  (espec.fp
                    (rg más))
                  (grup.nom.fp
                    (ncfp000 consideraciones)))
(a1)

S.co_[
      S_[ Zarrabeitia puso la rebeldía , ]
```

```
      coord_[ y ]
      S*_[ Delgado la gallardía ]
      .
    ]
(d2)
```

En las coordinaciones distributivas hay dos nexos, en lugar de uno. Ambos son hermanas de los elementos coordinados, tal como se muestra en la figura 5.5.

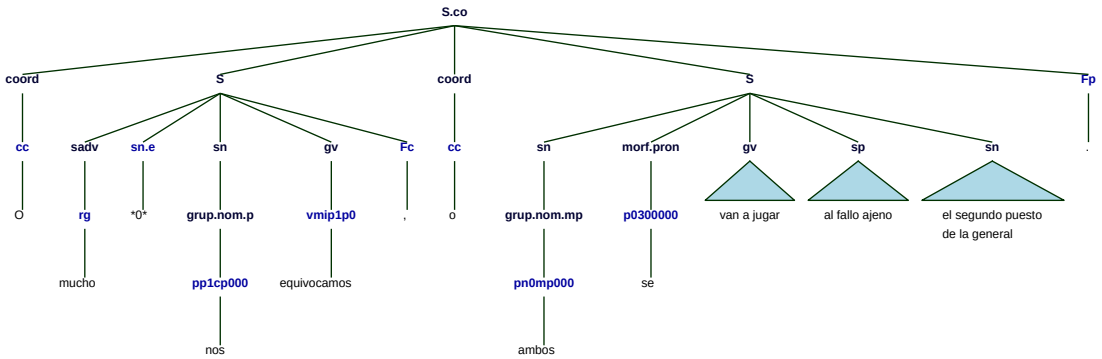

Figura 5.5: Coordinación distributiva

Hay estructuras que tratamos como las coordinadas distributivas (*no sólo ... sino (también); no tanto ... sino;* etc). En este caso los nexos son también hermanas de los elementos coordinados.

```
S_[ ...
    gv_[ parecía ]
    sadv_[ a_salvo ]
    sp.co_[
      coord_[no_sólo_rg]
      sp_[ de sus calamidades ,]
      coord_[ sino_también_cc ]
      sp_[de su historia ]]]
(c1)
```

Otras secuencias que se asimilan a las coordinadas distributivas son las del tipo *(des)de ... a/hasta* cuando no tienen sentido locativo (ni temporal) estricto, sino que marcan el primer y el último elemento de una serie, de una enumeración:

```
S*_[
   sp_[ Por todas partes ,]
   sn_[
      grup.nom.co_[
         grup.nom.fp_[ fotos ]
         coord_[ y ]
         grup.nom.mp_[ pósters ]
         sp_[ de
            sn_[
```

```
                grup.nom.mp_[ ídolos
                    s.a.mp_[ portorriqueños ]
                    sn.co_[
                       :
                       coord_[ desde_sps00 ]
                       sn_[el "alcalde" del Bronx, Fernando_Ferrer,]
                       coord_[ a_sps00 ]
                       sn_[
                          espec.fs_[ la ]
                          grup.nom.fs_[ bomba
                              s.a.fs[ sexy ]
                              sp_[ del momento ]
                              sn_[
                                    :
                                    Iris_Chacón ]]]]]]]]]]
(n1)
```

Un último caso de estructura que asimilamos a la coordinación distributiva es el que se da en la siguiente frase:

```
S_[
...
   gv_[ tiene ]
   sn_[
      espec.ms_[ un ]
      grunp.nom.ms_[ aire
         s.a.ms.co_[
            coord_[ entre_sps00 ]
            s.a.ms_[ rancio ]
            coord_[ y ]
            s.a.ms_[ "kitch" ] ]]]]
(n1)
```

5.3.3. Nodos raíz

La delimitación de las oraciones se hace por criterios ortográficos, siguiendo a Leech, Barnett, y Kahrel (1996). Se considera oración toda secuencia de palabras entre dos puntos o entre signos de puntuación fuertes (signos interrogativos y/o exclamativos); los puntos suspensivos pueden también marcar final de oración.

Los nodos que pueden aparecer como raíz de las oraciones son cuatro: **S, S.co, S*** y **S*.co**[23]. Las dos primeras etiquetas se utilizan para las estructuras oracionales con sujeto y predicado no incluidas en una estructura mayor (exceptuando el caso de la coordinación de oraciones). Se corresponden con las denominadas oraciones independientes o principales. Si estas estructuras carecen de verbo conjugado, reciben la etiqueta **S*** y **S*co**.

5.3.4. La elipsis verbal

En aquellos casos en que el verbo de la oración no aparece de modo explícito y sería necesario tenerlo en cuenta para el análisis sintáctico en el marco que proponemos hemos introducido el símbolo * en la etiqueta de la oración. Las estructuras en que * puede

[23]La coordinación se ha comentado en la sección anterior; los casos de S* se comentan en la sección siguiente.

aparecer son todas las estructuras oracionales[24], que aparecen en el cuadro 5.4. Hay que tener en cuenta, además, que todas ellas pueden coordinarse:

S*	S.F.C*
S.F.R*	S.F.A*
S.F.AComp*	S.F.ACond*
S.F.AConc*	S.F.ACons*
S.NF.A*	S.NF.C*
S.NF.P*	

Cuadro 5.4: Estructuras con verbo elíptico

A continuación aparecen ejemplos de elipsis verbal.
En casos de coordinación:

```
S.co_[
   S_[Don_Lope sorprende a los amantes]
   coord_[y]
   S.co[
     S_[la múltiple ... feminista]
     coord_[y]
     S.co_[
        S_[reduce el amor de Tristana
           y Horacio a una relación epistolar]
        coord [y]
        S*[la vida, a un continuo accidente]]]]
(a24)

S.co_[
   S_[
       sn_[Esos jueces]
       gv_[son]
       sn_[el Centro_para_el_control_de_las__Enfermedades]
       sp-CC_[en Atlanta]
   coord_[y]
   S*_[
       sn_[el_Instituto_para_Investigaciones_Víricas]
       ,_Fc
       sp_[en Moscú]]]
(a21)
```

En oraciones principales:

```
S*_[
    sn_[
        espec.ms_[Un_di0ms0]
        grup.nom.ms_[
            tonto_ncms000
            s.a.ms_[inmortal_aq0cs0]]]
    ._Fp
]
(a26)

S*_[
   ¿_Fia
```

[24]Véase la sección 5.3.5 para el significado de estas etiquetas.

```
   sp[
      prep_[por]
      sn_[qué]]
   ?_Fit
]
(t5)

S*_[
    sa_[
        s.a.mp_[
            indignos
            sp_[de la civilización que les cobija]]]]
(c2)

S*_[
    sadv_[Claro_rg]
    S.F.C_[
        conj.subord_[que_cs]
        gv_[
            tendría_vmic3s0
            que_cs
            infinitiu_[haber_van0000]]
        ...] ]
(a26)
```

En oraciones subordinadas:

```
...
    sp_[
        prep_[con_sps00]
        sn.co[
          sn.co[
             sn_[
                espec.fs_[más_rg]
                grup.nom.fs_[realeza_ncfs000]]
             coord_[y_cc]
             sn_[
                grup.nom.ms_[
                   s.a.ms_[mejor_aq0cs0]
                   porte_ncms000]]]
        S.F.AComp*_[
             conj.subord_[que_cs]
             sn_[
                espec.fs_[la_da0fs0]
                grup.nom.fs_[
                    s.a.fs_[propia_aq0fs0]
                    reina_ncfs000
                S.NF.P_[
                    sodomizada_aq0fsp
                    sp_[
                       prep_[por_sps00]
                       sn_[
                          grup.nom_[Godoy_np00000]]]]]]]]]
    ._Fp
```

Casos de completivas que aparecen entre signos fuertes de puntuación y que también se etiquetan como **S*** son:

```
S*_[
    ¿_Fia
    conj.subord_[Que_cs]
    sn.e_[ 0 ]
    gv_[experimente_vmsp3s0]
    sadv_[primero_rg]
    sn_[
        espec.mp_[los_da0mp0]
        grup.nom.mp_[rigores_ncmp000
            sp.de_[
                del_spcms
                grup.nom.ms_[clima_ncms000]]]]
    ...
    ?_Fit
   ]
```

Otro caso en que se utilizará la etiqueta **S*** es el de la oración *tampoco ellos podrán jamás recuperar su antiguo ser, retroceder a ese estadio, y saben...*; en la segunda perífrasis el auxiliar está elíptico, por lo que aparece la etiqueta S*:

```
S.co[
   S.co[
     S [
       sadv [tampoco]
       sn [ellos]
       gv [podrán jamás recuperar]
       sn [su antiguo ser]]
     ,_Fc
     S* [
       gv [infinitiu [retroceder]]
       sp [a ese estadio...]]]
   coord [y]
   S [ se saben...]]
```

5.3.5. Tipos de S.

Presentamos en esta sección los tipos de oraciones que se consideran en **Cast3LB**, así como las etiquetas utilizadas para su anotación y los criterios de identificación.

El nodo **S** se utiliza sólo para oraciones principales. Para las subordinadas, lo que se ha hecho ha sido incorporar más elementos a la etiqueta **S**. Por un lado se han considerado las oraciones subordinadas no finitas (con verbo no conjugado), cuya etiqueta empieza por **S.NF.** y, por otro, las finitas (con verbo en forma personal), cuya etiqueta empieza con **S.F.**. Todas estas oraciones pueden coordinarse o tener el verbo elíptico, y por tanto, las respectivas etiquetas pueden tener, además, los sufijos **.co** y *.

5.3.5.1. Oraciones no finitas

El cuadro 5.5 recoge tipos de oraciones no-finitas.

Los elementos que pueden incluir las subordinadas no finitas son o bien sólo la forma verbal (infinitivo, participio o gerundio) o bien la forma verbal con sus complementos. En ningún caso, sin embargo, se incluye un nodo vacío para el sujeto.

Las subordinadas no finitas pueden depender de la oración o de algún elemento de la oración.

subordinada no finita completiva	S.NF.C
subordinada no finita adjetiva	S.NF.P
subordinada no finita absoluta	S.NF.PA
subordinada no finita relativa	S.NF.R
subordinada no finita adverbial	S.NF.A

Cuadro 5.5: Tipos de oraciones no finitas

Ejemplos de estos tipos de oraciones son:

1. Subordinada no finita completiva (**S.NF.C**). Estas subordinadas tienen como núcleo un infinitivo, que puede o no llevar complementos.

```
S_[ sn_[ Tony_Rominger ] gv_[ quiere ] S.NF.C_[ infinitiu_[ parecerse
     ] sadv_[ cada vez más ] sp_[ a Miguel_Indurain ] ] .  ] (d2)
```

2. Subordinada no finita adjetiva (**S.NF.P**) y cláusula absoluta (**S.NF.PA**): El núcleo de estas construcciones es, por lo general un adjetivo con etiqueta morfológica *aq0..p*, aunque en las cláusulas absolutas es un participio con etiqueta verbal.

```
S_[
     sadv_[ Generalmente ,]
     gv_[ son ]
     sn_[ los
          grup.nom.mp_[ anticuerpos
                        S.NF.P_[ fabricados
                                 sp_[ por el sistema inmunológico ]
                                ]
                       ]
     sn_[ los vencedores ]
     ...
  ]
(dc10)
```

```
S_[
   S.NF.PA [
        abrumado
        sn_[el lector]]
   se pregunta... ]
(a1)
```

3. Subordinada no finita adverbial (**S.NF.A**): El núcleo de estas oraciones es un gerundio, con o sin complementos.

```
S_[
     sn_[ Otros ]
     gv_[ creen ]
     S.F.C_[ que
           S.NF.A_[
```

```
                gerundi_[ regalando ]
                sn_[ cosas prácticas ]
                  ]
            gv_[ quedan ]
            sadv_[ bien ]
            ].  ]
(c1)
```

En ocasiones, las forma verbal puede estar elíptica, como en el siguiente caso:

```
S_[
 ...
    S.NF.A.co [
       S.NF.A_[ apoyándose en ellos ]
       coord_[ y ]
       S.NF.A*_[ ellos en ti]
] ]
(d2)
```

4. Subordinada no finita de relativo (**S.NF.R**):

```
S_[ ...
   gv_[ tiene ]
   sn_[
      espec.ms_[ un ]
      grup.nom.ms_[ equipo
         s.a.fs_[ cantera ]
         S.NF.R.co_[
            relatiu_[ donde ]
            S.NF.R_[ observar el desarrollo
                     de los jóvenes ]
            coord_[ y ]
            S.NF.R_[ proyectarles hacia la élite]]]]]
(d2)
```

5.3.5.2. Oraciones finitas

Entre las subordinadas finitas se ha establecido la siguiente clasificación: subordinadas finitas completivas, relativas y adverbiales. Las adverbiales, a su vez, se han dividido en dos grupos: por un lado aquéllas que funcionan como complemento circunstancial (las locativas, temporales, modales, causales y finales) y por otro las condicionales, concesivas, consecutivas y comparativas. Este último grupo de subordinadas no funciona tanto como un circunstancial (ningún sintagma puede expresar estas nociones) sino como complemento de todo el predicado o de toda la oración. Dado que la representación que hacemos de la oración es muy plana, la única forma de distinguir estos dos tipos de subordinadas adverbiales es a través de la etiqueta. Para el primer grupo, la etiqueta es **S.F.A** y para el segundo **S.F.ACond, S.F.AConc, S.F.ACons, S.F.AComp** respectivamente. Al igual que ocurría con las oraciones no finitas, las subordinadas finitas pueden también coordinarse y tener el verbo elíptico, y, por tanto, las etiquetas pueden contener también los sufijos **.co** y *****. El cuadro 5.6 recoge los tipos de oraciones finitas.

oración	S
subordinada finita completiva	S.F.C
subordinada finita adjetiva	S.F.R
subordinada finita adverbial	S.F.A
subordinada finita adverbial comparativa	S.F.AComp
subordinada finita adverbial condicional	S.F.ACond
subordinada finita adverbial concesiva	S.F.AConc
subordinada finita adverbial consecutiva	S.F.ACons

Cuadro 5.6: Tipos de oraciones finitas

Ejemplos de estos tipos de oraciones son:

1. Oración (**S**)

```
S_[ Los mbitis se reproducen de forma sexual . ]
(dc1)

S.co_[
      S_[ Zarrabeitia puso la rebeldía , ]
      coord_[ y ]
      S*_[ Delgado la gallardía ]
     ]
(d2)
```

S no establece ningún tipo de distinción referida a las tradicionalmente llamadas modalidades oracionales. Así, por ejemplo, las oraciones interrogativas directas y las oraciones enunciativas, reciben la misma etiqueta.

```
S_[ ¿ Qué podía haber inducido al poeta para verse arrastrado a aquel
      descabellado lance ? ]
(dc10)
```

Y del mismo modo se etiquetan las oraciones exhortativas, desiderativas y dubitativas.

2. Subordinada finita completiva (**S.F.C**):

   Se incluyen en este grupo, las subordinadas con función sustantiva, introducidas por la conjunción *que*, por la conjunción *si*[25], por pronombres interrogativos (en oraciones interrogativas indirectas) y las relativas de *quien* sin antecedente.

```
S_[
     sn_[Javier_Mínguez ]
     gv_[mandó ]
```

[25] En interrogativas indirectas parciales o en oraciones dubitativas del tipo *no sé si...*.

```
     S.F.C_[ que sus corredores atacasen a_partir_de los tres puertos
             de montaña de ayer ]
     .
  ]
(d2)

S_[
     ...
     sn_[ nadie ]
     sn_[ se ]
     gv_[ explicaba ]
     S-F-C_[ cómo habían tenido tiempo de envejecer ]
     .
  ]
(t1)

S_[
     S.F.C_[ quienes nos dedicamos a estos menesteres ]
     neg_[ no ]
     gv_[ podemos trabajar ]
     sp_[ sin entusiasmo] ]
(a1)
```

Las completivas pueden ser término de una preposición, como en el siguiente caso:

```
S_[
   sp_[
     prep_[ Por ]
     sn_[
       espec.fs_[ la ]
       grup.nom.fs_[
          s.a.fs_[ sencilla ]
          razón
          sp_[
             prep_[ de ]
             S.F.C_[ que es casi imposible que ...]]]]]]
(a22)
```

Una aparición de la conjunción *que* que no introduce ninguna subordinada se da en los usos imperativos del subjuntivo, como en la siguiente frase: *Si la señora Aguirre quiere de_veras castigar la suciedad de Madrid en general [...]* **que** *multe al ministro Asunción* (c1). En estos casos, la conjunción forma un nodo unario que depende directamente de la estructura oracional.

3. Subordinada finita adjetiva (**S.F.R**):

   Se incluyen en este grupo las oraciones introducidas por un pronombre o adverbio relativo con y sin antecedente. Las relativas sin antecedente precedidas de artículo no se consideran subordinadas sustantivas, sino adjetivas; la sustantivación es sintáctica y se marca con el artículo. En principio el relativo es el primer elemento de la subordinada, aunque también puede formar parte de un sintagma preposicional.

```
S_[
     sn_[ Ese nombre de bandido generoso del bebé ]
```

```
     gv_[ ha contagiado ]
     sp_[ al papá
          S.F.R_[ , que no puede dejar de ser bandido ... ]]
  ]
(d2)


S_[
     sn_[ Unzaga ]
     gv_[ realizó ]
     sn_[ una exhibición
          sp_[ de
              sn_[ lo
                  S.F.R_[ que debe hacer un gregario para su líder]]]]
     .
   ]
(d2)

S_[
  ...
  sn_[las fondas lúgubres del puerto
        S.F.R.co
          relatiu_donde
          coord_[lo_mismo]
          S.F.R.[ se comía como un rey]
          coord_[ o ]
          S.F.R_ [ se moría de repente ...]]
]
(t1)

S_[
   sn[ la inestabilidad política
        S.F.R_[
              sp_[ en la que ]
              sn.e_[ 0 ]
              gv_[vivimos]
              ]]
   neg_[ no ]
   gv_[ es ]
   sp_[de fondo ]]
(r2)
```

4. Subordinada finita adverbial (**S.F.A**):

   Este grupo de oraciones incluye las subordinadas temporales, locativas, modales, causales y finales. Todas ellas son hijas de un nodo **S**. El criterio para identificarlas es la conjunción que las encabeza. Dado que en este nivel de análisis no hacemos ninguna distinción semántica no se incluye en la etiqueta ninguna marca específica, de modo que todos los tipos de subordinadas adverbiales anteriormente mencionados reciben el mismo tratamiento.

```
S_[
     S.F.A_[ Cuando no es así ]
     sn.e_[ 0 ]
     gv_[ queda ]
     S.NF.P_[ reflejado en los resultados ]
```

```
     .
  ]
(d2)
```

5. Subordinada finita adverbial condicional (**S.F.ACond**):

```
S_[
     S.F.ACond_[ Si se lo hubiese propuesto ,]
     gv_[habría alcanzado ]
     sp_[ a Camargo ]
     ...
  ]
(d2)
```

6. Subordinada finita adverbial concesiva (**S.F.AConc**):

```
S_[
     S.F.AConc_[ Aunque el número de adultos machos y hembras
                          puede ser igual , ]
     sn_[ la relación de hembras disponibles ...]
     gv_[ puede llegar a ser ]
     sp_[ de hasta 7 a 1 ...]
     .
   ]
(dc1)
```

7. Subordinada finita adverbial consecutiva (**S.F.ACons**):

```
S_[
     sn_[ Algunos de los jóvenes ]
     morf.pron_[ se ]
     gv_[ convirtieron ]
     sp_[ en patriarcas venerables ]
     sp_[ con
          sn_[
               sn_[ tanta premura , ]
               S.F.ACons_[ que nadie se explicaba...]
             ]
        ]
    .
   ]
(t1)
```

8. Subordinada finita adverbial comparativa (**S.F.AComp**):

```
S_[ ...
     sn_[ la frecuencia de oscilación del reloj volante ]
     gv_[ resultó ]
     S.NF-C_[
          infinitiu_[ ser ]
          sa_[
               sa_[ más alta ]
               S.F.AComp* [ que la del reloj situado en tierra ]
             ] ]
     .
   ]
(dc2)
```

5.3.6. El nodo INC

En ocasiones, el hilo gramatical se ve interrumpido por elementos que quedan fuera de la oración. Para la etiquetación de estos elementos hemos establecido un nodo con la etiqueta **INC**[26] que tendrá la particularidad de tener como única hija, un nodo con la etiqueta sintagmática correspondiente al tipo de estructura.

Ejemplos de este nodo son:

```
Quizá se inscriba en esa mayor humanización del suizo que, dicen,
se detecta en la Vuelta (d2).

S_[
     sadv_[ Quizá ]
     morfema.verbal_[ se ]
     gv_[ inscriba ]
     sp_[
          prep_[ en ]
          sn_[
               espec_[ esa ]
               grup.nom-fs_[
                    s.a.fs_[ mayor ]
                    humanización
                    sp-de_[ del suizo ]
                    S.F.R_[ que
                         INC_[
                              S_[
                                   sn.e_[ 0 ]
                                   gv_[ dicen ]]]
                         morfema.verbal_[ se ]
                         gv_[ detecta ]
                         sp_[ en la Vuelta ]
                          ]]]]
     .]

- - Es un puta - - dijo. (t1)
S_[
     - -
     sn.e_[ 0 ]
     gv_[ Es ]
     sn_[ una puta ]
     - -
     INC_[
          S_[
               sn.e_[ 0 ]
               gv_[ dijo ]
            ]].]
```

5.3.7. Otros constituyentes

En el interior de cada uno de los tipos de oraciones (principales o subordinadas) pueden aparecer los siguientes nodos: **sn, gv, sp. sadv, sa, conj.subord, coord, infinitiu, gerundi, interjeccio, neg, morfema.verbal, morf.pron**. El cuadro 5.7 presenta una breve descripción de cada uno de ellos.

[26] Similar a la etiqueta *PRN* utilizada en el PennTreeBank (Bies et al., 1995).

nodo	glosa
sn	sintagma nominal
gv	grupo verbal
sp	sintagma preposicional
sadv	sintagma adverbial
sa	sintagma adjetivo
conj.subord	conjunción subordinante
coord	conjunción coordinante
infinitiu	verbo en infinitivo
gerundi	verbo en gerundio
interjeccio	interjección
neg	adverbio de negación
morfema.verbal	uso particular de la forma SE
morf.pron	usos particulares de los pronombres átonos

Cuadro 5.7: Nodos en el seno de las oraciones

La estructura de los sintagmas es la siguiente: la mayoría de ellos tiene un núcleo léxico (sintagma nominal, adverbial y adjetivo) que se corresponde con la categoría léxica que le da nombre. Estos núcleos pueden recibir complementos antepuestos o postpuestos. A continuación presentamos ejemplos de todos ellos.

5.3.7.1. El sintagma nominal

El sintagma nominal **sn** tiene como núcleo un nombre o un pronombre. Puede ser un sintagma unario, pero también puede tener una estructura muy compleja. Un caso particular de sintagma nominal es el elíptico (**sn.e**), utilizado sólo para los casos en que el sujeto de las oraciones finitas no aparece explícitamente.

1. núcleo de **sn**

   a) el nombre (común o propio)

```
S_[
     sn.e_[ 0 ]
     gv_[ vio ]
     sn_[
          grup-nom.mp_[
               cromos
               sp_[ de revistas ]
                      ]
        ]
(t1)
```

   b) el pronombre

```
S_[
     sn_[ Nadie ]
     sn_[ lo ]
     gv_[ entendió ]
     .
  ]
(t1)
```

2. **sn** unario.
   El sintagma nominal suele ser unario si está constituido por un nombre propio o un pronombre.

