Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios

Rodrigo Alarcón Martínez

TESIS DOCTORAL UPF / 2009

DIRECTORES DE LA TESIS

Dra. Carme Bach Martorell (Departament de Traducció i Ciències del Llenguatge, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra)

Dr. Gerardo Sierra Martínez (Grupo de Ingeniería Lingüística, Instituto de Ingeniería, Universidad Nacional Autónoma de México)

UNIVERSITAT POMPEU FABRA

Für Nadine

Agradecimientos

Agradezco el apoyo, la orientación y la paciencia de los directores de esta tesis, Carme Bach y Gerardo Sierra, así como todo el soporte que me han otorgado los miembros del Instituto Universitario de Lingüística Aplicada y del Grupo de Ingeniería Lingüística.

Agradezco igualmente a mi madre, mi familia y mis amigos por su incondicional apoyo.

Esta tesis fue posible gracias a las siguientes becas:

- Beca de estudios de posgrado, Consejo Nacional de Ciencia y Tecnología (CONACYT), México. Referencia 179210.
- Beca de estudios de doctorado, Consejo Nacional de Ciencia y Tecnología (CONACYT), México. Proyecto 82050.
- Beca de estudios de posgrado, Fundación Carolina, España.

Resumen

El desarrollo de herramientas computacionales de ayuda en tareas lexicográficas y terminográficas es un creciente interés dentro del campo del procesamiento del lenguaje natural. Algunas herramientas se han desarrollado para la extracción automática de términos en textos especializados. Además, algunos estudios se han enfocado en el desarrollo de métodos para adquirir conocimiento definitorio sobre términos, tomando en cuenta la idea de que los términos suelen definirse en estructuras denominadas *contextos definitorios*, donde se emplean una serie de patrones que pueden ser reconocidos automáticamente.

Partiendo de esta premisa, en esta tesis presentamos un sistema basado en reglas lingüísticas para la extracción de contextos definitorios sobre textos especializados en español. Este sistema esta formado por cuatro procesos: la extracción de ocurrencias de patrones definitorios; el filtro de contextos no relevantes; la identificación de los elementos constitutivos en los candidatos, es decir, el término y la definición; y la organización jerárquica de los resultados con el fin de encontrar los mejores candidatos.

Abstract

The development of computational tools to help on lexicography and terminography tasks is indeed a growing interest on natural language processing field. Some tools have been developing for the extraction of terms from specialised texts. In addition, some studies have been focused on developing methods for acquiring definitional knowledge about terms, considering that the terms are commonly defined in structures called *definitional contexts*, which employ some patterns that can be automatically identified.

Taking into account this premise, in this thesis we present a system based on linguistic rules for the extraction of definitional contexts in Spanish specialised texts. The system includes four processes: the extraction of textual fragments with definitional patterns; the filtering of non-relevant contexts; the identification of the constituent elements in the candidates, i.e., terms and definitions; and the ranking of the results to identify the best candidates.

Índice

Lista de figuras

Lista de tablas

1. Introducción

Uno de los problemas básicos de cualquier área de conocimiento es la organización y descripción de sus conceptos. En cualquier campo, el registro de su terminología es un proceso indispensable para la comunicación del conocimiento que involucra. Es indispensable, además, para la enseñanza y aprendizaje de dicha área, al igual que para la interacción con otros campos. En este sentido, la terminografía ocupa un lugar importante en la resolución de este problema, ya que se encarga, por un lado, de la elaboración de ontologías donde se representa la red conceptual de un área específica, y por otro lado, de la elaboración de diccionarios donde se explica el significado de los términos.

El avance en el desarrollo de herramientas computacionales ha facilitado en gran medida el trabajo terminográfico y ha provisto de herramientas para la compilación de corpus lingüísticos especializados, donde se almacena digitalmente una gran cantidad de documentos técnicos, así como sistemas para la explotación de dichos corpus, como son los sistemas para la extracción automática de términos. Actualmente, existe además un creciente interés por el desarrollo de sistemas para la extracción automática de información que sea útil para describir el significado de los términos. Diversos estudios han coincidido en la idea de que, cuando un autor de un texto especializado define un término, lo hace mediante estructuras sintácticas recurrentes comúnmente denominadas *contextos definitorios* (CDs). A grandes rasgos, los CDs suelen estar constituidos por un término y una definición, los cuales, a su vez, suelen estar conectados por cierto tipo de patrones definitorios que pueden ser reconocidos automáticamente.