```
S_[ sn_[ grup.nom.ms_[ Javier_Mínguez ] ]
    gv_[ mandó ]
    S.F.C_[ que ...]  .  ]
(d2)
```

3. **sn** formado por determinante y núcleo:

```
S_[
    ...
    S.F.C_[
         conj.subord_[ que ]
         sn_[
              espec.mp_[ sus ]
              grup.nom.mp_[ corredores ]
           ]].]
(d2)
```

4. complementos de **sn**. Los elementos que pueden aparecer como complementos del sintagma nominal son el sintagma adjetivo, el sintagma preposicional, las subordinadas relativas y otros sintagmas nominales en aposición, además de subordinadas completivas precedidas de preposición; o combinaciones de todos ellos.

   Estos elementos son hermanas del núcleo y sólo en algunos casos en que un complemento depende de dos nombres coordinados los complementos se han adjuntado a **sn**.

   Los siguientes ejemplos muestran la variedad de complementos que puede recibir el nombre.

   a) **sintagma adjetivo (s.a.)**

```
...
sn_[ espec.mp_[ los ]
     grup.nom.mp_[
          caserones
          s.a.mp_[ desiertos ]
                 ]
   ]
...
(t1)
```

   b) **sintagma preposicional (sp)**.

```
...
sn_[ espec.fs_[ la ]
     grup.nom.fs_[
          excelencia
          sp_[ del soneto ]
                 ]
   ]
...
```

```
(t1)

...
sp_[
     prep_[ con ]
     sn.co[
          sn_[ los cuellos ]
          coord_[ y ]
          sn_[ los puños]
          sp_[
               prep_[ como ]
               sn_[ hostias recién planchadas ]
             ]]]
(t1)
```

c) **subordinada relativa (S.F.R)**

```
...
S.NF.A_[
     gerundi_[ temiendo ]
     sn_[
          espec.fs_[ una ]
          grup.nom.fs_[
               revelación
               S.F.R_[ que lo perturbara de por vida ]
                     ]]]
(t1)
```

d) **sintagma nominal en aposición**

```
S_[
     sn_[
          espec.fs_[ la ]
          grup.mon.fs_[ Orquídea_de_Oro ]
          sn_[
               Fc_,
               espec.ms_[ el ]
               grup.nom.ms_[ galardón más codiciado de la
                             poesía nacional ]
               Fc_,
             ]
        ]
     sn_[ le ]
     gv_[ fue adjudicada ]
     ...
   ]
(t1)
```

e) **subordinada de gerundio**

```
sn_[
   grup.nom.mp.co_[
      grup.nom.mp_[ Aviones
         S.NF.A.co_[
            S.NF.A_[ gerundi_[ aterrizando ]]
            coord_[ o ]
            S.NF.A_[ gerundi_[ despegando ]]]]
      ,
      grup.nom.mp_[ cristales
            S.NF.A_[ gerundi_[ aterrizando ]
```

```
                              sp_[ sin motivo ]]]]]
    (a12)
```

f) **combinaciones de complementos**

```
    ...
    sn_[
         espec.ms_[ el ]
         grup.nom.ms_[
              escándalo
              s.a.ms_[ público ]
              S.F.R_[ que provocó aquella decisión insólita ]
                    ]
       ]
    ...
    (t1)
```

5. un elemento sustantivado también puede formar un sintagma nominal:

```
S_[
     ...
     gv_[ eran ]
     sn_[
          espec.ms_[ los ]
          sp_[
               prep_[ de ]
               sn_[ las fondas lúgubres del puerto ]
             ]
        ]
     ...
]
(t1)
```

5.3.7.2. El grupo verbal

Tal como ya se ha indicado en 5.3.1 no hemos considerado la existencia de un nodo *sintagma verbal* o *predicado* de la oración, por lo que la anotación manual de corpus no implica muchos cambios en este sentido.

El único elemento que se ha tenido en cuenta aquí afecta a las perífrasis verbales que no vienen reconocidas como tales desde el analizador sintáctico. Como se comentó en el capítulo 4 hay algunos casos de perífrasis ambiguas, es decir, elementos que según el contexto en que aparecen deben considerarse perífrasis pero que en otros casos deben considerarse como dos formas verbales independientes. Como norma general, las perífrasis ambiguas no vienen reconocidas por la gramática, por lo que sólo la información proporcionada por el contexto puede indicar si debe construirse la agrupación perifrástica o no, como en el caso de *querer + infinitivo*, que es perífrasis verbal si tiene un sujeto inanimado y si no es posible sustituir el infinitivo por un pronombre. Un ejemplo lo proporciona la siguiente frase *¿Qué querrá decir incendiar el conjunto?* (c2).

Un fenómeno que puede producirse es que las perífrasis tengan elementos incrustados entre el auxiliar y la forma no finita, como en el siguiente caso: *tampoco ellos podrán jamás recuperar...*. Dado que no se alterara el orden superficial de las palabras, el adverbio queda incrustado en el grupo verbal, lo que significa representarlo de este modo:

```
S [
   sadv [tampoco]
   sn [ellos]
   gv [
     podrán
     sadv [jamás]
     recuperar]
   ...]
(a1)
```

Por lo general, los elementos incrustados en las perífrasis verbales son adverbios y sintagmas nominales en función de sujeto. Un caso extremo lo presenta la siguiente frase *debe o no uno abrigarse*, cuyo análisis es:

```
...
   gv [
     debe
     INC[
       S*[
       coord [o]
       neg [no]
     ]
     sn [uno]]
     abrigarse]
(a11)
```

5.3.7.3. El sintagma preposicional

El sintagma preposicional puede aparecer prácticamente en cualquier nodo del árbol de análisis: como hija directa de oración o en el interior de cualquier sintagma, como muestran los siguientes ejemplos:

```
S_[
   sn_[ La carrera ]
   gv_[ está ]
   sp_[ en sus manos]
   sp_[ hasta esos extremos]
   .]
(d2)
```

```
S_[
   sn_[ El suizo ]
   gv_[ imitó ]
   sp_[ a Miguel_Indurain ]
   sp_[ en su forma
           sp_[ de interpretar una jornada
                        sp_[de alta montaña ]]]
    ]
(d2)
```

5.3.7.4. El sintagma adverbial

Los sintagmas adverbiales suelen ser hijas de estructuras oracionales, aunque también pueden ocupar otras posiciones en la oración.

```
S_[
  ...
  S.F.R_[
      sp_[ a las que ]
      sn_[ Montaigne ]
      gv_[llamó ]
      sadv_[ sencillamente ]
      sn_[ " ensayos " ]]
      .
  ]
(a1)

S_[
   sn_[ El libro ]
   gv_[ es ]
   sa_[
     sadv[ muy ]
     divertido ]]

(a1)
```

5.3.7.5. El sintagma adjetivo

El sintagma adjetivo suele aparecer en el interior de un sintagma nominal, aunque también puede depender directamente de una estructura oracional.

```
S_[ ¿ No va a estar
          sn_[ el
               s.a.ms.co [ misterioso
                           ,
                           enigmático
                           ,
                           extraño
                           coord_[ y ]
                           genial ]
               delantero ]
         en el Bernabéu y en Atenas ?  ]
 (d2)

S_[
      coord_[ Y ]
      sn_[ Romario ]
      sa_[ , tan sorprendente en la vida como en el área
                  adversaria ... ]
(d2)
```

5.3.8. Tratamiento de los signos de puntuación

En la tradición de la lingüística de corpus, los signos de puntuación se consideran una clase más de elementos del texto y suelen tratarse como las otras clases de palabras, es decir, reciben su propia etiqueta. En el caso de **Cast3LB**, estos elementos mantienen esta etiqueta y no se añade al árbol de análisis ningún nodo suplementario.

Para su tratamiento en los árboles de análisis se han propuesto distintas soluciones: Bies et al. (1995), Abeillé, Toussenel, y Chéradame (2002) o Moreno, López, y Sánchez (1999). Lo cierto es que los signos de puntuación fuertes suelen ser el último elemento de

la oración en todos los casos; pero en lo que respecta a los signos que aparecen dentro de la oración, las soluciones propuestas difieren ampliamente.

Si nos centramos en los dos puntos (:), en Bies et al. (1995) se consideran sólo dos posibilidades: los dos puntos en las aposiciones y en otros contextos distintos llamados por los autores *colorful environments*. En Moreno, López, y Sánchez (1999) se tratan como el resto de signos de puntuación: *they will be the sister or the daugther of their closer constituent, depending on their position inside the sentence*. Por último, en Abeillé, Toussenel, y Chéradame (2002) se establece que los signos de puntuación en general no se incluyen en ningún constituyente; y sin embargo, algunas comas aparecen en el interior de cláusulas subordinadas.

El tratamiento de estos elementos no es un tema trivial; y menos si se pretende proporcionar una anotación consistente. En el marco de **Cast3LB** hemos establecido la siguiente casuística y adoptado la solución que comentamos para cada caso. En ocasiones, los signos de puntuación se utilizan para delimitar constituyentes (actúan como signos parentéticos); en este caso, aparecen como primer y último elemento del nodo parentizado. En los casos de coordinación asindética, los signos de puntuación ocupan el lugar de las conjunciones coordinantes o son hermanas de los elementos coordinados. Por último, si no se da ninguno de los casos anteriores, los signos de puntuación internos son hijas del nodo más alto a su derecha en el árbol.

En lo referente a los dos puntos, éstos suelen delimitar e identificar una estructura específica. Según la Real Academia Española de la Lengua, los principales usos de este signo de puntuación son: al principio o al final de una enumeración; como elementos introductorios de discurso directo; como delimitadores de un ejemplo respecto del resto de la oración; y, en general, como conectores clausales que pueden expresar distintas relaciones (causa, consecuencia, conclusión, resumen, explicación de lo que antecede, etc.).

Con el objetivo de sistematizar el análisis de este signo, se extrajeron 35 oraciones que lo contenían. Tras un análisis detallado de las mismas se estableció la siguiente tipología:

1. Los dos puntos introducen discurso directo, como en el siguiente ejemplo:

```
S_[
   sn_[ La gran cuestión ]
   gv_[ era ]
   S.F.C_[
          :
          ¿
          morf.pron_[ se ]
          gv_[ tendrá
               sn_[ Romario ]
               que ir ]
          sp_[ a Río ]
          ? ] ]
(d2)
```

   En este caso se tratan como si fueran una conjunción subordinante, por lo que son el primer elemento de la cláusula.

2. En otros casos, los dos puntos son el inicio (o el final) de una enumeración. Es su uso más frecuente en el corpus, y los analizamos como primer (o último) elemento de la serie que introduce, tal como muestran los siguientes ejemplos:

```
S_[
   sn.e_[ 0 ]
   gv_[ amenaza ]
   sp_[
     con
     S.NF.C_[
      infinitiu_[ ganarlo ]
      sn_[ todo
         sn.co_[
            :
            sn_[ la general ]
            ,
            sn_[ la montaña ]
            ,
            sn_[ la clasificación por puntos] ] ] ] ]
(d2)


S*_[
   sp_[ A su lado , ]
   sn_[ el apoyo
      sp_[ de
         sn [
            grup.nom.co [ un amigo, compadre, compañero
                          leal y alter_ego : ]
            Stoichkov ] ] ] . ]
(d2)
```

3. Por último, los dos puntos introducen un elemento complejo de la oración, por lo general una estructura oracional cuyo referente es todo el texto anterior, como ocurre en la oración *Entonces decidió cambiar la perspectiva: desprestigió a la heroína devolviéndola a su hogar y la convirtió en una mujer común, salvo en su soberana belleza.* En este caso, los dos puntos se incluyen en el nodo más alto a su derecha y todo el constituyente que encabezan se adjunta al elemento que refiere:

```
S_[
  sadv_[ Entonces ]
  gv_[ decidió ]
  S.NF.C_[
   S.NF.C_[ cambiar la perspectiva ]
   S.co_[
     :
     S_[ desprestigió ... hogar ]
     coord_[ y ]
     S_[ la convirtió ... belleza ] ] ] . ]
(a1)
```

Cuatro de los ejemplos hallados en el corpus no se corresponden con ninguno de los usos anteriormente descritos, aunque comparten algunas de sus características. En concreto, los cuatro son enumeraciones que tienen como referente un elemento que no les precede de

forma inmediata. La solución adoptada ha sido la adjunción del nodo que introducen a toda la estructura anterior, como en la frase: *Los demás ya son conocidos: la composición del resto del podio, la probabilidad de que Toni_Rominger se apunte también la montaña y la regularidad* (d2).

```
S_[
   S_[ Los demás ya son conocidos]
   sn.co_[ : la composición ... regularidad ]]
```

5.3.9. La adjunción de nodos

En ocasiones aparecen estructuras que dependen de los nodos S (o de sus equivalentes). Dada la representación plana que estamos utilizando, resulta imposible poder marcar esta relación. Para ello utilizamos la estructura de adjunción, que consiste en duplicar el nodo madre. Ejemplos de adjunción son los siguientes: *Bien es verdad que la idea podría tener una vertiente social y hasta religiosa importante: el ocio empobrecido obligaría a pasar las jornadas en el hogar con lo que ello supondría para el aumento de la natalidad y para el rejuvenecimiento de la población, que ya se sabe lo que ocurrió en Nueva_York cuando lo del apagón; además, podría surgir algún sucedáneo del padre Peyton que indujese a rezar el rosario en familia en_vez_de jugar a Monopoly en las tardes de ocio forzoso.*

El análisis de esta oración sería:

```
S [...
    S [el ocio ...]
    sp [con lo que ello...]]
(a18)
```

Otros casos en que esto sería aplicable aparecen en las frases siguientes cuyo análisis parcial se muestra a continuación:
Visto_que lo de menguar la soldada para repartir las pocas habas del puchero no parecía muy factible en todos los casos, llegó el pepero señor Aznar, agarró la sartén sin temor a quemarse y lanzó lo de que para salir de la crisis lo que había que hacer era trabajar más, más días y más horas, escuernarse sin tregua, o_sea, en un estajanovismo digno de mejor causa.
A primera hora de la mañana hacía gimnasia con Lali_Ruiz en el monitor de la derecha, mientras descifraba las noticias americanas de la ABC, lo que en ocasiones provocaba algunos momentos de tirantez ; al_fin_y_al_cabo, 30 flexiones estirando la pierna izquierda y otras 30 con la derecha no son el mejor caldo_de_cultivo para recibir los tambores de guerra en el Golfo.

```
S.co [
     S.F.A [Visto_que lo de menguar... en todos los casos]
     S.co [
        S [llegó...]
        ,_Fc]
        S [agarró...]
        coord [y]
        S [lanzó...]]
(a18)
```

```
S [
    S [A primera hora ... americanas de la ABC ,]
    sn [lo que en ocasiones...]]
(a2)
```

Recibe también este tratamiento la comparación. La estructura comparativa se adjunta al nodo que contiene el adverbio comparativo, como en las frases siguientes:
Claro que en su época no había genios, como aquel baturro sordo , que pintó a la otra Cayetana, aquella gran española que andaba con más realeza y mejor porte que la propia reina sodomizada por Godoy .

```
...
S.F.R [
     relatiu [que]
     gv [andaba]
     sp [
        prep [con]
        sn.co [
             sn.co [
                 sn [más realeza]
                 coord [y]
                 sn [mejor porte]]
             S.F.AComp* [que la propia reina...]]]]
(a20)
```

coincide menos con los instantes y tiempos del teatro que con los puntos de vista...

```
...
   gv_[coincide]
   sp-CREG_[
       sp_[
           sadv_[menos]
           sp_[
               prep_[con]
               sn_[
                   espec_[los]
                   grup.nom.mp.co[
                       grup.nom.mp_[instantes]
                       coord_[y]
                       grup.nom.mp_[tiempos]
                       sp_[del teatro]]]]]
       S.F.AComp*[
           conj.subord_[que]
           sp_[con los puntos de vista...]]]
...
(a24)
```

5.4. Conclusión

En este capítulo se han presentado los criterios de anotación de un banco de árboles sintácticos del español que se está construyendo actualmente. Dada la complejidad del problema, se ha optado por anotar exclusivamente aquellos elementos que están presentes de forma explícita en la oración. Esta decisión ha implicado un tratamiento simplificado de algunos aspectos sintácticos como la coordinación y determinados tipos de subordinadas,

que se dejan para una fase posterior. Se han dado soluciones a problemas específicos del análisis sintáctico aplicado a corpus irrestrictos: se ha abandonado el concepto de oración entendido como sujeto y predicado, para optar por el de una lista de constituyentes. La razón estriba en el carácter libre del orden de constituyentes en español. La introducción de información sobre las funciones suplirá este déficit. Se han dado, además, criterios para el etiquetado de los signos de puntuación, que han permitido la consistencia notacional. Con el desarrollo de este *Treebank* se pretende disponer de datos para realizar una investigación en profundidad orientada al desarrollo de una gramática computacional del castellano de alto nivel.

Capítulo 6

Conclusiones

El trabajo de investigación presentado se enmarca en el área de procesamiento automático de corpus. Se trata de una propuesta de codificación de los niveles morfosintáctico y sintáctico del español, lingüísticamente fundamentada y acorde con las propuestas de estándares de codificación vigentes. Se ha realizado un estudio detallado de los problemas de análisis y codificación y se han contrastado con el procesamiento efectivo de un corpus de pruebas, el corpus CLiC-TALP, con el objetivo de dar respuesta a un abanico amplio de problemas.

Las aportaciones que se han realizado atañen tanto a la mejora de recursos existentes como al desarrollo de nuevas estructuras de datos y la definición de parámetros para el desarrollo de nuevos recursos. Con este trabajo se cierra el procesamiento automático de textos irrestrictos en español desde el análisis morfológico hasta la sintaxis superficial. Esta propuesta de anotación de corpus está fundamentada desde el punto de vista lingüístico y puede utilizarse como base para otros trabajos en el área.

En lo que respecta a la mejora y sistematización de recursos ya existentes se ha trabajado en dos direcciones: la redefinición del tagset o etiquetario previo del analizador morfológico y la introducción de conocimiento lingüístico en un sistema de desambiguación automática.

En el primer caso, se han estudiado las diferentes clasificaciones de palabras realizadas desde la teoría gramatical y se ha justificado en cada caso la decisión tomada. En concreto, cabe destacar la labor realizada en el ámbito de la sistematización de las categorías cerradas, para las cuales no existe acuerdo en el terreno estrictamente lingüístico. Se ha dado una fundamentación lingüística a los temas planteados tratando de reflejar al máximo las peculiaridades de la lengua y, al mismo tiempo, se ha tenido en cuenta la viabilidad computacional de la solución adoptada.

En el segundo caso, nuestra aportación ha consistido en incorporar a un desambiguador automático restricciones basadas en conocimiento lingüístico que tienen en cuenta no sólo las categorías y subcategorías afectadas por la ambigüedad, sino también, y muy especialmente, los rasgos morfológicos de estas categorías, los lemas y las propias palabras. El resultado de la incorporación de conocimiento lingüístico a este sistema automático ha sido una mejora en los resultados cercana al 2 %.

En lo concerniente a la creación de nuevos recursos para el procesamiento automático del español, nuestra aportación ha consistido en el desarrollo de una gramática para el análisis superficial (GramEsp) y un corpus con anotación morfosintáctica (corpus CLiC-TALP) validado manualmente. Aunque GramEsp está concebida para el análisis superficial del español, se ha explotado al máximo la capacidad del formalismo para mostrar la estructura interna de los elementos. Se trata de una gramática de amplia cobertura que opera con un analizador robusto, de modo que es posible analizar textos irrestrictos. El desarrollo de esta gramática ha implicado la adaptación del concepto de chunk para el español, dado que esta unidad de análisis es dependiente de la lengua y no existe para ella una definición estándar universalmente válida. Los chunks definidos responden a criterios lingüísticos, y tienen como finalidad conseguir un análisis al nivel máximo de profundidad con la limitación que supone disponer sólo de la información formal proporcionada por la palabra y/o la etiqueta morfosintáctica.

Nuestra aportación en la creación del corpus CLiC-TALP ha consistido en la definición de los criterios de desambiguación manual del mismo y la definición de etiquetas específicas no contempladas en el análisis automático. Cabe destacar su utilidad práctica tanto desde el punto de vista del estudio de la lengua como desde una perspectiva aplicada ya que constituye un gold estándar para los sistemas de aprendizaje de desambiguación automática.

Por último, tras estos procesos para el tratamiento automático del español, se han establecido los criterios para la anotación sintáctica manual de corpus. El punto de partida es el análisis sintáctico superficial proporcionado por GramEsp y el resultado será un Treebank de 100.000 palabras (Cast3LB). La anotación sintáctica está prevista en distintas fases y aquí se ha presentado la primera: la parentización y etiquetación de los constituyentes. Se han proporcionado criterios para el tratamiento de distintos fenómenos como la coordinación o la etiquetación de los signos de puntuación. La mayoría de problemas surge de la representación arbórea de los treebanks, de modo que es especialmente importante proporcionar criterios coherentes y consistentes. El esquema de anotación que se propone pretende ser básico o neutro en el sentido de no seguir ninguna teoría concreta, pero proporcionando información apta para el estudio de la lengua española desde cualquier perspectiva, sea ésta estrictamente lingüística o computacional.

Bibliografía

Abaitua, J. 2000. Tratamiento de corpora bilingües. En *Seminario La ingeniería lingüística en la sociedad de la información*. Fundación Duques de Soria. disponible: http://www.serv-inf.deusto-es/abaitua/konzeptu/ta/soria00.htm.

Abeillé, A., F. Toussenel, y M. Chéradame. 2002. Corpus le Monde. Annotation en constituants. Guide pour les correcteurs. Informe técnico, LLF, UFRL. dernière mise à jour: 10-juillet-2002.

Abeillé, A., L. Clément, y A. Kinyon. 2000. Building a treebank for French. En *Proccedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, páginas 87–94, Athens, Greece.

Abeillé, A., L. Clément, y A. Kinyon. 2003. Building a treebank for French. En A. Abeillé, editor, *Building and Using syntactically annotated corpora*, Language and Speech. Kluwer, Dordrecht.

Abney, S. 1991. Parsing by Chunks. En R. Berwick S. Abney, y C. Tenny, editores, *Principle-Based Parsing*. Kluwer Academic. disponible: http://www.sfs.nphil.uni-tuebingen.de/ abney/.

Abney, S. 1995. Chunks and Dependencies: Bringing Processing Evidence to Bear on Syntax. En *Computational Linguistics ond the Foundations of Linguistic Theory*.

Abney, S. 1996a. Part-of-Speech Tagging and Partial Parsing. En K. Church S. Young, y G. Bloothooft, editores, *Corpus-Based Methods in Language and Speech*. Kluwer Academic. disponible: http://citeseer.nj.nec.com/abney96partspeech.html.

Abney, S. 1996b. Partial Parsing via Finite-State Cascades. En *Proceedings of the ESSLLI'96 Robust Parsing Workshop*. disponible: http://www.sfs.nphil.uni-tuebingen.de/ abney/.

Aduriz, I. 2000. *EUSMG: morfologiatik sintaxira murriztapen gramatika erabiliz*. Ph.D. tesis, Euskal Herriko Unibertsitatea.

Aduriz, I., I. Aldezabal, M. Aranzabe, B. Arrieta, J.M. Arriola, A. Atutxa, A. Diaz de Ilarraza, K. Gojenola, M. Oronoz, y K. Sarasola. 2002. Construcción de un corpus etiquetado sintácticamente para el euskara. *Procesamiento del Lenguaje Natural*, (29), Septiembre.

Aduriz, I., M. Aranzabe, J.M. Arriola, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa, y Z. Urizar. 2003. Methodology and steps towards the construction of a Corpus of written Basque tagged in morphological, syntactic, and semantic levels for the automatic processing (IXA Corpus of Basque, ICB). En *Proceedings of the Corpus Linguistics*, Lancaster.

Aduriz, I., J.M. Arriola, X. Artola, K. Gojenola A. Díaz de Ilarraza, y M. Maritxalar. 1997. Morphosyntactic Disambiguation for Basque based on the Constraint Grammar Formalism. En *Proceedings of the 2nd Conference on Recent Advances in Natural Language Processing (RANLP)*, páginas 282–287, Tzigov Chark, Bulgaria.

Afonso, S., E. Bick, R. Haber, y D. Santos. 2002. 'Floresta Sintá(c)tica': a Treebank for Portuguese. En *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02)*, páginas 1698–1703, Las Palmas de Gran Canaria, Spain, May.

Ageno, A. 2003. *An Island-Driven Parsing System.* Ph.D. tesis, Departament de Llenguatges i Sistemes Informàtics de la Universitat Politècnica de Catalunya .

Alarcos, E. 1987. *Estudios de gramática Funcional del Español.* Biblioteca Románica Hispánica. Gredos, Madrid, 3 edición.

Alarcos, E. 1994. *Gramática de la lengua española.* Espasa Calpe, Madrid.

Alcina, J. y J. M. Blecua. 1989. *Gramática española.* Ariel, Barcelona.

Alcoba, S. 1999. La flexión verbal. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española.* Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 75, páginas 4915–4991.

Allen, J., 1995. *Natural language Understanding*, capítulo Grammars and Parsing, páginas 40–79. The Benjamin/Cummings Publishing Company, Inc.

Alonge, A., N. Calzorali, P. Vossen, I. Blocksma, I. Castellón, M.A. Martí, y Peters W. 1998. The Linguistic Design of the EuroWordNet. Kluwer.

Alonso-Cortés, A. 1999. Las construcciones exclamativas. La interjección y las expresiones vocativas. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española.* Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 62, páginas 3993–4050.

Arévalo, M. 2001. Gramática para la detección y clasificación de entidades con nombre. Master's thesis, Departamento de Lingüística, Universidad de Barcelona.

Arévalo, M., L. Alonso, M. Taulé, y M.A. Martí. 2001. Documentación sobre el analizador morfológico para el castellano (AMCAST). X-tract wp 01/01, Universitat de Barcelona.

Arévalo, M., X. Carreras, L. Màrquez, M.A. Martí, L. Padró, y M.J. Simón. 2002. A Proposal for Wide-Coverage Spanish Named Entity Recognition. *Procesamiento del Lenguaje Natural*, (28):63–80, Mayo.

Arévalo, M., M. Taulé, y M.A. Martí. 2001. Documentación sobre el analizador morfológico para el catalán (AMCAT). X-tract wp 08/01, Universitat de Barcelona.

Atserias, J. y H. Rodríguez. 1998. TACAT: TAgged Corpus Text Analyzer. Informe técnico, Software Department (LSI). Technical University of Catalonia (UPC).

Bello, A. 1847. *Gramática de la Lengua Castellana destinada al uso de los americanos. Con las Notas de Rufino José Cuervo.* Arco/Libros, S.A. Estudio y Edición de Ramón Trujillo, 1988.

Bemova, A., J. Hajic, B. Hladka, y J. Panevova. 1999. Morphological and Syntactic Tagging of The Prague Dependency Treebank. Journées Atala, Corpus annotés pour la syntaxe, Paris, June. disponible: http://talana.linguist.jussieu.fr/treebanks99/.

Benveniste, C.B. 1998. *Estudios lingüísticos sobre la relación entre oralidad y escritura.* Gedisa, Barcelona.

Bertomeu, N. y L. Mayol. 2003. Tractament de l'ambigüitat determinant pronom adjectiu adverbi dels quantificadors en un sistema de desambiguació morfológica automàtica. Informe Técnico X-Tract-II WP-01/03, Universitat de Barcelona.

Böhmova, A. y E. Hajicova. 1999. How Much of the Underlying Syntactic Structure can be Tagged Automatically. Journées Atala, Corpus annotés pour la syntaxe, Paris, June. disponible: http://talana.linguist.jussieu.fr/treebanks99/.

Böhmova, A., J. Panevová, y P. Sgall. 1999. Syntactic Tagging: Procedure for the Transition from the Analytic to the Tectogrammatical Tree Structures. En *Proceedings og the Second Workshop on Text, Speech, Dialogue*, Marianske lazne, Czech Republic. disponible http://ufal.mff.cuni.cz/pdt/pdt_05.html.

Bies, A., M. Ferguson, K. Katz, y R. MacIntyre. 1995. Bracketing Guidelines for Treebank II Style Penn Treebank Project. Informe técnico, LDC.

Bod, R. 2003. An Efficient Implementation of a New DOP Model. En *proceedings of the 10th EACL Conference*, Budapest.

Boguslavsky, I., I. Chardin, S. Grigorieva, N. Grigoriev, L. Iomdin, L. Kreidlin, y N. Frid. 2002. Development of a Dependency Treebank for Russian and its possible Applications in NLP. En *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02)*, páginas 852–856, Las Palmas de Gran Canaria, Spain, May.

Bosco, C., V. Lombardo, D. Vassallo, y L. Lesmo. 2000. Building a treebank for Italian: a Data-driven Annotation Schema. En *Proccedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, páginas 99–105, Athens, Greece.

Bosque, I. 1991. *Las categorías gramaticales.* Numero 11 en Textos de Apoyo. Lingüística. Ed. Síntesis.

Bosque, I. 1999. El nombre común. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española.* Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 1, páginas 3–75.

Bosque, I. y V. Demonte. 1999. *Gramática Descriptiva de la Lengua Española.* Espasa-Calpe, Real Academia Española de la Lengua, Madrid.

Brants, S., S. Dipper, S. Hansen, W. Lezius, y G. Smith. 2002. The TIGER treebank. En *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Brants, T. y O. Plaehn. 2000. Interactive Corpus Annotation. En *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*, Athens, Greece.

Brants, T., W. Skut, y H. Uszkoreit. 2003. Syntactic Annotation of a German Newspaper Corpus. En A. Abeillé, editor, *Building and Using syntactically annotated corpora*, Language and Speech. Kluwer, Dordrecht. disponible: http://treebank.linguist.jussieu.fr/toc.html.

Brucart, J.M. 1999. La elipsis. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española.* Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 43, páginas 2787–2863.

Cabrera, J. C. Moreno. 1987. *Fundamentos de sintaxis general.* Numero 4 en Textos de Apoyo, Lingüística. Síntesis.

Carmona, J., S. Cervell, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé, y J. Turmo. 1998. An Environment for Morphosyntactic Processing of Unrestricted Spanish Text. En *Proceedings of the First Conference on Language Resources and Avaluation. LREC'98*, páginas 915–922, Granada.

Carreras, X., L. Màrquez, y L. Padró. 2003. Named Entity Recognition For Catalan Using Only Spanish Resources and Unlabelled Data. En *Proceedings of the 10th Conference of the European Chapter of the Association of Computational Linguistics*, Budapest, Hungary.

Carroll, J., T. Briscoe, y A. Sanfilippo. 1998. Parser Evaluation: a Survey and a New proposal. En *Proceedings of the First Conference on Language Resources and Avaluation. LREC'98*, páginas 447–454, Granada.

Carroll, J., G. Minnen, y T. Briscoe. 1999. Corpus Annotation for Parser Evaluation. Journées Atala, Corpus annotés pour la syntaxe, Paris, June. disponible: http://talana.linguist.jussieu.fr/treebanks99/.

Carroll, J., G. Minnen, y T. Briscoe. 2003. Parser evaluation using a grammatical relation annotation scheme. En A. Abeillé, editor, *Building and Using syntactically annotated corpora*, Language and Speech. Kluwer, Dordrecht. disponible: http://treebank.linguist.jussieu.fr/toc.html.

Cartagena, N. 1999. Los tiempos compuestos. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española.* Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 45, páginas 2935–2975.

Ces. 2000. Corpus Encoding Standard. Disponible: http://www.cs.vassar.edu/CES/. Version 1.5. Last modified 20 March 2000.

Civit, M. 2000. Guía para la anotación morfológica de corpus. Informe Técnico X-Tract WP-00/06, Universitat de Barcelona. disponible: http://clic.fil.ub.es/∼civit.

Civit, M. 2002. Guía para la anotación sintáctica de Cast3LB: un corpus del español con anotación sintáctica, semántica y pragmática. Informe Técnico X-Tract-II WP-02/01, 3LB WP 02-01, Universitat de Barcelona. disponible: http://clic.fil.ub.es/∼civit.

Civit, M., I. Castellón, y M.A. Martí. 2001a. Creación, etiquetación y desambiguación de un corpus de referencia del español. *Procesamiento del Lenguaje Natural*, (27):21–28, Septiembre. disponible: http://clic.fil.ub.es/∼civit.

Civit, M., I. Castellón, y M.A. Martí. 2001b. Joven periodista triste busca casa frente al mar o la ambigüedad en la anotación de corpus. Congreso Internacional sobre nuevas tendencias de la lingüística, Granada, November. disponible: http://clic.fil.ub.es/∼civit.

Civit, M. y M.A. Martí. 2002. Design Principles for a Spanish Treebank. En *Proceedings of the First Workshop on Treebanks and Linguistics Theories (TLT2002)*, páginas 61–77, September. disponible: http://clic.fil.ub.es/∼civit.

Civit, M., M.A. Martí, B.Navarro, N. Bufí, B. Fernández, y R. Marcos. 2003. Issues in the Syntactic Annotation of Cast3LB. En *Proceedings of the LINC03 Workshop*, Budapest. disponible: http://clic.fil.ub.es/∼civit.

Civit, M., M.A. Martí, y L. Padró. 2003. Using hybrid probabilistic-linguistic knowledge to improve pos-tagging performance. En *Proceedings of the Corpus Linguistics 2003*, Lancaster, UK. disponible: http://clic.fil.ub.es/∼civit.

Cotton, S. y S. Bird. 2000. An integrated Framework for Treebanks and Multilayer Annotations. En *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*, Athens, Greece.

Davies, M. 2002. Un corpus anotado de 100.000.000 palabras del español histórico y moderno. *Procesamiento del Lenguaje Natural*, (29):21–27, Septiembre.

de Bruyne, J. 1999. Las preposiciones. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 10, páginas 657–704.

de Miguel, E. 1999. El aspecto léxico. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 46, páginas 2977–3060.

Demonte, V. 1999. El Adjetivo. Clases y usos. La posición del adjetivo en el sintagma nominal. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua*

Española. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 3, páginas 129–215.

EAGLES. 1996a. Preliminary Recommendations on Corpus Typology. EAG–TCWG–CTYP/P.

EAGLES. 1996b. Recommendations for the Morphosyntactic Annotation of Corpora. EAG–TCWG–MAC/R, Mar. disponible: http://www.ilc.pi.cnr.it/EAGLES96/browse.html.

EAGLES. 1996c. Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A common Proposal and Applications to European Languages. EAG—CLWG—MORPHSYN/R. disponible: http://www.ilc.pi.cnr.it/EAGLES96/browse.html.

EAGLES. 1997. The EAGLES Spoken Language Working Group and the 'Handbook of Standards and Resources for Spoken Language Systems. disponible: http://coral.lili.uni-bielefeld.de/ gibbon/EAGLES/elra.news.mar97/.

Eguren, L. J. 1999. Pronombres y adverbios demostrativos. Las relaciones deícticas. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 14, páginas 929–972.

Engelson, S. P. y I. Dagan. 1996. Minimizing Manual Annotation Cost in Supervised Training from Corpora. En E. Riloff S. Wermter y G. Scheler, editores, *Connectionist, Statistical an Symbolic Approaches to Learning for Natural Language Processing*, Lecture Notes in Artificial Intelligence, 1040. Springer.

Fernández, M. y A. Anula. 1995. *Sintaxis y Cognición. Introducción al conocimiento,el procesamiento y los déficits sintácticos*. Letras Universitarias. Síntesis, Madrid.

Fernández, M.J. 1999a. El nombre propio. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 2, páginas 77–128.

Fernández, O. 1999b. El pronombre personal. Formas y distribuciones. Pronombres átonos y tónicos. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, capítulo 19.

Garside, R. 1987. The CLAWS word-tagging system. En R. Garside G. Leech, y G. Sampson, editores, *The Computational Analysis of English*. Longman, capítulo 1, páginas 30–41.

Gómez, L. 1992. *Valores gramaticales de SE*. Cuadernos de Lengua española. Arco Libros, Madrid.

Gómez, L. 1999. Los verbos auxiliares. Las perífrasis de infinitivo. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 51, páginas 3323–3389.

Hajic, J. 1998. Building a Syntactically Annotated Corpus: the Prague dependency Treebank. *Issues of Valency and meaning*, páginas 106–132.

Hajic, J. y B. Hladká. 1998. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. En *Proceedings of the 36th Annual Meeting of the ACL and the 17th ICCL*, páginas 483–490, Montréal. disponible http://ufal.mff.cuni.cz/pdt/pdt_05.html.

Instituto-Cervantes. 1996. Informe sobre recursos lingüísticos para el español (II). Corpus escritos y orales disponibles y en desarrollo en España. Observatorio español de industrias de la lengua, Instituto Cervantes.

Karlsson, F., A. Voutilainen, J. Heikkilä, y A. Anttila. 1995. *Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin / New York.

Kermes, H. y S. Evert. 2003. Text analysis meets corpus linguistics. En *Proceedings of the Corpus Linguistics 2003*, Lancaster, UK.

Kovacci, O. 1999. El adverbio. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 11, páginas 705–786.

Kurohashi, S. y M.Nagao. 1998. Building a japanese parsed corpus while improving the parsing system. disponible: http://citeseer.nj.nec.com/kurohashi98building.html.

Leech, G. 1997a. Grammatical Tagging. En R. Garside G. Leech, y T. McEnery, editores, *Corpus Annotation. Linguistic Information from Computer Text Corpora*. Longman, capítulo 2, páginas 19–33.

Leech, G. 1997b. Introducing Corpus Annotation. En R. Garside G. Leech, y T. McEnery, editores, *Corpus Annotation. Linguistic Information from Computer Text Corpora*. Longman, capítulo 1, páginas 1–18.

Leech, G., R. Barnett, y P. Kahrel. 1996. Recommendations for the Syntactic Annotation of Corpora, March. disponible: http://www.ilc.cnr.it/EAGLES96/browse.html.

Leonetti, M. 1999. El artículo. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 12, páginas 787–890.

Llisterri, J. 1997. Transcripción, etiquetado y codificación de corpus orales. Seminario de Industrias de la Lengua, Soria. disponible: http://liceu.uab.es/ joaquim/publicacions/FDS97.html.

Marciniak, M., A. Mykowiecka, A. Przepiórkowski, y A. Kupsc. 2003. Construction of an HPSG Treebank for Polish. En A. Abeillé, editor, *Building and Using syntactically annotated corpora*, Language and Speech. Kluwer, Dordrecht. disponible: http://treebank.linguist.jussieu.fr/toc.html.

Marcos, F.A. 1999. Los cuantificadores: los numerales. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 18, páginas 1189–1208.

Marcus, M., G. Kim, M.A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, y B. Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. En *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, New Jersey, March.

Marcus, M., B. Santorini, y M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*. disponible: http://www.cis.upenn.edu/ treebank/home.html.

Màrquez, L. 1999. *Part-of-Speech Tagging: A Machine-Learning Approach based on Decision Trees*. Phd. thesis, Software Department (LSI). Technical University of Catalonia (UPC).

Màrquez, L. y L. Padró. 1997. A Flexible POS Tagger Using an Automatically Acquired Language Model. En *Proceedings of the joint EACL/ACL'97*, páginas 238–245, Madrid.

Marsá, F. 1984. *Cuestiones de sintaxis española*. Ariel.

Martínez, J. A. 1999. La concordancia. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 42, páginas 2695–2786.

Martín, A. 1999. *Una propuesta de codificación morfosintáctica para corpus de referencia en lengua española*, volumen 3. Estudios de Lingüística Española (ELiEs). disponible: http://elies.rediris.es/elies3/.

Matínez, R. 1999. *Alineación automática de corpus paralelos: uns propuesta metodológica y su aplicación a un dominio de especificidad*. Ph.D. tesis, Universidad de Deusto.

McEnery, T. y A. Wilson. 1996a. *Corpus Linguistics*. Edinburgh University Press. Introductory course on corpus linguistic, based on the book, disponible: http://www.ling.lanc.ac.uk/monkey/ihe/linguistics/contents.htm.

McEnery, T. y A. Wilson. 1996b. *Corpus Linguistics*. Edinburgh University Press, 2d, 2001 edición.

Montemagni, S., F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M.T. Pazienza, D.Saracino, F. Zanzotto, N. Mana, F. Pianesi, y R. Delmonte. 2003. Building the Italian

Syntactic-Semantic Treebank. En A. Abeillé, editor, *Building and Using syntactically annotated corpora*, Language and Speech. Kluwer, Dordrecht. disponible: http://treebank.linguist.jussieu.fr/toc.html.

Moreno, A., R. Grishman, S. López, F. Sánchez, y S. Sekine. 2000. A Treebank of Spanish and its Application to Parsing. En *Proccedings of the Second Conference on Language Resources and Evaluation (LREC2000)*, páginas 107–111, Athens, Greece.

Moreno, A., S. López, y F. Sánchez. 1999. Spanish Tree Bank: Specifications (Version 5). Informe técnico, Laboratorio de Lingüística Informática (UAM).

Moreno, A. y S. López. 1999. Developing a Spanish TreeBank. Journées Atala, Corpus annotés pour la syntaxe, Paris, June. disponible: http://talana.linguist.jussieu.fr/treebanks99/.

Moreno, A., S. López, F. Sánchez, y R. Grishman. 2003. Developing a Spanish Treebank. En A. Abeillé, editor, *Building and Using syntactically annotated corpora*, Language and Speech. Kluwer, Dordrecht. disponible: http://treebank.linguist.jussieu.fr/toc.html.

Moreno, J.C. 2000. *Curso Universitario de Lingüística general*, volumen 1. Síntesis, 2a. edición.

Màrquez, L., L. Padró, y H. Rodríguez. 2001. *Mètodes robustos en l'anàlisi del llenguatge. El processament de text no restringit.* Universitat Oberta de Catalunya, UOC.

Navarro, B., M. Civit, M.A. Martí, B. Fernández, y R. Marcos. 2003. Syntactic, semantic and pragmatic annotation in Cast3LB. En *Proceedings of the Corpus Linguistics*, Lancaster. disponible: http://clic.fil.ub.es/∼civit.

Oflazer, K., B. Say, D.Z. Hakkani-Tür, y G. Tür. 2003. Building a Turkish Treebank. En A. Abeillé, editor, *Building and Using syntactically annotated corpora*, Language and Speech. Kluwer, Dordrecht. disponible: http://treebank.linguist.jussieu.fr/toc.html.

Ooi, Vincent B. Y. 1998. *Computer Corpus Lexicography.* Edimburgh University Press.

Padró, L. 1996a. A Constraint Satisfaction Alternative for POS Tagging. En *Proceedings of NLP+AI/TAL+AI*, Université de Moncton, New Bronswick, Canada. disponible: http://www.lsi.upc.es/p̃adro/.

Padró, L. 1996b. POS Tagging Using Relaxation Labelling. En *Proceedings of the COLING96*, Copenhagen, Denmark. disponible: http://www.lsi.upc.es/p̃adro/.

Padró, L. 1998. *A Hybrib Environment for Syntax-Semantic Tagging.* Ph.D. tesis, Software Department (LSI). Technical University of Catalonia (UPC).

Pavón, M. V. 1999. Clases de partículas: preposición, conjunción y adverbio. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española.* Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 9, páginas 565–655.

Payrató, L. 1996. Transcripció del discurs col·loquial. En L. Payrató E. Boix M.R. Lloret, y M. Lorente (Eds.), editores, *Corpus, Corpora. Actes del 1r i 2on col·loquis lingüístics de la Universitat de Barcelona*. PPU, Barcelona.

Payrató, L. y N. Alturo. 2002. *Corpus Oral de conversa col·loquial. Materials de treball*. Publicacions de la Universitat de Barcelona.

Picallo, M. C. y G. Rigau. 1999. El posesivo y las relaciones posesivas. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 15, páginas 973–1023.

RAE. 1973. *Esbozo de una nueva gramática de la lengua española*. RAE.

RAE. 2000. *Ortografía de la lengua española*. RAE.

Ridruejo, E. 1999. Modo y Modalidad. El modo en las subordinadas sustantivas. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 49, páginas 3209–3251.

Rigau, G. 1999. La estructura del sintagma nominal. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 5, páginas 311–362.

Rocio, V., M. Alves, G. Lopes, F. Xavier, y G. Vicente. 2003. Automated creation of a partial treebank of medieval Portuguese. En A. Abeillé, editor, *Building and Using syntactically annotated corpora*, Language and Speech. Kluwer, Dordrecht.

Rojo, G., 2001. *Lingüística con corpus*, capítulo La explotación de la Base de Datos del español actual (BDS). Gramática Española: Enseñanza e Investigación. Universidad de Salamanca.

Rojo, G. y A. Veiga. 1999. El tiempo verbal. Los tiempos simples. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 44, páginas 2867–2934.

Sampson, G. 1995. *English for the Computer. The SUSANNE corpus and Analytic Scheme*. Clarendon Press, Oxford.

Sampson, G. 2001. *Empirical Linguistics*. Continuum, London and New York.

Santalla, P. 2000. *An AGFL Formal Grammar for Phrase Level Analysis in Spanish*. Ph.D. tesis, Universidade de Santiago de Compostela.

Santamarta, L., N. Lindberg, y B. Gambäck. 1995. Towards building a Swedish Treebank. En *Proceedings of the 10th Nordic Conference of Computational Linguistics*. disponible: http://nodali.sics.se/NoDaLiDa/1995_hki.html.

Santorini, B., 1990. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project.* disponible http://cis.upenn.edu/ treebank/.

Sebastián, N., M.A. Martí, M.F. Carreiras, y F. Cuetos. 2000. *LEXESP: Léxico Informatizado del Español.* Edicions de la Universitat de Barcelona.

Sima'an, K., A. Itai, Y. Winter, A. Altman, y N.Nativ. 2001. Building a Treebank of Modern Hebrew Text. En *Traitement Automatique des Langues*, volumen 42. disponible: http://citeseer.nj.nec.com/499058.html.

Simov, K., P. Osenova, M. Slavcheva, S. Kolkovska, E. Balabanova, D. Doikoff, K. Ivanova, A. Simov, y M. Kouylekov. 2002. Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. En *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC02)*, páginas 1729–1736, Las Palmas de Gran Canaria, Spain, May.

Sánchez, C. 1999. Los cuantificadores: clases de cuantificadores y estructuras cuantificativas. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española.* Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 16, páginas 1025–1128.

Sánchez, F., J. Porta, J.L. Sancho, A.Nieto, A. Ballester, A. Fernández, J. Gómez, L. Gómez, E. Raigal, y R. Ruiz. 1999. La anotación de los corpus CREA y CORDE. *Procesamiento de Lenguaje Natural*, (25):175–182, Septiembre.

Svartvik, J. y R. Quirk. 1980. *A Corpus of English Conversation.* Lund Studies in English 56. Lund: Liber/Gleerups.

Taylor, A., M. Marcus, y B. Santorini. 2003. The Penn Treebank: an overview. En A. Abeillé, editor, *Building and Using syntactically annotated corpora*, Language and Speech. Kluwer, Dordrecht. disponible: http://treebank.linguist.jussieu.fr/toc.html.

Turmo, J. 2002. *An Information Extraction System Portable to New Domains.* Ph.D. tesis, Departament de Llenguatges i Sistemes Informàtics de la Universitat Politècnica de Catalunya .

Véronis, J. 2000. Sense Tagging: don't look for the meaning but for the use. En *Computational Lexicography and Multimedia Dictionaries, COMLEX*, páginas 1–9, Kato Achia, Greece. disponible: http://www.up.univ-mrs.fr/∼veronis/.

Véronis, J. 2001a. Annotation automatique de corpus: panorama et état de la technique. En J.M. Pierrel, editor, *Ingénierie des langues.* Éditions Hermès, Paris, capítulo 4. disponible: http://www.up.univ-mrs.fr/∼veronis/.

Véronis, J. 2001b. Sense Tagging: does it make sense? En *The Corpus Linguistics 2001 Conference*, Lancaster, U.K. disponible: http://www.up.univ-mrs.fr/∼veronis/.

Yablonsky, S.A. 2000. Russian Monitor Corpora: Composition, Linguistic Encoding and Internet . En *Proceedings of the Second International Conference on Language and Evaluation LREC-2000*, Athens, Greece.

Yllera, A. 1999. Las perífrasis verbales de gerundio y participio. En I. Bosque y V. Demonte, editores, *Gramática Descriptiva de la Lengua Española*. Espasa-Calpe, Real Academia Española de la Lengua, Madrid, capítulo 52, páginas 3391–3441.

Apéndices

Apéndice A

Locuciones

Presentamos en este apéndice, expresiones que hemos tratado como locuciones y expresiones que hemos tratado por separado.

A.1. Locuciones conjuntivas subordinantes

(A.1) *... yo solita [...] soy un teatro lírico.* **Así_que_cs**, *por caridad, ¿es que no hay disponible...* (c1).

... acaba de aprobar una reforma [...] **para_que_cs** *los padres puedan registrar...* (c1).

los más suaves propinan un patadón a la papelera, **a_la_vez_que_cs** *sueltan un aullido* (c2).

En_tanto_que_cs *culé, yo debiera desear que esa pareja perdurara años...* (d1).

...y, **como_quiera_que_cs** *las precipitaciones se están reduciendo paulatinamente...* (dc1).

... ella puede producirle otra nidada rápidamente, **en_el_caso_de_que_cs** *un depredador destruya la primera* (dc1).

A.2. Locuciones conjuntivas coordinantes

(A.2) *Cuando no los vemos* **sino_que_cc** *oímos en nuestro interior su sordo zumbido de bacterias inquietas, los médicos hablan de tinnitus* (a12).

A.3. Locuciones adverbiales

(A.3) *... como el Mendicuti anda* **en_vilo_rg**... (c1).

De la Historia **en_general_rg** *y de la Historia de cada país* **en_particular_rg**. **Por_ejemplo_rg**: *el próximo nieto de ...* (c1).

Y **desde_luego_rg** *alguna chiquilla vistosa se tendría que llamar...* (c1).

Parecerán medio comanches, pero, **hoy_por_hoy_rg**, *la Historia de este país...* (c1).

Una señora **como_es_debido_rg** *diría que,* **más_o_menos_rg**, *los tres son...* (c2).

Claro que cuando Felipe_González (que **a_lo_que_parece_rg** *no ha perdido* **del_todo_rg** *el carisma aquel que tuvo)...* (c2).

ella misma terminó declarándose escritora **a_secas_rg** (c2).

en las ocasiones en que caí (entonces **por_casualidad_rg***) en casa de ...* (c2).
Su atracción por los precipicios ha concluido superponiéndose o, **al_menos_rg***, equilibrándose con...* (c2).
En el Barça, que juega al toque, a veces mal y otras [...] **de_maravilla_rg***, ...* (d1).
el temor a que, **de_repente_rg***, le salga bien* (d1).
A_la_larga_rg *ella puede incluso mejorar* (dc1).
en función de una larga serie de atribuciones que desempeña **a_la_perfección_rg** (dc1).
la condensación de ingredientes que **en_principio_rg** *son verdaderos* (a10).
No me extraña que el martillo, el yunque y el estribo puedan **de_golpe_rg** *ponerse a armar follón sólo porque sí, porque les viene en gana* (a12).
Iba arrastrando un poco los pies porque aún no se había repuesto **del_todo_rg** *del último ataque* (a14).

A.4. Locuciones preposicionales

(A.4) *Los neonatos tienen que tener la oportunidad de convertirse,* **gracias_a_sps00** *sus nombres ...* (c1).
un periódico deportivo se lamentaba ayer de ello, en futbolístico reflejo, como si las múltiples repeticiones de los duelos Detroit-Chicago o Boston-Nueva_ York **a_lo_largo_de_sps00** *toda la temporada no fueran desde siempre la sal de la temporada normal* (d1).
la relación de hembras disponibles sexualmente **con_respecto_a_sps00** *los machos puede llegar a ser de 7 a 1* (dc1).
Porque, **pese_a_sps00** *sus variadas formas, todos estos animales tienen su piel* (dc1).
A_pesar_de_sps00 *su precocidad, ...* (dc1).
y **al_margen_del_spcms** *polvo acumulado sobre todas las superficies, las habitaciones casi vacías...* (t6).

A.5. Unidades léxicas complejas

Hay estructuras que pueden considerarse locuciones y que aparecen así tratadas en la bibliografía y que aquí no hemos tratado de este modo. Algunas estás formadas por la secuencia <adverbio + preposición> y creemos que pueden tratarse por separado. Hay adverbios que se consideran transitivos, que suelen aparecer seguidos de la preposición *de* porque su significado es relativo[1] como *después, antes, cerca, lejos,* etc.

(A.5) *a tu padre,* **antes de** *mandaros al hospicio, le gustaba marcaros* (t2).
pues el solo pasar **cerca de** *un hombre podía mancillarles la honra* (t1).
la víctima logró escurrirse de **debajo de** *la asesina fumadora* (t5).
todo esto se podría pasar a la pantalla **después de** *haberlo percibido* (a10).
En el Tercer_ Mundo, por_ el_ contrario, seguro que no andan muy **lejos del** *famoso lema de los mosqueteros: una rata para todos y todos para una* (a12).

[1] Frente a otros adverbio con significado absoluto como *ayer, ahora, aquí.*

Esta vez queremos que a todo el mundo le quede claro que queremos gobernar en Andalucía y que el 12_de_junio por la noche pondremos nuestro programa **encima de** *la mesa para llegar a un acuerdo con los socialistas y gobernar* (r2).

Por otra parte, algunos de estos casos admiten que el término de la preposición sea una completiva:

(A.6) *pero eso era* **antes de que** *ingresáramos en el reino de a alta tecnología* (a12).
Casi un cuarto de siglo **después de que** *Buñuel sometiera la -Tristana- de Galdós al tormento -y también al hallazgo- de su personalísima versión cinematográfica, vuelve a representarse -esta vez en la escena- lo que Don Benito concibió como pura materia novelable* (a24).

Otra de las secuencias que pueden tratarse como locuciones y que aquí hemos analizado por separado es la formada por <preposición + *que*>, como en los ejemplos siguientes:

(A.7) *No levantó los ojos hacia el espejo* **hasta que** *adquirió la certeza de que...* (t6).
él recordaba que le habían ascendido dos veces **desde que** *estaban juntos* (t6).

Por último, también tratamos por separado expresiones temporales como:

(A.8) **hace años que** *no veo a Rosa* (c2).
Y **alguna vez** *me he preguntado si...* (a12).
los ruidos que los oídos de ese buen hombre recrean para él **cada vez que** *intenta conciliar el sueño* (a12).

Apéndice B

GramEsp

B.1. Control del formato de salida

Nodos de análisis incluidos en las listas para el control de la salida del analizador
@HIDEN a-ms a-fs a-mp a-fp adv
@HIDEN pron-ms pron-fs pron-mp pron-fp adv-interrog pposs-ms pposs-fs
@HIDEN pposs-mp pposs-fp cuyo-ms cuyo-fs cuyo-mp cuyo-fp cual-s cual-p
@HIDEN quien-s quien-p prel prel-adv
@HIDEN j-ms j-fs j-mp j-fp
@HIDEN grup-complex-spec-ms grup-complex-spec-fs grup-complex-spec-mp
@HIDEN grup-complex-spec-fp pos-ms pos-fs pos-mp pos-fp exc-ms exc-fs
@HIDEN exc-mp exc-fp vser vaux parti parti-ms parti-fs parti-mp parti-fp
@HIDEN grup-c-nom-fp grup-c-nom-mp
@HIDEN cuantif dem-fp dem-fs dem-mp pron
@HIDEN dem-ms ger-pas geraux geraux-ser
@HIDEN indef-fp indef-fs indef-mp indef-ms
@HIDEN inf inf-pas infaux infaux-ser int-fp int-fs int-mp int-ms
@HIDEN n-fp n-fs n-mp n-ms pinterrog-s pinterrog-p
@HIDEN num-fp num-fs num-mp num-ms parti-aux parti-ser
@HIDEN parti-flex paton-p paton-s pdem-fp pdem-fs pdem-mp pdem-ms
@HIDEN pindef-fp pindef-fs pindef-mp pindef-ms pinterrog
@HIDEN pinterrog-fp pinterrog-fs pinterrog-mp pinterrog-ms
@HIDEN verb verb-pass vser w-fp w-fs w-mp w-ms
@HIDEN prepc-ms psubj-fp psubj-fs psubj-mp psubj-ms ptonic
@HIDEN paton-fp paton-fs paton-mp paton-ms ger psubj-s
@HIDEN prel-fp prel-fs prel-mp prel-ms morf-pron

@FLAT grup-nom-ms grup-nom-fs grup-nom-mp grup-nom-fp

@GROUP pnum-ms pnum-fs pnum-mp pnum-fp
@GROUP num-ms num-fs num-mp num-fp

@GROUP data sa

@NOTOP

Prioridad en la aplicación de las reglas de anáisis:

@PRIOR grup-verb
@PRIOR s-a-ms s-a-fs s-a-mp s-a-fp parti-flex
@PRIOR sn
@PRIOR verb vaux vser
@PRIOR grup-nom-ms grup-nom-fs grup-nom-mp grup-nom-fp
@PRIOR sadv
@PRIOR espec-ms espec-fs espec-mp espec-fp

B.2. Pseudoterminales

B.2.1. Adjetivos

a-ms → *aq0000.*
a-mp → *aq0000.*
a-ms → *aq0cn0.*
a-fs → *aq0cn0.*
a-mp → *aq0cn0.*
a-fp → *aq0cn0.*
a-mp → *aq0cp0.*
a-ms → *aq0cs0.*
a-fp → *aq0fp0.*
a-fs → *aq0fs0.*
a-mp → *aq0mp0.*
a-ms → *aq0ms0.*
a-fs → *aq0000.*
a-fp → *aq0000.*
a-fp → *aq0cp0.*
a-fs → *aq0cs0.*
a-ms → *aq0msp.*
a-fs → *aq0fsp.*
a-mp → *aq0mpp.*
a-fp → *aq0fpp.*
a-ms → *ao0ms0.*
a-fs → *ao0fs0.*
a-mp → *ao0mp0.*
a-fp → *ao0fp0.*

B.2.2. Nombres

n-ms → *nc00000.*
n-fs → *nc00000.*
n-mp → *nc00000.*
n-fp → *nc00000.*
n-ms → *ncms000.*
n-mp → *ncmp000.*
n-fs → *ncfs000.*
n-fp → *ncfp000.*
n-ms → *nccs000.*
n-mp → *nccp000.*
n-fs → *nccs000.*
n-fp → *nccp000.*
n-mp → *ncmn000.*
n-fp → *ncfn000.*
n-ms → *ncmn000.*
n-fs → *ncfn000.*
n-ms → *nccn000.*
n-fs → *nccn000.*
n-mp → *nccn000.*

n-fp → *nccn000*.
w-ms → *np00000*.
w-mp → *np00000*.
w-fs → *np00000*.
w-fp → *np00000*.

B.2.3. Adverbios

adv → *rg*.
neg → *rn*.

B.2.4. Preposiciones

prep → *sps00*.
prepc-ms → *spcms*.

B.2.5. Conjunciones

conj-subord → *cs*.
coord → *cc*.

B.2.6. Verbos principales

Formas indicativas:

verb → *vmip1s0*.
verb → *vmip2s0*.
verb → *vmip3s0*.
verb → *vmip1p0*.
verb → *vmip2p0*.
verb → *vmip3p0*.
verb → *vmii1s0*.
verb → *vmii2s0*.
verb → *vmii3s0*.
verb → *vmii1p0*.
verb → *vmii2p0*.
verb → *vmii3p0*.
verb → *vmis1s0*.
verb → *vmis2s0*.
verb → *vmis3s0*.
verb → *vmis1p0*.
verb → *vmis2p0*.
verb → *vmis3p0*.
verb → *vmif1s0*.
verb → *vmif2s0*.
verb → *vmif3s0*.
verb → *vmif1p0*.
verb → *vmif2p0*.
verb → *vmif3p0*.
verb → *vmic1s0*.
verb → *vmic2s0*.
verb → *vmic3s0*.
verb → *vmic1p0*.
verb → *vmic2p0*.
verb → *vmic3p0*.

Formas imperativas:

verb → *vmm02s0.*
verb → *vmm03s0.*
verb → *vmm01p0.*
verb → *vmm02p0.*
verb → *vmm03p0.*

Formas subjuntivas:

verb → *vmsp1s0.*
verb → *vmsp2s0.*
verb → *vmsp3s0.*
verb → *vmsp1p0.*
verb → *vmsp2p0.*
verb → *vmsp3p0.*
verb → *vmsi1s0.*
verb → *vmsi2s0.*
verb → *vmsi3s0.*
verb → *vmsi1p0.*
verb → *vmsi2p0.*
verb → *vmsi3p0.*
verb → *vmsf1s0.*
verb → *vmsf2s0.*
verb → *vmsf3s0.*
verb → *vmsf1p0.*
verb → *vmsf2p0.*
verb → *vmsf3p0.*

Verbo *ser* como principal:

verb → *vsip1s0.*
verb → *vsip2s0.*
verb → *vsip3s0.*
verb → *vsip1p0.*
verb → *vsip2p0.*
verb → *vsip3p0.*
verb → *vsii1s0.*
verb → *vsii2s0.*
verb → *vsii3s0.*
verb → *vsii1p0.*
verb → *vsii2p0.*
verb → *vsii3p0.*
verb → *vsis1s0.*
verb → *vsis2s0.*
verb → *vsis3s0.*
verb → *vsis1p0.*
verb → *vsis2p0.*
verb → *vsis3p0.*
verb → *vsif1s0.*
verb → *vsif2s0.*
verb → *vsif3s0.*
verb → *vsif1p0.*
verb → *vsif2p0.*
verb → *vsif3p0.*
verb → *vsic1s0.*
verb → *vsic2s0.*
verb → *vsic3s0.*
verb → *vsic1p0.*
verb → *vsic2p0.*
verb → *vsic3p0.*
verb → *vsmp2s0.*
verb → *vsmp3s0.*
verb → *vsmp1p0.*
verb → *vsmp2p0.*
verb → *vsmp3p0.*
verb → *vssp1s0.*
verb → *vssp2s0.*
verb → *vssp3s0.*
verb → *vssp1p0.*
verb → *vssp2p0.*
verb → *vssp3p0.*
verb → *vssi1s0.*
verb → *vssi1s0.*
verb → *vssi2s0.*
verb → *vssi2s0.*

verb → *vssi3s0*.
verb → *vssi3s0*.
verb → *vssi1p0*.
verb → *vssi1p0*.
verb → *vssi2p0*.
verb → *vssi2p0*.
verb → *vssi3p0*.
verb → *vssi3p0*.
verb → *vssf1s0*.
verb → *vssf2s0*.
verb → *vssf3s0*.
verb → *vssf1p0*.
verb → *vssf2p0*.
verb → *vssf3p0*.
verb → *vsm02s0*.
verb → *vsm03s0*.
verb → *vsm01p0*.
verb → *vsm02p0*.
verb → *vsm03p0*.

Verbo *haber* como principal:
verb → *vaip3s0*.
verb → *vaii3s0*.
verb → *vais3s0*.
verb → *vaif3s0*.
verb → *vaic3s0*.
verb → *vam03s0*.
verb → *vasp3s0*.
verb → *vasi3s0*.
verb → *vasf3s0*.

B.2.7. Determinantes

espec-ms → *grup-complex-spec-ms*.
espec-mp → *grup-complex-spec-mp*.
espec-fs → *grup-complex-spec-fs*.
espec-fp → *grup-complex-spec-fp*.
espec-ms → *cuantif*.
espec-fs → *cuantif*.
espec-mp → *cuantif*.
espec-fp → *cuantif*.
espec-ms → *num-ms*.
espec-fs → *num-fs*.
espec-mp → *num-mp*.
espec-fp → *num-fp*.
espec-ms → *dem-ms*.
espec-mp → *dem-mp*.
espec-ms → *pos-ms*.
espec-mp → *pos-mp*.
espec-ms → *int-ms*.
espec-mp → *int-mp*.
espec-ms → *exc-ms*.
espec-mp → *exc-mp*.
espec-ms → *indef-ms*.
espec-mp → *indef-mp*.
espec-ms → *num-ms*.
espec-mp → *num-mp*.
espec-ms → *j-ms*.
espec-mp → *j-mp*.
espec-fs → *dem-fs*.
espec-fp → *dem-fp*.
espec-fs → *pos-fs*.
espec-fp → *pos-fp*.
espec-fs → *int-fs*.
espec-fp → *int-fp*.
espec-fs → *exc-fs*.
espec-fp → *exc-fp*.
espec-fs → *indef-fs*.
espec-fp → *indef-fp*.
espec-fs → *num-fs*.

espec-fp → num-fp.
espec-fs → j-fs.
espec-fp → j-fp.

B.2.8. Pronombres

psubj-s → pp1csn00.
psubj-s → pp2csn00.
psubj-ms → pp3ms000.
psubj-fs → pp3fs000.
psubj-mp → pp1mp000.
psubj-fp → pp1fp000.
psubj-mp → pp2mp000.
psubj-fp → pp2fp000.
psubj-mp → pp3mp000.
psubj-fp → pp3fp000.
psubj-ms → pp3ns000.
psubj-ms → pp2cs00p.
psubj-mp → pp2cp00p.
psubj-fs → pp2cs00p.
psubj-fp → pp2cp00p.

ptonic → pp1cso00(mí).
ptonic → pp2cso00(ti).
ptonic → pp3cno00(sí).

pdem-ms → pd0ms000.
pdem-fs → pd0fs000.
pdem-ms → pd0ns000.
pdem-mp → pd0mp000.
pdem-fp → pd0fp000.
pdem-ms → pd0cs000.
pdem-fs → pd0cs000.
pdem-mp → pd0cp000.
pdem-fp → pd0cp000.

adv-interrog → pt000000.
pinterrog-ms → pt0ms000.
pinterrog-fs → pt0fs000.
pinterrog-mp → pt0mp000.
pinterrog-fp → pt0fp000.
pinterrog-s → pt0cs000.
pinterrog-p → pt0cp000.
pinterrog → pt0cn000.

pposs-ms → px1ms020.
pposs-fs → px1fs0s0.
pposs-ms → px1ms0p0.
pposs-fs → px1fs0p0.
pposs-ms → px2ms0s0.
pposs-fs → px2fs0s0.
pposs-ms → px2ms0p0.
pposs-fs → px2fs0p0.
pposs-ms → px3ms000.
pposs-fs → px3fs000.

pposs-mp → px1mp0s0.
pposs-fp → px1fp0s0.
pposs-mp → px1mp0p0.
pposs-fp → px1fp0p0.
pposs-mp → px2mp0s0.
pposs-fp → px2fp0s0.
pposs-mp → px2mp0p0.
pposs-fp → px2fp0p0.
pposs-mp → px3mp000.
pposs-fp → px3fp000.

cuyo-ms → pr0ms000(cuyo).
cuyo-fs → pr0fs000(cuya).
cuyo-mp → pr0mp000(cuyos).
cuyo-fp → pr0fp000(cuyas).
cual-s → pr0cs000(cual).
cual-p → pr0cp000(cuales).
quien-s → pr0cs000(quien).
quien-p → pr0cp000(quienes).
prel → pr0cn000.
prel-ms → pr0ms000(cuanto).
prel-mp → pr0mp000(cuantos).
prel-fs → pr0fs000(cuanta).
prel-fp → pr0fp000(cuantas).
prel-adv → pr000000.

pindef-ms → *pi0ms000*.
pindef-fs → *pi0fs000*.
pindef-mp → *pi0mp000*.
pindef-fp → *pi0fp000*.
pindef-fp → *pi0cp000*.
pindef-mp → *pi0cp000*.
pindef-ms → *pi0cs000*.

pnum-mp → *pn0cp000*.
pnum-fp → *pn0cp000*.
pnum-ms → *pn0cs000*.
pnum-fp → *pn0fp000*.
pnum-fs → *pn0fs000*.
pnum-mp → *pn0mp000*.
pnum-ms → *pn0ms000*.

patons → *paton-s*.
patons → *paton-p*.
patons → *paton-mp*.
patons → *paton-fp*.
patons → *paton-ms*.
patons → *paton-fs*.
patons → *paton*.
paton-s → *pp1cs000*.
paton-s → *pp2cs000*.
patons → *pp3cn000*.
morf-pron → *p0300000*.
morf-pron → *p010s000*.
morf-pron → *p010p000*.
morf-pron → *p020s000*.
morf-pron → *p020p000*.
morfema-verbal → *p0000000*.
paton-s → *pp3csd00*.
paton-s → *pp3csa00*.
paton-p → *pp1cp000*.
paton-p → *pp2cp000*.
paton-mp → *pp3mpa00*.
paton-fp → *pp3fpa00*.
paton-fs → *pp3fsa00*.
paton-ms → *pp3msa00*.
paton → *pp3cna00*.
paton-p → *pp3cpd00*.
paton-p → *pp3cpa00*.

pron-ms → *pinterrog-s*.
pron-fs → *pinterrog-s*.
pron-mp → *pinterrog-p*.
pron-fp → *pinterrog-p*.
pron-ms → *pinterrog*.
pron-fs → *pinterrog*.
pron-ms → *psubj-ms*.
pron-fs → *psubj-fs*.
pron-mp → *psubj-mp*.
pron-fp → *psubj-fp*.
pron-ms → *pdem-ms*.
pron-fs → *pdem-mp*.
pron-mp → *pdem-fs*.
pron-fp → *pdem-fp*.
pron-ms → *pinterrog-ms*.
pron-fs → *pinterrog-fs*.
pron-mp → *pinterrog-mp*.
pron-fp → *pinterrog-fp*.
pron-ms → *pposs-ms*.
pron-fs → *pposs-fs*.
pron-mp → *pposs-mp*.
pron-fp → *pposs-fp*.
pron-ms → *pindef-ms*.
pron-fs → *pindef-mp*.
pron-mp → *pindef-fs*.
pron-fp → *pindef-fp*.

B.2.9. Verbos auxiliares y semiauxiliares

vaux → *vaip1s0*.
vaux → *vaip2s0*.
vaux → *vaip3s0*.
vaux → *vaip1p0*.
vaux → *vaip2p0*.
vaux → *vaip3p0*.
vaux → *vaii1s0*.
vaux → *vaii2s0*.
vaux → *vaii3s0*.

vaux → *vaii1p0*.
vaux → *vaii2p0*.
vaux → *vaii3p0*.
vaux → *vais1s0*.
vaux → *vais2s0*.
vaux → *vais3s0*.
vaux → *vais1p0*.
vaux → *vais2p0*.
vaux → *vais3p0*.
vaux → *vaif1s0*.
vaux → *vaif2s0*.
vaux → *vaif3s0*.
vaux → *vaif1p0*.
vaux → *vaif2p0*.
vaux → *vaif3p0*.
vaux → *vaic1s0*.
vaux → *vaic2s0*.
vaux → *vaic3s0*.
vaux → *vaic1p0*.
vaux → *vaic2p0*.
vaux → *vaic3p0*.
vaux → *vam02s0*.
vaux → *vam03s0*.
vaux → *vam01p0*.
vaux → *vam02p0*.
vaux → *vam03p0*.

vaux → *vasp1s0*.
vaux → *vasp2s0*.
vaux → *vasp3s0*.
vaux → *vasp1p0*.
vaux → *vasp2p0*.
vaux → *vasp3p0*.
vaux → *vasi1s0*.
vaux → *vasi2s0*.
vaux → *vasi3s0*.
vaux → *vasi1p0*.
vaux → *vasi2p0*.
vaux → *vasi3p0*.
vaux → *vasi1s0*.
vaux → *vasi2s0*.
vaux → *vasi3s0*.
vaux → *vasi1p0*.
vaux → *vasi2p0*.
vaux → *vasi3p0*.
vaux → *vasf1s0*.
vaux → *vasf2s0*.
vaux → *vasf3s0*.
vaux → *vasf1p0*.
vaux → *vasf2p0*.
vaux → *vasf3p0*.

vser → *vsip1s0*.
vser → *vsip2s0*.
vser → *vsip3s0*.
vser → *vsip1p0*.
vser → *vsip2p0*.
vser → *vsip3p0*.
vser → *vsii1s0*.
vser → *vsii2s0*.
vser → *vsii3s0*.
vser → *vsii1p0*.
vser → *vsii2p0*.
vser → *vsii3p0*.
vser → *vsis1s0*.
vser → *vsis2s0*.
vser → *vsis3s0*.
vser → *vsis1p0*.

vser → *vsis2p0*.
vser → *vsis3p0*.
vser → *vsif1s0*.
vser → *vsif2s0*.
vser → *vsif3s0*.
vser → *vsif1p0*.
vser → *vsif2p0*.
vser → *vsif3p0*.
vser → *vsic1s0*.
vser → *vsic2s0*.
vser → *vsic3s0*.
vser → *vsic1p0*.
vser → *vsic2p0*.
vser → *vsic3p0*.
vser → *vsmp2s0*.
vser → *vsmp3s0*.
vser → *vsmp1p0*.

vser → *vsmp2p0.*
vser → *vsmp3p0.*
vser → *vssp1s0.*
vser → *vssp2s0.*
vser → *vssp3s0.*
vser → *vssp1p0.*
vser → *vssp2p0.*
vser → *vssp3p0.*
vser → *vssi1s0.*
vser → *vssi1s0.*
vser → *vssi2s0.*
vser → *vssi2s0.*
vser → *vssi3s0.*
vser → *vssi3s0.*
vser → *vssi1p0.*
vser → *vssi1p0.*
vser → *vssi2p0.*
vser → *vssi2p0.*
vser → *vssi3p0.*
vser → *vssi3p0.*
vser → *vssf1s0.*
vser → *vssf2s0.*
vser → *vssf3s0.*
vser → *vssf1p0.*
vser → *vssf2p0.*
vser → *vssf3p0.*
vser → *vsm02s0.*
vser → *vsm03s0.*
vser → *vsm01p0.*
vser → *vsm02p0.*
vser → *vsm03p0.*

B.2.10. Formas no personales del verbo

Participios:
parti → *vmp00sm.*
parti-ms → *vmp00sm.*
parti-mp → *vmp00pm.*
parti-fs → *vmp00sf.*
parti-fp → *vmp00pf.*
parti-aux → *vap00sm.*
parti-ser → *vsp00sm.*

Gerundios:
ger → *vmg0000.* %comiendo amando
ger → *vag0000.* %siendo
geraux → *vag0000.*
geraux-ser → *vsg0000.*

Infinitivos:
infaux-ser → *vsn0000.*
infaux → *van0000.*
inf → *van0000.* %haber
inf → *vsn0000.* %ser
inf → *vmn0000.* %comer amar

B.3. Reglas de sintagma nominal

B.3.1. Grup-nom

grup-nom-ms → *n-ms.*
grup-nom-mp → *n-mp.*
grup-nom-fs → *n-fs.*
grup-nom-fp → *n-fp.*

grup-nom-ms → *n-ms, n-fs.* %gas mostaza
grup-nom-mp → *n-mp, n-fs.*
grup-nom-fs → *n-fs, n-fs.* %la palabra cultura
grup-nom-fs → *n-fs, n-ms.*

grup-nom-ms → *n-ms, sp-de.*
grup-nom-mp → *n-mp, sp-de.*
grup-nom-fs → *n-fs, sp-de.*
grup-nom-fp → *n-fp, sp-de.*

grup-nom-ms → *pnum-ms, sp-de.* % cientos de miles de personas
grup-nom-mp → *pnum-mp, sp-de.*
grup-nom-fs → *pnum-fs, sp-de.*
grup-nom-fp → *pnum-fp, sp-de.*

grup-nom-mp → *w-mp.*
grup-nom-ms → *w-ms.*
grup-nom-fp → *w-fp.*
grup-nom-fs → *w-fs.*
grup-nom → *w-ms.*
grup-nom → *w-fs.*
grup-nom → *w-mp.*
grup-nom → *w-fp.*

grup-nom-ms → *w-ms, w-ms.*
grup-nom-ms → *n-ms, w-ms.* %El presidente Chirac
grup-nom-ms → *n-ms, s-a-ms.*
grup-nom-mp → *n-mp, s-a-mp.*
grup-nom-ms → *w-ms, s-a-ms.*
grup-nom-mp → *w-mp, s-a-mp.*
grup-nom-ms → *s-a-ms, grup-nom-ms.*
grup-nom-mp → *s-a-mp, grup-nom-mp.*

grup-nom-fs → *w-fs, w-fs.*

grup-nom-fs → *n-fs, w-fs.*
grup-nom-fs → *n-fs, s-a-fs.*
grup-nom-fp → *n-fp, s-a-fp.*
grup-nom-fs → *w-fs, s-a-fs.*
grup-nom-fp → *w-fp, s-a-fp.*
grup-nom-fs → *s-a-fs, grup-nom-fs.*
grup-nom-fp → *s-a-fp, grup-nom-fp.*

grup-nom-mp → *grup-c-nom-mp.*
grup-nom-fp → *grup-c-nom-fp.*

grup-nom-mp → *grup-c-nom-mp.*
grup-nom-fp → *grup-c-nom-fp.*

grup-nom-mp → *grup-c-nom-mp, s-a-mp.*
grup-nom-fp → *grup-c-nom-fp, s-a-fp.*

grup-nom-fp → *pnum-fp, pnum-fp.* %mil seiscientas
grup-nom-mp → *pnum-mp, pnum-mp.*
grup-nom-fp → *pnum-fp, pnum-fp, pnum-fp.* %mil seiscientas setenta
grup-nom-mp → *pnum-mp, pnum-mp, pnum-mp.*
grup-nom-fp → *pnum-fp, pnum-fp, coord, pnum-fp.*
grup-nom-mp → *pnum-mp, pnum-mp, coord, pnum-mp.*
grup-nom-fp → *pnum-fp, pnum-fp, pnum-fp, coord, pnum-fp.*
grup-nom-mp → *pnum-mp, pnum-mp, pnum-mp, coord, pnum-mp.*

grup-nom-ms → *n-ms, pos-ms.* %amigo suyo
grup-nom-mp → *n-mp, pos-mp.*
grup-nom-fs → *n-fs, pos-fs.*
grup-nom-fp → *n-fp, pos-fp.*

grup-nom-ms → *n-ms, pdem-ms.* %el chico ése
grup-nom-mp → *n-mp, pdem-mp.*
grup-nom-fs → *n-fs, pdem-fs.*
grup-nom-fp → *n-fp, pdem-fp.*

B.3.1.1. Coordinación léxica de nombres

grup-c-nom-mp → *n-mp, coord, n-fs.* %cartones y lata
grup-c-nom-mp → *n-mp, coord, n-fp.* %cartones y latas
grup-c-nom-mp → *n-mp, coord, n-ms.* %cartones y latón
grup-c-nom-mp → *n-mp, coord, n-mp.* %cartones y periódicos

grup-c-nom-mp → *n-ms, coord, n-fs.* %cartón y lata
grup-c-nom-mp → *n-ms, coord, n-fp.* %cartón y latas
grup-c-nom-mp → *n-ms, coord, n-ms.* %cartón y latón
grup-c-nom-mp → *n-ms, coord, n-mp.* %cartón y periódicos

grup-c-nom-mp → *n-fs, coord, n-mp.* %lata y cartones
grup-c-nom-mp → *n-fp, coord, n-mp.* %latas y cartones
grup-c-nom-mp → *n-fs, coord, n-ms.* %lata y cartón
grup-c-nom-mp → *n-fp, coord, n-ms.* %latas y cartón

grup-c-nom-fp → *n-fs, coord, n-fs.* %lata y caja
grup-c-nom-fp → *n-fs, coord, n-fp.* %lata y cajas
grup-c-nom-fp → *n-fp, coord, n-fs.* %cajas y lata
grup-c-nom-fp → *n-fp, coord, n-fp.* %cajas y latas

%coordinación en forma: nom, Fc, n-coord
grup-c-nom-mp → *n-ms, Fc, grup-c-nom-mp.*
grup-c-nom-mp → *n-fs, Fc, grup-c-nom-mp.*
grup-c-nom-mp → *n-mp, Fc, grup-c-nom-mp.*
grup-c-nom-mp → *n-fp, Fc, grup-c-nom-mp.*

grup-c-nom-fp → *n-fs, Fc, grup-c-nom-fp.*
grup-c-nom-fp → *n-fp, Fc, grup-c-nom-fp.*

B.3.1.2. Combinaciones de especificadores

grup-complex-spec-ms → *indef-ms,num-ms.* %otros doce meses
grup-complex-spec-mp → *indef-mp,num-mp.*
grup-complex-spec-ms → *numer-ms,num-ms.*
grup-complex-spec-mp → *numer-mp,num-mp.*
grup-complex-spec-mp → *num-mp, coord, num-mp.* %dos o tres libros
grup-complex-spec-fp → *num-fp, coord, num-fp.* %dos o tres cosas
grup-complex-spec-mp → *num-ms, coord, num-mp.* %uno o dos
grup-complex-spec-fp → *num-fs, coord, num-fp.* %una o dos
grup-complex-spec-ms → *indef-ms, coord, indef-ms.* %uno y otro
grup-complex-spec-fs → *indef-fs, coord, indef-fs.* %una y otra
grup-complex-spec-mp → *indef-mp, coord, indef-mp.* %unos y otros
grup-complex-spec-fp → *indef-fp, coord, indef-fp.* %unas y otras
grup-complex-spec-ms → *indef-ms,dem-ms.* %todo este
grup-complex-spec-mp → *indef-mp,dem-mp.*
grup-complex-spec-ms → *indef-ms,j-ms.* %todo el
grup-complex-spec-mp → *indef-mp,j-mp.*

grup-complex-spec-ms → *cuantif(casi), indef-ms,j-ms.* %casi todo el
grup-complex-spec-mp → *cuantif(casi), indef-mp,j-mp.*
grup-complex-spec-ms → *indef-ms,indef-ms.*
grup-complex-spec-mp → *indef-mp,indef-mp.* %algunos pocos
grup-complex-spec-ms → *indef-ms,pos-ms.*
grup-complex-spec-mp → *indef-mp,pos-mp.*
grup-complex-spec-ms → *dem-ms, num-ms.* %estos tres
grup-complex-spec-mp → *dem-mp, num-mp.*
grup-complex-spec-ms → *pos-ms, num-ms.* %tus tres
grup-complex-spec-mp → *pos-mp, num-mp.*
grup-complex-spec-ms → *j-ms, num-ms.* %los dos
grup-complex-spec-mp → *j-mp, num-mp.*
grup-complex-spec-ms → *j-ms, indef-ms, indef-ms.*
grup-complex-spec-mp → *j-mp, indef-mp, indef-mp.*
grup-complex-spec-ms → *j-ms, indef-ms.*
grup-complex-spec-mp → *j-mp, indef-mp.*

grup-complex-spec-fs → *indef-fs,num-fs.*
grup-complex-spec-fp → *indef-fp,num-fp.*
grup-complex-spec-fs → *num-fs,num-fs.*
grup-complex-spec-fp → *num-fp,num-fp.*
grup-complex-spec-fs → *indef-fs,dem-fs.*
grup-complex-spec-fp → *indef-fp,dem-fp.*
grup-complex-spec-fs → *indef-fs,j-fs.*
grup-complex-spec-fp → *indef-fp,j-fp.*
grup-complex-spec-fs → *cuantif(casi), indef-fs,j-fs.*
grup-complex-spec-fp → *cuantif(casi), indef-fp,j-fp.*
grup-complex-spec-fs → *indef-fs,indef-fs.*
grup-complex-spec-fp → *indef-fp,indef-fp.*
grup-complex-spec-fs → *indef-fs,pos-fs.*
grup-complex-spec-fp → *indef-fp,pos-fp.*
grup-complex-spec-fs → *dem-fs, num-fs.*
grup-complex-spec-fp → *dem-fp, num-fp.*
grup-complex-spec-fs → *pos-fs, num-fs.*
grup-complex-spec-fp → *pos-fp, num-fp.*
grup-complex-spec-fs → *j-fs, num-fs.*
grup-complex-spec-fp → *j-fp, num-fp.*
grup-complex-spec-fs → *j-fs, indef-fs, indef-fs.*
grup-complex-spec-fp → *j-fp, indef-fp, indef-fp.* %las bastantes pocas
grup-complex-spec-fs → *j-fs, indef-fs.*
grup-complex-spec-fp → *j-fp, indef-fp.*

grup-complex-spec-fs → *j-fs, num-fs.*
grup-complex-spec-fp → *j-fp, num-fp.*

grup-complex-spec-ms → *j-ms, num-ms.*
grup-complex-spec-mp → *j-mp, num-mp.*

grup-complex-spec-fs → *pos-fs, num-fs.*
grup-complex-spec-fp → *pos-fp, num-fp.*
grup-complex-spec-ms → *pos-ms, num-ms.*
grup-complex-spec-mp → *pos-mp, num-mp.*

grup-complex-spec-fs → *pos-fs, indef-fs.*
grup-complex-spec-fp → *pos-fp, indef-fp.*
grup-complex-spec-ms → *pos-ms, indef-ms.*
grup-complex-spec-mp → *pos-mp, indef-mp.*

grup-complex-spec-fs → *indef-fs, num-fs.*
grup-complex-spec-fp → *indef-fp, num-fp.*
grup-complex-spec-ms → *indef-ms, num-ms.*
grup-complex-spec-mp → *indef-mp, num-mp.*

grup-complex-spec-fp → *num-fp, num-fp.* %mil seiscientas
grup-complex-spec-mp → *num-mp, num-mp.*
grup-complex-spec-fp → *num-fp, num-fp, num-fp.* %mil seiscientas setenta
grup-complex-spec-mp → *num-mp, num-mp, num-mp.*
grup-complex-spec-fp → *num-fp, num-fp, coord, num-fp.*
grup-complex-spec-mp → *num-mp, num-mp, coord, num-mp.*
grup-complex-spec-fp → *num-fp, num-fp, num-fp, coord, num-fp.*
grup-complex-spec-mp → *num-mp, num-mp, num-mp, coord, num-mp.*

espec-ms → *grup-complex-spec-ms.*
espec-mp → *grup-complex-spec-mp.*
espec-fs → *grup-complex-spec-fs.*
espec-fp → *grup-complex-spec-fp.*

B.3.1.3. Reglas para la sustantivación

sn → *j-ms, s-a-ms.*
sn → *j-fs, s-a-fs.*
sn → *j-mp, s-a-mp.*
sn → *j-fp, s-a-fp.*

sn → *j-ms, grup-sp.*
sn → *j-fs, grup-sp.*
sn → *j-mp, grup-sp.*

sn → *j-fp, grup-sp.*

B.4. Reglas para el sintagma adjetivo

sa → *s-a-ms.*
sa → *s-a-fs.*
sa → *s-a-mp.*
sa → *s-a-fp.*

s-a-ms → *a-ms.*
s-a-ms → *s-a-ms, Fc, s-a-ms.*
s-a-ms → *s-a-ms, Fc, s-a-ms, Fs.*
s-a-ms → *s-a-ms, coord, s-a-ms.*
s-a-ms → *sadv, a-ms.*
s-a-ms → *a-ms, s-a-ms.*

s-a-fs → *a-fs.*
s-a-fs → *s-a-fs, Fc, s-a-fs.*
s-a-fs → *s-a-fs, Fc, s-a-fs, Fs.*
s-a-fs → *s-a-fs, coord, s-a-fs.*
s-a-fs → *sadv, a-fs.*

s-a-mp → *a-mp.*
s-a-mp → *s-a-mp, Fc, s-a-mp.*
s-a-mp → *s-a-mp, Fc, s-a-mp, Fs.*
s-a-mp → *s-a-mp, coord, s-a-mp.*
s-a-mp → *sadv, a-mp.*
s-a-mp → *a-mp, s-a-mp.*

s-a-fp → *a-fp.*
s-a-fp → *s-a-fp, Fc, s-a-fp.*
s-a-fp → *s-a-fp, Fc, s-a-fp, Fs.*
s-a-fp → *s-a-fp, coord, s-a-fp.*
s-a-fp → *sadv, a-fp.*
s-a-fp → *a-fp, s-a-fp.*

B.5. Reglas para el sintagma adverbial

sadv → *adv-interrog.*
sadv → *adv.*

sadv → *adv(cerca), prep(de), sn.*
sadv → *adv(lejos), prep(de), sn.*
sadv → *adv(arriba), prep(de), sn.*
sadv → *adv(abajo), prep(de), sn.*
sadv → *adv(después), prep(de), sn.*
sadv → *adv(antes), prep(de), sn.*
sadv → *adv(fuera), prep(de), sn.*
sadv → *adv(dentro), prep(de), sn.*
sadv → *adv(delante), prep(de), sn.*
sadv → *adv(detrás), prep(de), sn.*
sadv → *adv(encima), prep(de), sn.*
sadv → *adv(debajo), prep(de), sn.*
sadv → *adv(más), prep(de), sn.*
sadv → *adv(menos), prep(de), sn.*
sadv → *adv(enfrente), prep(de), sn.*
sadv → *adv(frente), prep(a), sn.*
sadv → *adv(junto), prep(a), sn.*

sadv → *adv(cerca), prep(de), sadv.*
sadv → *adv(lejos), prep(de), sadv.*
sadv → *adv(arriba), prep(de), sadv.*
sadv → *adv(abajo), prep(de), sadv.*
sadv → *adv(después), prep(de), sadv.*
sadv → *adv(antes), prep(de), sadv.*
sadv → *adv(fuera), prep(de), sadv.*
sadv → *adv(dentro), prep(de), sadv.*
sadv → *adv(delante), prep(de), sadv.*
sadv → *adv(detrás), prep(de), sadv.*
sadv → *adv(encima), prep(de), sadv.*
sadv → *adv(debajo), prep(de), sadv.*
sadv → *adv(más), prep(de), sadv.*
sadv → *adv(menos), prep(de), sadv.*
sadv → *adv(enfrente), prep(de), sadv.*
sadv → *adv(frente), prep(a), sadv.*
sadv → *adv(junto), prep(a), sadv.*

B.6. Reglas para el grupo preposicional

grup-sp → *pp1cso00(conmigo).*
grup-sp → *pp2cso00(contigo).*
grup-sp → *pp3cso00(consigo).*

grup-sp → *prepc-ms, grup-nom-ms.*
grup-sp → *prepc-ms, pindef-ms.*
grup-sp → *prepc-ms, pposs-ms.*
grup-sp → *prep, numero.*
grup-sp → *prep, s-a-ms.*
grup-sp → *prep, s-a-mp.*
grup-sp → *prep, s-a-fs.*
grup-sp → *prep, s-a-fp.*
grup-sp → *prep, sadv.*
grup-sp → *prep, sn.*
grup-sp → *prep, ptonic.*
grup-sp → *prepc-ms, s-a-ms.*

grup-sp → *prep, infinitiu.*
grup-sp → *prepc-ms, infinitiu.*
grup-sp → *prep, inf.*
grup-sp → *prep, data.*
grup-sp → *prepc-ms, w.*

grup-sp → *prep(entre), sn, coord(y), sn.*

B.7. Reglas para el grupo verbal

B.7.1. Formas no perifrásticas

grup-verb → *verb.*
grup-verb → *verb-pass.*

verb → *vaux, parti.*

verb-pass → *vser, parti-flex.*

B.7.2. Perífrasis verbales: formas simples

verb → *vaic3s0(habría), cs(que), infinitiu.*

verb → *vmip3s0(debe), infinitiu.*
verb → *vmip1p0(debemos), infinitiu.*
verb → *vmip3p0(deben), infinitiu.*
verb → *vmip2s0(debes), infinitiu.*

verb → *vmip1s0(debo), infinitiu.*
verb → *vmip2p0(debéis), infinitiu.*

verb → *vmip3s0(debe), sps00(de), infinitiu.*
verb → *vmip1p0(debemos), sps00(de), infinitiu.*
verb → *vmip3p0(deben), sps00(de), infinitiu.*
verb → *vmip2s0(debes), sps00(de), infinitiu.*
verb → *vmip1s0(debo), sps00(de), infinitiu.*
verb → *vmip2p0(debéis), sps00(de), infinitiu.*

verb → *vmip1p0(tenemos), cs(que), infinitiu.*
verb → *vmip1s0(tengo), cs(que), infinitiu.*
verb → *vmip2p0(tenéis), cs(que), infinitiu.*
verb → *vmip3s0(tiene), cs(que), infinitiu.*
verb → *vmip3p0(tienen), cs(que), infinitiu.*
verb → *vmip2s0(tienes), cs(que), infinitiu.*

verb → *vaip3s0(ha), sps00(de), infinitiu.*
verb → *vaip2p0(habeís), sps00(de), infinitiu.*
verb → *vaip3p0(han), sps00(de), infinitiu.*
verb → *vaip2s0(has), sps00(de), infinitiu.*
verb → *vaip1s0(he), sps00(de), infinitiu.*
verb → *vaip1p0(hemos), sps00(de), infinitiu.*

verb → *vmip1p0(podemos), infinitiu.*
verb → *vmip2p0(podéis), infinitiu.*
verb → *vmip3s0(puede), infinitiu.*
verb → *vmip3p0(pueden), infinitiu.*
verb → *vmip2s0(puedes), infinitiu.*
verb → *vmip1s0(puedo), infinitiu.*

verb → *vmip3s0(va), sps00(a), infinitiu.*
verb → *vmip2p0(vais), sps00(a), infinitiu.*
verb → *vmip1p0(vamos), sps00(a), infinitiu.*
verb → *vmip3p0(van), sps00(a), infinitiu.*
verb → *vmip2s0(vas), sps00(a), infinitiu.*
verb → *vmip1s0(voy), sps00(a), infinitiu.*

verb → *vmip1p0(empezamos), sps00(a), infinitiu.*
verb → *vmip2p0(empezáis), sps00(a), infinitiu.*
verb → *vmip3s0(empieza), sps00(a), infinitiu.*
verb → *vmip3p0(empiezan), sps00(a), infinitiu.*
verb → *vmip2s0(empiezas), sps00(a), infinitiu.*
verb → *vmip1s0(empiezo), sps00(a), infinitiu.*

verb → *vmip1p0(comenzamos), sps00(a), infinitiu.*
verb → *vmip2p0(comenzáis), sps00(a), infinitiu.*
verb → *vmip3s0(comienza), sps00(a), infinitiu.*
verb → *vmip3p0(comienzan), sps00(a), infinitiu.*
verb → *vmip2s0(comienzas), sps00(a), infinitiu.*
verb → *vmip1s0(comienzo), sps00(a), infinitiu.*

verb → *vmip3s0(echa), sps00(a), infinitiu.*
verb → *vmip1p0(echamos), sps00(a), infinitiu.*
verb → *vmip3p0(echan), sps00(a), infinitiu.*
verb → *vmip2s0(echas), sps00(a), infinitiu.*
verb → *vmip1s0(echo), sps00(a), infinitiu.*
verb → *vmip2p0(echáis), sps00(a), infinitiu.*

verb → *vmip3s0(rompe), sps00(a), infinitiu.*
verb → *vmip1p0(rompemos), sps00(a), infinitiu.*
verb → *vmip3p0(rompen), sps00(a), infinitiu.*
verb → *vmip2s0(rompes), sps00(a), infinitiu.*
verb → *vmip1s0(rompo), sps00(a), infinitiu.*
verb → *vmip2p0(rompéis), sps00(a), infinitiu.*

verb → *vmip1p0(estamos), sps00(a_punto_de), infinitiu.*
verb → *vmip1s0(estoy), sps00(a_punto_de), infinitiu.*
verb → *vmip3s0(está), sps00(a_punto_de), infinitiu.*
verb → *vmip2p0(estáis), sps00(a_punto_de), infinitiu.*
verb → *vmip3p0(están), sps00(a_punto_de), infinitiu.*
verb → *vmip2s0(estás), sps00(a_punto_de), infinitiu.*

verb → *vmip1p0(volvemos), sps00(a), infinitiu.*
verb → *vmip2p0(volvéis), sps00(a), infinitiu.*
verb → *vmip3s0(vuelve), sps00(a), infinitiu.*
verb → *vmip3p0(vuelven), sps00(a), infinitiu.*
verb → *vmip2s0(vuelves), sps00(a), infinitiu.*
verb → *vmip1s0(vuelvo), sps00(a), infinitiu.*

verb → *vmip1p0(solemos), infinitiu.*
verb → *vmip2p0(soléis), infinitiu.*
verb → *vmip3s0(suele), infinitiu.*
verb → *vmip3p0(suelen), infinitiu.*
verb → *vmip2s0(sueles), infinitiu.*
verb → *vmip3s0(suelo), infinitiu.*

verb → *vmip3s0(acostumbra), sps00(a), infinitiu.*

verb → *vmip1p0(acostumbramos), sps00(a), infinitiu.*
verb → *vmip3p0(acostumbran), sps00(a), infinitiu.*
verb → *vmip2s0(acostumbras), sps00(a), infinitiu.*
verb → *vmip1s0(acostumbro), sps00(a), infinitiu.*
verb → *vmip2p0(acostumbráis), sps00(a), infinitiu.*

verb → *vmip3s0(acaba), sps00(de), infinitiu.*
verb → *vmip1p0(acabamos), sps00(de), infinitiu.*
verb → *vmip3p0(acaban), sps00(de), infinitiu.*
verb → *vmip2s0(acabas), sps00(de), infinitiu.*
verb → *vmip1s0(acabo), sps00(de), infinitiu.*
verb → *vmip2p0(acabáis), sps00(de), infinitiu.*

verb → *vmip3s0(termina), sps00(de), infinitiu.*
verb → *vmip1p0(terminamos), sps00(de), infinitiu.*
verb → *vmip3p0(terminan), sps00(de), infinitiu.*
verb → *vmip1s0(termino), sps00(de), infinitiu.*
verb → *vmip2s0(terminas), sps00(de), infinitiu.*
verb → *vmip2p0(termináis), sps00(de), infinitiu.*

verb → *vmip3s0(deja), sps00(de), infinitiu.*
verb → *vmip1p0(dejamos), sps00(de), infinitiu.*
verb → *vmip3p0(dejan), sps00(de), infinitiu.*
verb → *vmip2s0(dejas), sps00(de), infinitiu.*
verb → *vmip1s0(dejo), sps00(de), infinitiu.*
verb → *vmip2p0(dejáis), sps00(de), infinitiu.*

verb → *vmip3s0(llega), sps00(a), infinitiu.*
verb → *vmip1p0(llegamos), sps00(a), infinitiu.*
verb → *vmip3p0(llegan), sps00(a), infinitiu.*
verb → *vmip2s0(llegas), sps00(a), infinitiu.*
verb → *vmip1s0(llego), sps00(a), infinitiu.*
verb → *vmip2p0(llegáis), sps00(a), infinitiu.*

verb → *vmip1p0(acertamos), sps00(a), infinitiu.*
verb → *vmip2p0(acertáis), sps00(a), infinitiu.*
verb → *vmip3s0(acierta), sps00(a), infinitiu.*
verb → *vmip3p0(aciertan), sps00(a), infinitiu.*
verb → *vmip2s0(aciertas), sps00(a), infinitiu.*
verb → *vmip1s0(acierto), sps00(a), infinitiu.*

verb → *vmip3s0(alcanza), sps00(a), infinitiu.*
verb → *vmip1p0(alcanzamos), sps00(a), infinitiu.*
verb → *vmip3p0(alcanzan), sps00(a), infinitiu.*

verb → *vmip2s0(alcanzas), sps00(a), infinitiu.*
verb → *vmip1s0(alcanzo), sps00(a), infinitiu.*
verb → *vmip2p0(alcanzáis), sps00(a), infinitiu.*

verb → *vmip3s0(tarda), sps00(en), infinitiu.*
verb → *vmip1p0(tardamos), sps00(en), infinitiu.*
verb → *vmip3p0(tardan), sps00(en), infinitiu.*
verb → *vmip2s0(tardas), sps00(en), infinitiu.*
verb → *vmip1s0(tardo), sps00(en), infinitiu.*
verb → *vmip2p0(tardáis), sps00(en), infinitiu.*

verb → *vmip3s0(acaba), gerundi.*
verb → *vmip1p0(acabamos), gerundi.*
verb → *vmip3p0(acaban), gerundi.*
verb → *vmip2s0(acabas), gerundi.*
verb → *vmip1s0(acabo), gerundi.*
verb → *vmip2p0(acabáis), gerundi.*

verb → *vmip3s0(anda), gerundi.*
verb → *vmip1p0(andamos), gerundi.*
verb → *vmip3p0(andan), gerundi.*
verb → *vmip2s0(andas), gerundi.*
verb → *vmip1s0(ando), gerundi.*
verb → *vmip2p0(andáis), gerundi.*

verb → *vmip1p0(comenzamos), gerundi.*
verb → *vmip2p0(comenzáis), gerundi.*
verb → *vmip3s0(comienza), gerundi.*
verb → *vmip3p0(comienzan), gerundi.*
verb → *vmip2s0(comienzas), gerundi.*
verb → *vmip1s0(comienzo), gerundi.*

verb → *vmip1p0(continuamos), gerundi.*
verb → *vmip3s0(continúa), gerundi.*
verb → *vmip3p0(continúan), gerundi.*
verb → *vmip2s0(continúas), gerundi.*
verb → *vmip1s0(continúo), gerundi.*
verb → *vmip2p0(continuáis), gerundi.*

verb → *vmip1p0(empezamos), gerundi.*
verb → *vmip2p0(empezáis), gerundi.*
verb → *vmip3s0(empieza), gerundi.*
verb → *vmip3p0(empiezan), gerundi.*
verb → *vmip2s0(empiezas), gerundi.*

verb → *vmip1s0(empiezo), gerundi.*

verb → *vmip1p0(estamos), gerundi.*
verb → *vmip1s0(estoy), gerundi.*
verb → *vmip3s0(está), gerundi.*
verb → *vmip2p0(estáis), gerundi.*
verb → *vmip3p0(están), gerundi.*
verb → *vmip2s0(estás), gerundi.*

verb → *vmip3s0(va), gerundi.*
verb → *vmip2p0(vais), gerundi.*
verb → *vmip1p0(vamos), gerundi.*
verb → *vmip3p0(van), gerundi.*
verb → *vmip2s0(vas), gerundi.*
verb → *vmip1s0(voy), gerundi.*

verb → *vmip3s0(lleva), gerundi.*
verb → *vmip1p0(llevamos), gerundi.*
verb → *vmip3p0(llevan), gerundi.*
verb → *vmip2s0(llevas), gerundi.*
verb → *vmip1s0(llevo), gerundi.*
verb → *vmip2p0(lleváis), gerundi.*

verb → *vmip1p0(seguimos), gerundi.*
verb → *vmip2p0(seguís), gerundi.*
verb → *vmip1s0(sigo), gerundi.*
verb → *vmip3s0(sigue), gerundi.*
verb → *vmip3p0(siguen), gerundi.*
verb → *vmip2s0(sigues), gerundi.*

verb → *vmip3s0(termina), gerundi.*
verb → *vmip1p0(terminamos), gerundi.*
verb → *vmip3p0(terminan), gerundi.*
verb → *vmip2s0(terminas), gerundi.*
verb → *vmip1s0(termino), gerundi.*
verb → *vmip2p0(termináis), gerundi.*

verb → *vmip1s0(vengo), gerundi.*
verb → *vmip1p0(venimos), gerundi.*
verb → *vmip2p0(venís), gerundi.*
verb → *vmip3s0(viene), gerundi.*
verb → *vmip3p0(vienen), gerundi.*
verb → *vmip2s0(vienes), gerundi.*

B.7.3. Perífrasis verbales: formas complejas

verb → vaux, vmp00sm(debido), infinitiu.
verb → vaux, vmp00sm(debido), sps00(de), infinitiu.
verb → vaux, vmp00sm(tenido), cs(que),infinitiu.
verb → vaux, vmp00sm(habido), sps00(de),infinitiu.
verb → vaux, vmp00sm(podido), infinitiu.
verb → vaux, vmp00sm(ido), sps00(a), infinitiu.
verb → vaux, vmp00sm(empezado), sps00(a), infinitiu.
verb → vaux, vmp00sm(comenzado), sps00(a),infinitiu.
verb → vaux, vmp00sm(echado), sps00(a), infinitiu.
verb → vaux, vmp00sm(roto), sps00(a), infinitiu.
verb → vaux, vmp00sm(estado), sps00(a_punto_de), infinitiu.
verb → vaux, vmp00sm(vuelto), sps00(a), infinitiu.
verb → vaux, vmp00sm(acostumbrado), sps00(a), infinitiu.
verb → vaux, vmp00sm(acabado), sps00(de), infinitiu.
verb → vaux, vmp00sm(terminado), sps00(de), infinitiu.
verb → vaux, vmp00sm(dejado), sps00(de), infinitiu.
verb → vaux, vmp00sm(llegado), sps00(a), infinitiu.
verb → vaux, vmp00sm(acertado), sps00(a), infinitiu.
verb → vaux, vmp00sm(alcanzado), sps00(a), infinitiu.
verb → vaux, vmp00sm(tardado), sps00(en), infinitiu.

verb → vaux, vmp00sm(acabado), gerundi.
verb → vaux, vmp00sm(andado), gerundi.
verb → vaux, vmp00sm(comenzado), gerundi.
verb → vaux, vmp00sm(continuado), gerundi.
verb → vaux, vmp00sm(empezado), gerundi.
verb → vaux, vmp00sm(estado), gerundi.
verb → vaux, vmp00sm(ido), gerundi.
verb → vaux, vmp00sm(llevado), gerundi.
verb → vaux, vmp00sm(seguido), gerundi.
verb → vaux, vmp00sm(terminado), gerundi.
verb → vaux, vmp00sm(venido), gerundi.

inf → vmn0000(deber), infinitiu.
inf → vmn0000(deber), sps00(de), infinitiu.
inf → vmn0000(tener), cs(que), infinitiu.
inf → vmn0000(poder), infinitiu.
inf → vmn0000(ir), sps00(a), infinitiu.
inf → vmn0000(empezar), sps00(a), infinitiu.
inf → vmn0000(comenzar), sps00(a), infinitiu.
inf → vmn0000(echarse), sps00(a), infinitiu.

inf → *vmn0000(echar), sps00(a), infinitiu.*
inf → *vmn0000(romper), sps00(a), infinitiu.*
inf → *vmn0000(volver), sps00(a), infinitiu.*
inf → *vmn0000(soler), infinitiu.*
inf → *vmn0000(acostumbrar), sps00(a), infinitiu.*
inf → *vmn0000(acabar), sps00(de), infinitiu.*
inf → *vmn0000(terminar), sps00(de), infinitiu.*
inf → *vmn0000(dejar), sps00(de), infinitiu.*
inf → *vmn0000(venir), sps00(a), infinitiu.*
inf → *vmn0000(llegar), sps00(a), infinitiu.*
inf → *vmn0000(acertar), sps00(a), infinitiu.*
inf → *vmn0000(alcanzar), sps00(a), infinitiu.*
inf → *vmn0000(tardar), sps00(en), infinitiu.*
inf → *van0000(haber), sps00(de), infinitiu.*
inf → *vmn0000(estar), gerundi.*
inf → *vmn0000(acabar), gerundi.*
inf → *vmn0000(andar), gerundi.*
inf → *vmn0000(comenzar), gerundi.*
inf → *vmn0000(continuar), gerundi.*
inf → *vmn0000(empezar), gerundi.*
inf → *vmn0000(ir), gerundi.*
inf → *vmn0000(llevar), gerundi.*
inf → *vmn0000(terminar), gerundi.*
inf → *vmn0000(seguir), gerundi.*
inf → *vmn0000(venir), gerundi.*
inf → *infaux, vmp00sm(debido), infinitiu.*
inf → *infaux, vmp00sm(debido), sps00(de), infinitiu.*
inf → *infaux, vmp00sm(tenido), cs(que), infinitiu.*
inf → *infaux, vmp00sm(acabado), sps00(de), infinitiu.*
inf → *infaux, vmp00sm(acostumbrado), sps00(a), infinitiu.*
inf → *infaux, vmp00sm(comenzado), sps00(a), infinitiu.*
inf → *infaux, vmp00sm(echado), sps00(a), infinitiu.*
inf → *infaux, vmp00sm(llegado), sps00(a), infinitiu.*
inf → *infaux, vmp00sm(roto), sps00(a), infinitiu.*
inf → *infaux, vmp00sm(tardado), sps00(en), infinitiu.*
inf → *infaux, vmp00sm(terminado), sps00(de), infinitiu.*
inf → *infaux, vmp00sm(acertado), sps00(a), infinitiu.*
inf → *infaux, vmp00sm(alcanzado), sps00(a), infinitiu.*
inf → *infaux, vmp00sm(dejado), sps00(de), infinitiu.*
inf → *infaux, vmp00sm(empezado), sps00(a), infinitiu.*
inf → *infaux, vmp00sm(ido), sps00(a), infinitiu.*
inf → *infaux, vmp00sm(vuelto), sps00(a), infinitiu.*
inf → *infaux, vmp00sm(podido), infinitiu.*
inf → *infaux, vmp00sm(solido), infinitiu.*

inf → *infaux, vmp00sm(estado), gerundi.*
inf → *infaux, vmp00sm(acabado), gerundi.*
inf → *infaux, vmp00sm(andado), gerundi.*
inf → *infaux, vmp00sm(comenzado), gerundi.*
inf → *infaux, vmp00sm(continuado), gerundi.*
inf → *infaux, vmp00sm(empezado), gerundi.*
inf → *infaux, vmp00sm(ido), gerundi.*
inf → *infaux, vmp00sm(llevado), gerundi.*
inf → *infaux, vmp00sm(terminado), gerundi.*
inf → *infaux, vmp00sm(seguido), gerundi.*
inf → *infaux, vmp00sm(venido), gerundi.*

ger → *vmg0000(debiendo), infinitiu.*
ger → *vmg0000(debiendo), sps00(de), infinitiu.*
ger → *vmg0000(teniendo), cs(que), infinitiu.*
ger → *vmg0000(pudiendo), infinitiu.*
ger → *vmg0000(yendo), sps00(a), infinitiu.*
ger → *vmg0000(empezando), sps00(a), infinitiu.*
ger → *vmg0000(comenzando), sps00(a), infinitiu.*
ger → *vmg0000(echando), sps00(a), infinitiu.*
ger → *vmg0000(rompiendo), sps00(a), infinitiu.*
ger → *vmg0000(volviendo), sps00(a), infinitiu.*
ger → *vmg0000(acostumbrando),sps00(a), infinitiu.*
ger → *vmg0000(acabando), sps00(de), infinitiu.*
ger → *vmg0000(terminando), sps00(de), infinitiu.*
ger → *vmg0000(dejando), sps00(de), infinitiu.*
ger → *vmg0000(viniendo), sps00(a), infinitiu.*
ger → *vmg0000(llegando), sps00(a), infinitiu.*
ger → *vmg0000(acertando), sps00(a), infinitiu.*
ger → *vmg0000(alcanzando), sps00(a), infinitiu.*
ger → *vmg0000(tardando), sps00(en), infinitiu.*
ger → *vag0000(habiendo), sps00(de), infinitiu.*
ger → *vmg0000(acabando), gerundi.*
ger → *vmg0000(andando), gerundi.*
ger → *vmg0000(comenzando), gerundi.*
ger → *vmg0000(continuando), gerundi.*
ger → *vmg0000(empezando), gerundi.*
ger → *vmg0000(estando), gerundi.*
ger → *vmg0000(yendo), gerundi.*
ger → *vmg0000(llevando), gerundi.*
ger → *vmg0000(siguiendo), gerundi.*
ger → *vmg0000(terminando), gerundi.*
ger → *vmg0000(viniendo), gerundi.*

B.8. Reglas para los elementos restantes

B.8.1. Formas no personales del verbo

parti-flex → *parti-ms.*
parti-flex → *parti-fs.*
parti-flex → *parti-mp.*
parti-flex → *parti-fp.*

gerundi → *ger.*
gerundi → *ger-pas.*
gerundi → *geraux-ser.*
ger → *geraux, parti.* %habiendo comido
ger → *geraux, parti-ser.* %habiendo sido
ger-pas → *geraux-ser, parti-flex.*
ger-pas → *geraux, parti-aux, parti-flex.*

infinitiu → *inf.*
infinitiu → *inf-pas.*
infinitiu → *infaux-ser.*
inf → *infaux, parti.* %haber cantado

inf → *infaux, parti-ser.* %haber sido
inf-pas → *infaux-ser, parti-flex.* %ser amado/a/os/as
inf-pas → *infaux, parti-ser, parti-flex.* %haber sido amado

B.8.2. Reglas para los relativos

relatiu → *prel-mp, grup-nom-mp.*
relatiu → *prel-ms, grup-nom-ms.*
relatiu → *prel-fp, grup-nom-fp.* %cuantas personas
relatiu → *prel-fs, grup-nom-fs.* %cuantos asistentes

relatiu → *prepc-ms, cual-s.* %al/del cual
relatiu → *prep, j-ms, cual-s.* %con el cual
relatiu → *prep, j-fs, cual-s.* %con la cual
relatiu → *prep, j-mp, cual-p.* %con los cuales
relatiu → *prep, j-fp, cual-p.* %con las cuales
relatiu → *j-ms, cual-s.* %el cual
relatiu → *j-fs, cual-s.* %la cual
relatiu → *j-mp, cual-p.* %los cuales

relatiu → *j-fp, cual-p.* %las cuales
relatiu → *cuyo-ms, grup-nom-ms.* %cuyo amigo
relatiu → *cuyo-fs, grup-nom-fs.*
relatiu → *cuyo-mp, grup-nom-mp.*
relatiu → *cuyo-fp, grup-nom-fp.*
relatiu → *prep, cuyo-ms, grup-nom-ms.* %con cuyo amigo
relatiu → *prep, cuyo-fs, grup-nom-fs.*
relatiu → *prep, cuyo-mp, grup-nom-mp.*
relatiu → *prep, cuyo-fp, grup-nom-fp.*

B.9. Nodos entre signos de puntuación

sn → *espec-ms, Fe, grup-nom-ms, Fe.*
sn → *espec-fs, Fe, grup-nom-fs, Fe.*
sn → *espec-mp, Fe, grup-nom-mp, Fe.*
sn → *espec-fp, Fe, grup-nom-fp, Fe.*

sn → *espec-ms, Fpa, grup-nom-ms, Fpt.*
sn → *espec-fs, Fpa, grup-nom-fs, Fpt.*
sn → *espec-mp, Fpa, grup-nom-mp, Fpt.*
sn → *espec-fp, Fpa, grup-nom-fp, Fpt.*

s-a-ms → *Fe, s-a-ms, Fe.*
s-a-fs → *Fe, s-a-fs, Fe.*
s-a-mp → *Fe, s-a-mp, Fe.*
s-a-fp → *Fe, s-a-fp, Fe.*
s-a-ms → *Fpa, s-a-ms, Fpt.*
s-a-fs → *Fpa, s-a-fs, Fpt.*
s-a-mp → *Fpa, s-a-mp, Fpt.*
s-a-fp → *Fpa, s-a-fp, Fpt.*

sn → *Fe, sn, Fe.*
sn → *Fpa, sn, Fpt.*

Apéndice C

Etiquetas utilizadas en la anotación de Cast3LB

En este apéndice aparecen las etiquetas utilizadas para la etiquetación de corpus **Cast3LB**. En primer lugar aparecen las etiquetas morfológicas y seguidamente las de los constituyentes que se consideran en este corpus.

C.1. Etiquetas morfológicas

ao0fp0, ao0fs0, ao0mp0, ao0ms0, aq0cn0, aq0cp0,
aq0cs0, aq0fn0, aq0fp0, aq0fpp,
aq0fs0, aq0fsp, aq0mn0,
aq0mp0, aq0mpp, aq0ms0, aq0msp, cc, cs, da0fp0,
da0fs0, da0mp0,
da0ms0, da0ns0, dd0cp0, dd0cs0, dd0fp0, dd0fs0, dd0mp0,
dd0ms0, de0cn0, de0fp0, de0fs0, de0mp0, de0ms0, di0cn0,
di0cp0, di0cs0, di0fp0, di0fs0, di0mp0, di0ms0, dn0cp0,
dn0cs0,
dn0fp0, dn0fs0, dn0mp0, dn0ms0, dp1cps, dp1css, dp1fpp,
dp1fsp, dp1fss, dp1mpp, dp1mps, dp1msp, dp1mss, dp2cps,
dp2css, dp2fpp, dp2fsp, dp2mpp, dp2mps, dp2msp, dp2mss,
dp3cp0,
dp3cs0, dp3fp0, dp3fs0, dp3mp0, dp3ms0, dt0cn0, dt0fp0,
dt0fs0, Faa, Fat, Fc, Fd, Fe, Fg, Fh, Fia, Fit,
Fp, Fpa, Fpt, Fs, Fx,
Fz, I, nc00000, nccn000, nccp000, nccs000, ncfn000,
ncfp000, ncfs000, ncmn000, ncmp000, ncms000, np00000,
p0000000, p010p000, p010s000, p020p000, p020s000,
p0300000, pd0cp000, pd0cs000, pd0fp000,
pd0fs000, pd0mp000, pd0ms000, pd0ns000, pe000000,
pi0cp000, pi0cs000, pi0fp000, pi0fs000, pi0mp000,
pi0ms000, pn0cp000, pn0fp000, pn0fs000, pn0mp000,

pn0ms000, pp1cp000, pp1cs000, pp1csn00, pp1cso00,
pp1fp000, pp1mp000,
pp2cn00p, pp2cp000, pp2cp00p, pp2cs000, pp2cs00p,
pp2csn00, pp2cso00, pp2fp000, pp2mp000, pp3cn000,
pp3cna00, pp3cno00, pp3cpa00, pp3cpd00, pp3csa00,
pp3csd00, pp3fp000, pp3fpa00, pp3fs000, pp3fsa00,
pp3mp000, pp3mpa00,
pp3ms000, pp3msa00, pp3ns000, pr000000, pr0cn000,
pr0cp000, pr0cs000, pr0fp000, pr0fs000, pr0mp000,
pr0ms000, pt000000, pt0cp000, pt0cs000, pt0mp000,
pt0ms000, px1fs0p0, px1mp0p0, px1ms0p0, px1ms0s0,
px2fs0s0, px3fs000,
px3mp000, px3ms000, px3ns000, rg, rn, spcms, sps00,
vag0000, vaic1p0, vaic1s0, vaic2p0, vaic2s0, vaic3p0,
vaic3s0, vaif1p0, vaif1s0, vaif2p0, vaif2s0, vaif3p0,
vaif3s0, vaii1p0, vaii1s0,
vaii2p0, vaii2s0, vaii3p0, vaii3s0, vaip1p0, vaip1s0,
vaip2p0, vaip2s0, vaip3p0, vaip3s0, vais1p0, vais1s0,
vais2p0, vais2s0, vais3p0, vais3s0, vam01p0, vam02p0,
vam02s0, vam03p0, vam03s0, van0000,
vap00pf, vap00pm, vap00sf, vap00sm, vasf1p0, vasf1s0,
vasf2p0, vasf2s0, vasf3p0, vasi1p0, vasi1s0, vasi2p0,
vasi2s0, vasi3p0, vasi3s0, vasp1p0, vasp1s0, vasp2p0,
vasp2s0, vasp3p0, vasp3s0, vmg0000,
vmi0000, vmic1p0, vmic1s0, vmic2p0, vmic2s0, vmic3p0,
vmic3s0, vmif1p0, vmif1s0, vmif2p0, vmif2s0, vmif3p0,
vmif3s0, vmii1p0, vmii1s0, vmii2p0, vmii2s0, vmii3p0,
vmii3s0, vmip1p0, vmip1s0, vmip2p0,
vmip2s0, vmip3p0, vmip3s0, vmis1p0, vmis1s0, vmis2p0,
vmis2s0, vmis3p0, vmis3s0, vmm01p0, vmm02p0, vmm02s0,
vmm03p0, vmm03s0, vmn0000, vmp0000, vmp00pf, vmp00pm,
vmp00sf, vmp00sm, vmsf1p0, vmsf1s0,
vmsf2p0, vmsf2s0, vmsf3p0, vmsf3s0, vmsi1p0, vmsi1s0,
vmsi2p0, vmsi2s0, vmsi3p0, vmsi3s0, vmsp1p0, vmsp1s0,
vmsp2p0, vmsp2s0, vmsp3p0, vmsp3s0, vsg0000, vsic1p0,
vsic1s0, vsic2p0, vsic2s0, vsic3p0,
vsic3s0, vsif1p0, vsif1s0, vsif2p0, vsif2s0, vsif3p0,
vsif3s0, vsii1p0, vsii1s0, vsii2p0, vsii3p0, vsii3s0,
vsip1p0, vsip1s0, vsip2p0, vsip2s0, vsip3p0, vsip3s0,
vsis1s0, vsis3p0, vsis3s0, vsm01p0, vsm03p0,
vsm03s0, vsn0000, vsp00sm, vssf3s0, vssi1p0, vssi3p0,
vssi3s0, vssp1s0, vssp2p0, vssp2s0, vssp3p0, vssp3s0, W,
X, Y, Z, Zm, Zp,

C.2. Etiquetas para los constituyentes

INC, INC.co,
S, S.co, S*, S*.co,
S.F.A, S.F.A.co, S.F.AComp, S.F.AConc, S.F.ACond, S.F.ACons,
S.F.A*,
S.F.A.co, S.F.AComp*, S.F.AConc*, S.F.ACond*, S.F.ACons*,
S.F.AComp.co, S.F.AConc.co, S.F.ACond.co, S.F.ACons.co,
S.F.C, S.F.C*, S.F.C.co,
S.F.R, S.F.R*, S.F.R.co,
sn, sn.co, sn.e, S.NF.C, S.NF.A, S.NF.A.co, S.NF.C.co,
S.NF.P, S.NF.PA, S.NF.A*, S.NF.C*, S.NF.P*,
S.NF.P.co, S.NF.R, S.NF.R.co, conj.subord, coord, data,
espec.ms, espec.ms.co,
espec.mp, espec.mp.co, espec.fs, espec.fs.co, espec.fp,
espec.fp.co, gerundi,
grup.nom, grup.nom.co, grup.nom.s, grup.nom.p,
grup.nom.ms, grup.nom.ms.co,
grup.nom.mp, grup.nom.mp.co,
grup.nom.fs, grup.nom.fs.co,
grup.nom.fp, grup.nom.fp.co,
gv, infinitiu,
interjeccio, morf.pron, morfema.verbal, neg, numero, prep,
relatiu,
sa, sa.co,
s.a.ms, s.a.ms.co,
s.a.mp, s.a.mp.co,
s.a.fs, s.a.fs.co,
s.a.fp, s.a.fp.co,
sadv, sadv.co, sp,
sp.co, sp.de,

Apéndice D

Corpus CLiC-TALP desambiguado

En este apéndice se presenta el primer fragmento del corpus CLiC-TALP desambiguado, tras la fase de validación manual. La primera columna corresponde a las palabras; la segunda al lema; y la tercera a la etiqueta morfosintáctica.

Medardo_Fraile medardo_fraile NP00000
juega jugar VMIP3S0
a a SPS00
un uno DI0MS0
cinismo cinismo NCMS000
fácil fácil AQ0CS0
y y CC
divertido divertido AQ0MSP
. . Fp
No no RN
quiero querer VMIP1S0
decir decir VMN0000
que que CS
lo él PP3CNA00
sea ser VSM03S0
, , Fc
cínico cínico AQ0MS0
o o CC
divertido divertido AQ0MSP
, , Fc
sino_que sino_que CC
ante ante SPS00
un uno DI0MS0
mazo mazo NCMS000
de de SPS00
hojas hoja NCFP000
grabadas grabado AQ0FPP
coloca colocar VMIP3S0
un uno DI0MS0
cristal cristal NCMS000
bien bien RG
tallado tallado AQ0MSP
y y CC
lo él PP3MSA00
hace hacer VMIP3S0
girar girar VMN0000
para_que para_que CS
el el DA0MS0
sol sol NCMS000
rompa romper VMSP3S0
contra contra SPS00
él él PP3MS000
sus su DP3CP0
rayos rayo NCMP000
. . Fp
Resulta resultar VMIP3S0
entonces entonces RG
que que CS
no no RN
practica practicar VMIP3S0
las el DA0FP0
historias historia NCFP000
del del SPCMS
espejo espejo NCMS000
, , Fc

del del SPCMS
tomavistas tomavistas NCMN000
o o CC
de de SPS00
todos todo DI0MP0
esos ese DD0MP0
inventos invento NCMP000
que que PR0CN000
tan tanto RG
poca poco DI0FS0
ilusión ilusión NCFS000
nos yo PP1CP000
producen producir VMIP3P0
ya ya RG
. . Fp
Hace hacer VMIP3S0
dar dar VMN0000
vueltas volver NCFP000
a a SPS00
su su DP3CS0
cristal cristal NCMS000
y y CC
la el DA0FS0
luz luz NCFS000
se él P0300000
descompone descomponer VMIP3S0
o o CC
se él P0300000
polariza polarizar VMIP3S0
o o CC
se él P0300000
desparrama desparramar VMIP3S0
y y CC
entonces entonces RG
nada nada PI0CS000
es ser VSIP3S0
como como CS
la el DA0FS0
costumbre costumbre NCFS000
tediosa tedioso AQ0FS0
a a SPS00
la el DA0FS0
que que PR0CN000
estamos estar VMIP1P0
acostumbrados acostumbrado AQ0MPP
. . Fp
Para para SPS00
entendernos entender VMN0000
, , Fc
diríamos decir VMIC1P0
esperpentos esperpento NCMP000
. . Fp
No no RN
marraríamos marrar VMIC1P0
mucho mucho RG
. . Fp
“ “ Fe
El_rey_y_el_país_con_granos
el_rey_y_el_país_con_granos
NP00000
“ “ Fe
es ser VSIP3S0
un uno DI0MS0
puro puro AQ0MS0
esperpento esperpento NCMS000
, , Fc
pero pero CC
no no RN
porque porque CS
el el DA0MS0
héroe héroe NCMS000
se él P0300000
haya haber VASP3S0
enfrentado enfrentar VMP00SM
a a SPS00
un uno DI0MS0
espejo espejo NCMS000
deformante deformante AQ0CS0
, , Fc
sino sino CC
porque porque CS
el el DA0MS0
escritor escritor NCMS000
va ir VMIP3S0
haciendo hacer VMG0000
guiar guiar VMN0000
un uno DI0MS0
berilo berilo NCMS000

clarísimo clarísimo AQ0MS0
sobre sobre SPS00
una uno DI0FS0
página página NCFS000
de de SPS00
la el DA0FS0
historia historia NCFS000
y y CC
lo el DA0NS0
que que PR0CN000
resulta resultar VMIP3S0
es ser VSIP3S0
una uno DI0FS0
infinita infinito AQ0FS0
sarta sarta NCFS000
de de SPS00
coherentes coherente AQ0CP0
absurdos absurdo NCMP000
, , Fc
como como CS
si si CS
un uno DI0MS0
diablillo diablillo NCMS000
enredador enredador NCMS000
hiciera hacer VMSI3S0
mofa mofa NCFS000
de de SPS00
todos todo DI0MP0
los el DA0MP0
principios principio NCMP000
solemnes solemne AQ0CP0
y y CC
nos yo PP1CP000
regalara regalar VMSI3S0
con con SPS00
el el DA0MS0
puro puro AQ0MS0
disparate disparate NCMS000
. . Fp
Claro claro RG
que que CS
el el DA0MS0
disparate disparate NCMS000
es ser VSIP3S0
contemplar contemplar VMN0000
aquella aquel DD0FS0
realidad realidad NCFS000
con con SPS00
la el DA0FS0
fría frío AQ0FS0
lógica lógica NCFS000
del del SPCMS
adormecido adormecido AQ0MSP
burgués burgués NCMS000
cuando cuando CS
lo el DA0NS0
que que PR0CN000
hay haber VAIP3S0
que que CS
hacer hacer VMN0000
es ser VSIP3S0
detener detener VMN0000
los el DA0MP0
rayos rayo NCMP000
disparados disparado AQ0MPP
desde desde SPS00
el el DA0MS0
perfil perfil NCMS000
de de SPS00
las el DA0FP0
aristas arista NCFP000
y y CC
no no RN
mirar mirar VMN0000
las el DA0FP0
figuras figura NCFP000
que que PR0CN000
la el DA0FS0
hoja hoja NCFS000
de de SPS00
papel papel NCMS000
tenía tener VMII3S0
impresas impreso AQ0FPP
. . Fp
La el DA0FS0
historia historia NCFS000
existió existir VMIS3S0
y y CC

vale valer VMIP3S0
como como CS
referente referente AQ0CS0
, , Fc
pero pero CC
los el DA0MP0
grabados grabado NCMP000
de de SPS00
colorines colorín NCMP000
son ser VSIP3P0
otra otro DI0FS0
cosa cosa NCFS000
que que CS
el el DA0MS0
solemne solemne AQ0CS0
testimonio testimonio NCMS000
de de SPS00
su su DP3CS0
presencia presencia NCFS000
. . Fp

Apéndice E

Corpus CLiC-TALP analizado sintácticamente

En este apéndice se presenta el primer fragmento del corpus CLiC-TALP con el análisis sintáctico superficial proporcionado por TACAT y GramEsp.

```
S_[
    sn_[
        grup-nom-fp_[
            Medardo_Fraile_np00000
        ]
    ]
    grup-verb_[
        juega_vmip3s0
    ]
    grup-sp_[
        prep_[
            a_sps00
        ]
        sn_[
            espec-ms_[
                un_di0ms0
            ]
            grup-nom-ms_[
                cinismo_ncms000
                s-a-ms_[
                    s-a-ms_[
                        fácil_aq0cs0
                    ]
                    coord_[
                        y_cc
                    ]
                    s-a-ms_[
                        divertido_aq0msp
                    ]
                ]
            ]
        ]
    ]
    ._Fp
]
S_[
    neg_[
        No_rn
```

```
]
grup-verb_[
    quiero_vmip1s0
]
infinitiu_[
    decir_vmn0000
]
conj-subord_[
    que_cs
]
patons_[
    paton_[
        lo_pp3cna00
    ]
]
grup-verb_[
    sea_vsm03s0
]
,_Fc
sa_[
    s-a-ms_[
        s-a-ms_[
            cínico_aq0ms0
        ]
        coord_[
            o_cc
        ]
        s-a-ms_[
            divertido_aq0msp
        ]
    ]
]
,_Fc
coord_[
    sino_que_cc
]
grup-sp_[
    prep_[
        ante_sps00
    ]
    sn_[
        espec-ms_[
            un_di0ms0
        ]
        grup-nom-ms_[
            mazo_ncms000
            sp-de_[
                prep_[
                    de_sps00
                ]
                sn_[
                    grup-nom-fp_[
                        hojas_ncfp000
                        s-a-fp_[
                            grabadas_aq0fpp
                        ]
                    ]
                ]
            ]
        ]
```

```
    ]
]
grup-verb_[
    coloca_vmip3s0
]
sn_[
    espec-ms_[
        un_di0ms0
    ]
    grup-nom-ms_[
        cristal_ncms000
        s-a-ms_[
            sadv_[
                bien_rg
            ]
            tallado_aq0msp
        ]
    ]
]
coord_[
    y_cc
]
patons_[
    lo_pp3msa00
]
grup-verb_[
    hace_vmip3s0
]
infinitiu_[
    girar_vmn0000
]
conj-subord_[
    para_que_cs
]
sn_[
    espec-ms_[
        el_da0ms0
    ]
    grup-nom-ms_[
        sol_ncms000
    ]
]
grup-verb_[
    rompa_vmsp3s0
]
grup-sp_[
    prep_[
        contra_sps00
    ]
    sn_[
        él_pp3ms000
    ]
]
sn_[
    espec-mp_[
        sus_dp3cp0
    ]
    grup-nom-mp_[
        rayos_ncmp000
    ]
```

```
    ]
    ._Fp
]
S_[
    grup-verb_[
        Resulta_vmip3s0
    ]
    sadv_[
        entonces_rg
    ]
    conj-subord_[
        que_cs
    ]
    neg_[
        no_rn
    ]
    grup-verb_[
        practica_vmip3s0
    ]
    sn_[
        espec-fp_[
            las_da0fp0
        ]
        grup-nom-fp_[
            historias_ncfp000
            sp-de_[
                del_spcms
                grup-nom-ms_[
                    espejo_ncms000
                ]
            ]
        ]
    ]
    ,_Fc
    grup-sp_[
        del_spcms
        grup-nom-ms_[
            tomavistas_ncmn000
        ]
    ]
    coord_[
        o_cc
    ]
    grup-sp_[
        prep_[
            de_sps00
        ]
        sn_[
            espec-mp_[
                todos_di0mp0
                esos_dd0mp0
            ]
            grup-nom-mp_[
                inventos_ncmp000
            ]
        ]
    ]
    relatiu_[
        que_pr0cn000
    ]
```

```
    sadv_[
        tan_rg
    ]
    sn_[
        espec-fs_[
            poca_di0fs0
        ]
        grup-nom-fs_[
            ilusión_ncfs000
        ]
    ]
    patons_[
        nos_pp1cp000
    ]
    grup-verb_[
        producen_vmip3p0
    ]
    sadv_[
        ya_rg
    ]
    ._Fp
]
S_[
    grup-verb_[
        Hace_vmip3s0
    ]
    infinitiu_[
        dar_vmn0000
    ]
    sn_[
        grup-nom-fp_[
            vueltas_ncfp000
        ]
    ]
    grup-sp_[
        prep_[
            a_sps00
        ]
        sn_[
            espec-ms_[
                su_dp3cs0
            ]
            grup-nom-ms_[
                cristal_ncms000
            ]
        ]
    ]
    coord_[
        y_cc
    ]
    sn_[
        espec-fs_[
            la_da0fs0
        ]
        grup-nom-fs_[
            luz_ncfs000
        ]
    ]
    grup-verb_[
        se_p0300000
```

```
        grup-verb_[
            descompone_vmip3s0
        ]
    ]
    coord_[
        o_cc
    ]
    grup-verb_[
        se_p0300000
        grup-verb_[
            polariza_vmip3s0
        ]
    ]
    coord_[
        o_cc
    ]
    grup-verb_[
        se_p0300000
        grup-verb_[
            desparrama_vmip3s0
        ]
    ]
    coord_[
        y_cc
    ]
    sadv_[
        entonces_rg
    ]
    sn_[
        nada_pi0cs000
    ]
    grup-verb_[
        es_vsip3s0
    ]
    grup-sp_[
        prep_[
            como_cs
        ]
        sn_[
            espec-fs_[
                la_da0fs0
            ]
            grup-nom-fs_[
                costumbre_ncfs000
                s-a-fs_[
                    tediosa_aq0fs0
                ]
            ]
        ]
    ]
    prep_[
        a_sps00
    ]
    espec-fs_[
        la_da0fs0
    ]
    relatiu_[
        que_pr0cn000
    ]
    grup-verb_[
```

```
        estamos_vmip1p0
    ]
    sa_[
        s-a-mp_[
            acostumbrados_aq0mpp
        ]
    ]
    ._Fp
]
S_[
    grup-sp_[
        prep_[
            Para_sps00
        ]
        infinitiu_[
            entendernos_vmn0000
        ]
    ]
    ,_Fc
    grup-verb_[
        diríamos_vmic1p0
    ]
    sn_[
        grup-nom-mp_[
            esperpentos_ncmp000
        ]
    ]
    ._Fp
]
S_[
    neg_[
        No_rn
    ]
    grup-verb_[
        marraríamos_vmic1p0
    ]
    sadv_[
        mucho_rg
    ]
    ._Fp
    sn_[
        "_Fe
        sn_[
            grup-nom-fp_[
                El_rey_y_el_país_con_granos_np00000
            ]
        ]
        "_Fe
    ]
    grup-verb_[
        es_vsip3s0
    ]
    sn_[
        espec-ms_[
            un_di0ms0
        ]
        grup-nom-ms_[
            s-a-ms_[
                puro_aq0ms0
            ]
```

```
            grup-nom-ms_[
                esperpento_ncms000
            ]
        ]
    ]
    ,_Fc
    coord_[
        pero_cc
    ]
    neg_[
        no_rn
    ]
    conj-subord_[
        porque_cs
    ]
    sn_[
        espec-ms_[
            el_da0ms0
        ]
        grup-nom-ms_[
            héroe_ncms000
        ]
    ]
    grup-verb_[
        se_p0300000
        grup-verb_[
            haya_vasp3s0
            enfrentado_vmp00sm
        ]
    ]
    grup-sp_[
        prep_[
            a_sps00
        ]
        sn_[
            espec-ms_[
                un_di0ms0
            ]
            grup-nom-ms_[
                espejo_ncms000
                s-a-ms_[
                    deformante_aq0cs0
                ]
            ]
        ]
    ]
    ,_Fc
    coord_[
        sino_cc
    ]
    conj-subord_[
        porque_cs
    ]
    sn_[
        espec-ms_[
            el_da0ms0
        ]
        grup-nom-ms_[
            escritor_ncms000
        ]
    ]
```

```
grup-verb_[
    va_vmip3s0
    gerundi_[
        haciendo_vmg0000
    ]
]
infinitiu_[
    guiar_vmn0000
]
sn_[
    espec-ms_[
        un_di0ms0
    ]
    grup-nom-ms_[
        berilo_ncms000
        s-a-ms_[
            clarísimo_aq0ms0
        ]
    ]
]
grup-sp_[
    prep_[
        sobre_sps00
    ]
    sn_[
        espec-fs_[
            una_di0fs0
        ]
        grup-nom-fs_[
            página_ncfs000
            sp-de_[
                prep_[
                    de_sps00
                ]
                sn_[
                    espec-fs_[
                        la_da0fs0
                    ]
                    grup-nom-fs_[
                        historia_ncfs000
                    ]
                ]
            ]
        ]
    ]
]
coord_[
    y_cc
]
espec-ms_[
    lo_da0ns0
]
relatiu_[
    que_pr0cn000
]
grup-verb_[
    resulta_vmip3s0
]
grup-verb_[
    es_vsip3s0
```

```
]
sn_[
    espec-fs_[
        una_di0fs0
    ]
    grup-nom-fs_[
        s-a-fs_[
            infinita_aq0fs0
        ]
        grup-nom-fs_[
            sarta_ncfs000
            sp-de_[
                prep_[
                    de_sps00
                ]
                sn_[
                    grup-nom-mp_[
                        s-a-mp_[
                            coherentes_aq0cp0
                        ]
                        grup-nom-mp_[
                            absurdos_ncmp000
                        ]
                    ]
                ]
            ]
        ]
    ]
]
,_Fc
conj-subord_[
    como_cs
]
conj-subord_[
    si_cs
]
sn_[
    espec-ms_[
        un_di0ms0
    ]
    grup-nom-ms_[
        diablillo_ncms000
    ]
]
sn_[
    grup-nom-ms_[
        enredador_ncms000
    ]
]
grup-verb_[
    hiciera_vmsi3s0
]
sn_[
    grup-nom-fs_[
        mofa_ncfs000
        sp-de_[
            prep_[
                de_sps00
            ]
            sn_[
                espec-mp_[
```

```
                        todos_di0mp0
                        los_da0mp0
                    ]
                    grup-nom-mp_[
                        principios_ncmp000
                        s-a-mp_[
                            solemnes_aq0cp0
                        ]
                    ]
                ]
            ]
        ]
    ]
    coord_[
        y_cc
    ]
    patons_[
        nos_pp1cp000
    ]
    grup-verb_[
        regalara_vmsi3s0
    ]
    grup-sp_[
        prep_[
            con_sps00
        ]
        sn_[
            espec-ms_[
                el_da0ms0
            ]
            grup-nom-ms_[
                s-a-ms_[
                    puro_aq0ms0
                ]
                grup-nom-ms_[
                    disparate_ncms000
                ]
            ]
        ]
    ]
    ._Fp
]
S_[
    sadv_[
        Claro_rg
    ]
    conj-subord_[
        que_cs
    ]
    sn_[
        espec-ms_[
            el_da0ms0
        ]
        grup-nom-ms_[
            disparate_ncms000
        ]
    ]
    grup-verb_[
        es_vsip3s0
    ]
```

```
infinitiu_[
    contemplar_vmn0000
]
sn_[
    espec-fs_[
        aquella_dd0fs0
    ]
    grup-nom-fs_[
        realidad_ncfs000
    ]
]
grup-sp_[
    prep_[
        con_sps00
    ]
    sn_[
        espec-fs_[
            la_da0fs0
        ]
        grup-nom-fs_[
            s-a-fs_[
                fría_aq0fs0
            ]
            grup-nom-fs_[
                lógica_ncfs000
                sp-de_[
                    del_spcms
                    grup-nom-ms_[
                        s-a-ms_[
                            adormecido_aq0msp
                        ]
                        grup-nom-ms_[
                            burgués_ncms000
                        ]
                    ]
                ]
            ]
        ]
    ]
]
conj-subord_[
    cuando_cs
]
espec-ms_[
    lo_da0ns0
]
relatiu_[
    que_pr0cn000
]
grup-verb_[
    hay_vaip3s0
    que_cs
    infinitiu_[
        hacer_vmn0000
    ]
]
grup-verb_[
    es_vsip3s0
]
infinitiu_[
```

```
    detener_vmn0000
]
sn_[
    espec-mp_[
        los_da0mp0
    ]
    grup-nom-mp_[
        rayos_ncmp000
        s-a-mp_[
            disparados_aq0mpp
        ]
    ]
]
grup-sp_[
    prep_[
        desde_sps00
    ]
    sn_[
        espec-ms_[
            el_da0ms0
        ]
        grup-nom-ms_[
            perfil_ncms000
            sp-de_[
                prep_[
                    de_sps00
                ]
                sn_[
                    espec-fp_[
                        las_da0fp0
                    ]
                    grup-nom-fp_[
                        aristas_ncfp000
                    ]
                ]
            ]
        ]
    ]
]
coord_[
    y_cc
]
neg_[
    no_rn
]
infinitiu_[
    mirar_vmn0000
]
sn_[
    espec-fp_[
        las_da0fp0
    ]
    grup-nom-fp_[
        figuras_ncfp000
    ]
]
relatiu_[
    que_pr0cn000
]
sn_[
```

```
        espec-fs_[
            la_da0fs0
        ]
        grup-nom-fs_[
            hoja_ncfs000
            sp-de_[
                prep_[
                    de_sps00
                ]
                sn_[
                    grup-nom-ms_[
                        papel_ncms000
                    ]
                ]
            ]
        ]
    ]
    grup-verb_[
        tenía_vmii3s0
    ]
    sa_[
        s-a-fp_[
            impresas_aq0fpp
        ]
    ]
    ._Fp
]
S_[
    sn_[
        espec-fs_[
            La_da0fs0
        ]
        grup-nom-fs_[
            historia_ncfs000
        ]
    ]
    grup-verb_[
        existió_vmis3s0
    ]
    coord_[
        y_cc
    ]
    grup-verb_[
        vale_vmip3s0
    ]
    grup-sp_[
        prep_[
            como_cs
        ]
        s-a-fs_[
            referente_aq0cs0
        ]
    ]
    ,_Fc
    coord_[
        pero_cc
    ]
    sn_[
        espec-mp_[
            los_da0mp0
```

```
        ]
        grup-nom-mp_[
            grabados_ncmp000
            sp-de_[
                prep_[
                    de_sps00
                ]
                sn_[
                    grup-nom-mp_[
                        colorines_ncmp000
                    ]
                ]
            ]
        ]
    ]
    grup-verb_[
        son_vsip3p0
    ]
    sn_[
        espec-fs_[
            otra_di0fs0
        ]
        grup-nom-fs_[
            cosa_ncfs000
        ]
    ]
    conj-subord_[
        que_cs
    ]
    sn_[
        espec-ms_[
            el_da0ms0
        ]
        grup-nom-ms_[
            s-a-ms_[
                solemne_aq0cs0
            ]
            grup-nom-ms_[
                testimonio_ncms000
                sp-de_[
                    prep_[
                        de_sps00
                    ]
                    sn_[
                        espec-fs_[
                            su_dp3cs0
                        ]
                        grup-nom-fs_[
                            presencia_ncfs000
                        ]
                    ]
                ]
            ]
        ]
    ]
    ._Fp
]
```

Apéndice F

Corpus CLiC-TALP anotado sintácticamente

En este apéndice se presenta el primer fragmento del corpus CLiC-TALP tras la anotación sintáctica manual.

```
(
  (S
    (sn
      (grup.nom.ms
        (np00000 Medardo_Fraile)))
    (gv
      (vmip3s0 juega))
    (sp
      (prep
        (sps00 a))
      (sn
        (espec.ms
          (di0ms0 un))
        (grup.nom.ms
          (ncms000 cinismo)
          (s.a.ms.co
            (s.a.ms
              (aq0cs0 fácil))
            (coord
              (cc y))
            (S.NF.P
              (aq0msp divertido))))))
    (Fp .)))
(
  (S.co
    (S
      (sn.e *0*)
      (neg
        (rn No))
      (gv
        (vmip1s0 quiero))
      (S.NF.C
        (infinitiu
          (vmn0000 decir))
        (S.F.C
          (conj.subord
```

```
          (cs que))
        (sn
          (grup.nom
            (pp3cna00 lo)))
        (gv
          (vsm03s0 sea))
        (sa.co
          (Fc ,)
          (sa
            (aq0ms0 cínico))
          (coord
            (cc o))
          (S.NF.P
            (aq0msp divertido))
          (Fc ,))))))
  (coord
    (cc sino_que))
  (S.co
    (sp
      (prep
        (sps00 ante))
      (sn
        (espec.ms
          (di0ms0 un))
        (grup.nom.ms
          (ncms000 mazo)
          (sp
            (prep
              (sps00 de))
            (sn
              (grup.nom.fp
                (ncfp000 hojas)
                (S.NF.P
                  (aq0fpp grabadas))))))))
    (S
      (sn.e *0*)
      (gv
        (vmip3s0 coloca))
      (sn
        (espec.ms
          (di0ms0 un))
        (grup.nom.ms
          (ncms000 cristal)
          (S.NF.P
            (sadv
              (rg bien))
            (aq0msp tallado)))))
    (coord
      (cc y))
    (S
      (sn.e *0*)
      (sn
        (grup.nom.ms
          (pp3msa00 lo)))
      (gv
        (vmip3s0 hace)
        (infinitiu
```

```
            (vmn0000 girar)))
          (S.F.A
            (conj.subord
              (cs para_que))
            (sn
              (espec.ms
                (da0ms0 el))
              (grup.nom.ms
                (ncms000 sol)))
            (gv
              (vmsp3s0 rompa))
            (sp
              (prep
                (sps00 contra))
              (sn
                (grup.nom.ms
                  (pp3ms000 él))))
            (sn
              (espec.mp
                (dp3cp0 sus))
              (grup.nom.mp
                (ncmp000 rayos))))))
    (Fp .)))
(
  (S
    (gv
      (vmip3s0 Resulta))
    (sadv
      (rg entonces))
    (S.F.C
      (conj.subord
        (cs que))
      (sn.e *0*)
      (neg
        (rn no))
      (gv
        (vmip3s0 practica))
      (sn
        (espec.fp
          (da0fp0 las))
        (grup.nom.fp
          (ncfp000 historias)
          (sp.co
            (sp
              (prep
                (spcms del))
              (sn
                (grup.nom.ms
                  (ncms000 espejo))))
            (Fc ,)
            (sp
              (prep
                (spcms del))
              (sn
                (grup.nom.ms
                  (ncmn000 tomavistas))))
```

```
            (coord
              (cc o))
            (sp
              (prep
                (sps00 de))
              (sn
                (espec.mp
                  (di0mp0 todos)
                  (dd0mp0 esos))
                (grup.nom.mp
                  (ncmp000 inventos)
                  (S.F.R
                    (relatiu
                      (pr0cn000 que))
                    (sn
                      (espec.fs
                        (sadv
                          (rg tan))
                        (di0fs0 poca))
                      (grup.nom.fs
                        (ncfs000 ilusión)))
                    (sn
                      (grup.nom.p
                        (pp1cp000 nos)))
                    (gv
                      (vmip3p0 producen))
                    (sadv
                      (rg ya)))))))))
    (Fp .)))
(
  (S.co
    (S.co
      (S
        (sn.e *0*)
        (gv
          (vmip3s0 Hace)
          (infinitiu
            (vmn0000 dar)))
        (sn
          (grup.nom.fp
            (ncfp000 vueltas)))
        (sp
          (prep
            (sps00 a))
          (sn
            (espec.ms
              (dp3cs0 su))
            (grup.nom.ms
              (ncms000 cristal)))))
      (coord
        (cc y))
      (S.co
        (S
          (sn
            (espec.fs
              (da0fs0 la))
            (grup.nom.fs
```

```
            (ncfs000 luz)))
          (morf.pron
            (p0300000 se))
          (gv
            (vmip3s0 descompone)))
        (coord
          (cc o))
        (S
          (sn.e *0*)
          (morf.pron
            (p0300000 se))
          (gv
            (vmip3s0 polariza)))
        (coord
          (cc o))
        (S
          (sn.e *0*)
          (morf.pron
            (p0300000 se))
          (gv
            (vmip3s0 desparrama)))))
    (coord
      (cc y))
    (S
      (sadv
        (rg entonces))
      (sn
        (grup.nom.ms
          (pi0cs000 nada)))
      (gv
        (vsip3s0 es))
      (sp
        (prep
          (cs como))
        (sn
          (espec.fs
            (da0fs0 la))
          (grup.nom.fs
            (ncfs000 costumbre)
            (s.a.fs
              (aq0fs0 tediosa))
            (S.F.R
              (sp
                (prep
                  (sps00 a))
                (sn
                  (espec.fs
                    (da0fs0 la))
                  (relatiu
                    (pr0cn000 que))))
              (sn.e *0*)
              (gv
                (vmip1p0 estamos))
              (S.NF.P
                (aq0mpp acostumbrados)))))))
    (Fp .)))
(
```

```
  (S
    (sp
      (prep
        (sps00 Para))
      (S.NF.C
        (infinitiu
          (vmn0000 entendernos)))
      (Fc ,))
    (sn.e *0*)
    (gv
      (vmic1p0 diríamos))
    (sn
      (grup.nom.mp
        (ncmp000 esperpentos)))
    (Fp .)))
(
  (S
    (sn.e *0*)
    (neg
      (rn No))
    (gv
      (vmic1p0 marraríamos))
    (sadv
      (rg mucho))
    (Fp .)))
(
  (S.co
    (S.co
      (S
        (sn
          (grup.nom.ms
            (Fe ")
            (np00000 El_rey_y_el_país_con_granos)
            (Fe ")))
        (gv
          (vsip3s0 es))
        (sn
          (espec.ms
            (di0ms0 un))
          (grup.nom.ms
            (s.a.ms
              (aq0ms0 puro))
            (ncms000 esperpento)))
        (Fc ,))
      (coord
        (cc pero))
      (S*
        (neg
          (rn no))
        (S.F.A.co
          (S.F.A
            (conj.subord
              (cs porque))
            (sn
              (espec.ms
                (da0ms0 el))
              (grup.nom.ms
                (ncms000 héroe)))
```

```
          (morf.pron
            (p0300000 se))
          (gv
            (vasp3s0 haya)
            (vmp00sm enfrentado))
          (sp
            (prep
              (sps00 a))
            (sn
              (espec.ms
                (di0ms0 un))
              (grup.nom.ms
                (ncms000 espejo)
                (s.a.ms
                  (aq0cs0 deformante)))))
          (Fc ,))
        (coord
          (cc sino))
        (S.F.A
          (conj.subord
            (cs porque))
          (sn
            (espec.ms
              (da0ms0 el))
            (grup.nom.ms
              (ncms000 escritor)))
          (gv
            (vmip3s0 va)
            (gerundi
              (vmg0000 haciendo)
              (infinitiu
                (vmn0000 guiar))))
          (sn
            (espec.ms
              (di0ms0 un))
            (grup.nom.ms
              (ncms000 berilo)
              (s.a.ms
                (aq0ms0 clarísimo))))
          (sp
            (prep
              (sps00 sobre))
            (sn
              (espec.fs
                (di0fs0 una))
              (grup.nom.fs
                (ncfs000 página)
                (sp
                  (prep
                    (sps00 de))
                  (sn
                    (espec.fs
                      (da0fs0 la))
                    (grup.nom.fs
                      (ncfs000 historia)))))))))
(coord
  (cc y))
```

```
(S
  (sn
    (espec.ms
      (da0ns0 lo))
    (S.F.R
      (relatiu
        (pr0cn000 que))
      (gv
        (vmip3s0 resulta))))
  (gv
    (vsip3s0 es))
  (sn
    (espec.fs
      (di0fs0 una))
    (grup.nom.fs
      (s.a.fs
        (aq0fs0 infinita))
      (ncfs000 sarta)
      (sp
        (prep
          (sps00 de))
        (sn
          (grup.nom.mp
            (s.a.mp
              (aq0cp0 coherentes))
            (ncmp000 absurdos))))))
  (sp
    (Fc ,)
    (conj.subord
      (cs como))
    (S.F.C.co
      (conj.subord
        (cs si))
      (S.F.C
        (sn
          (espec.ms
            (di0ms0 un))
          (grup.nom.ms
            (ncms000 diablillo)
            (s.a.ms
              (aq0ms0 enredador))))
        (gv
          (vmsi3s0 hiciera))
        (sn
          (grup.nom.fs
            (ncfs000 mofa)))
        (sp
          (prep
            (sps00 de))
          (sn
            (espec.mp
              (di0mp0 todos)
              (da0mp0 los))
            (grup.nom.mp
              (ncmp000 principios)
              (s.a.mp
                (aq0cp0 solemnes))))))))
```

```
          (coord
            (cc y))
          (S.F.C
            (sn.e *0*)
            (sn
              (grup.nom.p
                (pp1cp000 nos)))
            (gv
              (vmsi3s0 regalara))
            (sp
              (prep
                (sps00 con))
              (sn
                (espec.ms
                  (da0ms0 el))
                (grup.nom.ms
                  (s.a.ms
                    (aq0ms0 puro))
                  (ncms000 disparate))))))))
    (Fp .)))
(
  (S*
    (sadv
      (rg Claro))
    (S.F.C
      (conj.subord
        (cs que))
      (sn
        (espec.ms
          (da0ms0 el))
        (grup.nom.ms
          (ncms000 disparate)))
      (gv
        (vsip3s0 es))
      (S.NF.C
        (infinitiu
          (vmn0000 contemplar))
        (sn
          (espec.fs
            (dd0fs0 aquella))
          (grup.nom.fs
            (ncfs000 realidad)))
        (sp
          (prep
            (sps00 con))
          (sn
            (espec.fs
              (da0fs0 la))
            (grup.nom.fs
              (s.a.fs
                (aq0fs0 fría))
              (ncfs000 lógica)
              (sp
                (prep
                  (spcms del))
                (sn
                  (grup.nom.ms
```

```
              (S.NF.P
                (aq0msp adormecido))
              (ncms000 burgués))))))))
  (S.F.A
    (conj.subord
      (cs cuando))
    (sn
      (espec.ms
        (da0ns0 lo))
      (S.F.R
        (relatiu
          (pr0cn000 que))
        (gv
          (vaip3s0 hay)
          (cs que)
          (infinitiu
            (vmn0000 hacer)))))
    (gv
      (vsip3s0 es))
    (S.NF.C.co
      (S.NF.C
        (infinitiu
          (vmn0000 detener))
        (sn
          (espec.mp
            (da0mp0 los))
          (grup.nom.mp
            (ncmp000 rayos)
            (S.NF.P
              (aq0mpp disparados)
              (sp
                (prep
                  (sps00 desde))
                (sn
                  (espec.ms
                    (da0ms0 el))
                  (grup.nom.ms
                    (ncms000 perfil)
                    (sp
                      (prep
                        (sps00 de))
                      (sn
                        (espec.fp
                          (da0fp0 las))
                        (grup.nom.fp
                          (ncfp000 aristas)))))))))))
      (coord
        (cc y))
      (S.NF.C
        (neg
          (rn no))
        (infinitiu
          (vmn0000 mirar))
        (sn
          (espec.fp
            (da0fp0 las))
          (grup.nom.fp
```

```
                (ncfp000 figuras)
                (S.F.R
                  (relatiu
                    (pr0cn000 que))
                  (sn
                    (espec.fs
                      (da0fs0 la))
                    (grup.nom.fs
                      (ncfs000 hoja)
                      (sp
                        (prep
                          (sps00 de))
                        (sn
                          (grup.nom.ms
                            (ncms000 papel))))))
                  (gv
                    (vmii3s0 tenía))
                  (S.NF.P
                    (aq0fpp impresas)))))))))
    (Fp .)))
(
  (S.co
    (S.co
      (S
        (sn
          (espec.fs
            (da0fs0 La))
          (grup.nom.fs
            (ncfs000 historia)))
        (gv
          (vmis3s0 existió)))
      (coord
        (cc y))
      (S
        (sn.e *0*)
        (gv
          (vmip3s0 vale))
        (sp
          (prep
            (cs como))
          (sn
            (grup.nom.ms
              (ncms000 referente)))))
      (Fc ,))
    (coord
      (cc pero))
    (S
      (sn
        (espec.mp
          (da0mp0 los))
        (grup.nom.mp
          (ncmp000 grabados)
          (sp
            (prep
              (sps00 de))
            (sn
              (grup.nom.mp
```

```
            (ncmp000 colorines))))))
    (gv
      (vsip3p0 son))
    (sn
      (espec.ms
        (rg otra_cosa_que)
        (da0ms0 el))
      (grup.nom.ms
        (s.a.ms
          (aq0cs0 solemne))
        (ncms000 testimonio)
        (sp
          (prep
            (sps00 de))
          (sn
            (espec.fs
              (dp3cs0 su))
            (grup.nom.fs
              (ncfs000 presencia)))))))
  (Fp .)))
```