Si bien el desarrollo de sistemas para la extracción automática de CDs no ha alcanzado el nivel de desarrollo que ha tenido el de los sistemas para la extracción de términos, en la actualidad se cuenta con las bases teóricas y metodológicas que permiten la elaboración de metodologías para la extracción de información definitoria.

A partir de esta premisa, en esta tesis presentamos una metodología para el desarrollo de un sistema de extracción automática de CDs en textos especializados en español, basado principalmente en la

búsqueda de patrones definitorios. Un sistema que sólo obtuviera las ocurrencias de dichos patrones ya constituiría por sí mismo una herramienta de ayuda en el proceso terminográfico enfocado en la extracción de información relevante sobre términos. No obstante, el análisis manual de dichas ocurrencias supondría todavía un esfuerzo que podría simplificarse si el sistema contara con un procesamiento automático de las ocurrencias. En este sentido, el sistema que proponemos incluye no sólo la extracción de ocurrencias de patrones definitorios, sino también el filtrado automático de ocurrencias no relevantes, la identificación automática de los elementos constitutivos, es decir, los términos y las definiciones, y una clasificación posterior para tratar de determinar los mejores candidatos. Cabe señalar que, como veremos a lo largo de esta tesis, la metodología aquí propuesta no parte de la identificación automática previa de los términos que se quieren definir, sino de la búsqueda de ocurrencias de patrones definitorios y la posterior identificación automática de los términos y definiciones en dichas ocurrencias.

Un sistema de esta naturaleza tiene como principal campo de aplicación el ámbito terminográfico y el conocimiento especializado, ya que, principalmente, serviría para la elaboración de glosarios o diccionarios especializados, bases de datos de conocimiento léxico o bien para la elaboración de ontologías. No obstante, por su misma naturaleza de extracción de conocimiento especializado, el ámbito de aplicación también se extiende a áreas como la traducción especializada, en tanto que un traductor especializado necesita en algún momento saber el significado de un término particular, o bien a la misma enseñanza de cualquier dominio especializado, pensando en la necesidad de obtener información relevante sobre un término determinado.

1.1 Antecedentes

Esta investigación tiene como origen el trabajo realizado en el Grupo de Ingeniería Lingüística (GIL) de la Universidad Nacional Autónoma de México, donde una de las principales investigaciones se basa en el desarrollo de diccionarios onomasiológicos, esto es, diccionarios donde se puede encontrar un término a partir de la descripción del concepto.

Como parte de este proyecto y para facilitar el trabajo terminográfico que implica, se tiene contemplada la adquisición automática de términos y definiciones. Para lograr este objetivo, una parte de las investigaciones del GIL se ha centrado en el estudio de la extracción automática de CDs.

Este estudio constituye una continuación de una investigación realizada en el GIL a nivel de licenciatura, donde el tema central fue el análisis lingüístico de CDs en textos especializados en español. Constituye también una continuación de un primer acercamiento a la extracción automática de CDs, elaborada como proyecto de tesis en el marco del Doctorado de Ciencias del Lenguaje y Lingüística Aplicada, de la Universidad Pompeu Fabra.

1.2 Objetivos

El objetivo general de esta tesis es desarrollar una metodología para la extracción automática de CDs en un corpus anotado morfosintácticamente.

Para ello se han planteado los siguientes objetivos particulares:

1. Revisión de los principales conceptos involucrados en la investigación desde una perspectiva terminográfica: *contexto definitorio*, *término*, *definición* y *patrón definitorio*.
2. Revisión de estudios previos y trabajos en curso relacionados con la extracción automática de conocimiento definitorio.
3. Búsqueda, análisis y evaluación de patrones verbales recurrentes en CDs en español.
4. Desarrollo de una metodología para la extracción automática de CDs.
5. Evaluación de la metodología propuesta para la extracción automática de CDs.

1.3 Objeto de estudio

El objeto de estudio de esta tesis son CDs que incluyen algún patrón verbal definitorio asociado a alguno de los siguientes tipos de definición: *analítica*, *extensional*, *funcional* y *sinonímica*.

1.4 Hipótesis

Esta investigación parte de la premisa de que en textos especializados se pueden extraer automáticamente CDs en los cuales los autores definen términos mediante el uso de patrones definitorios. A partir de esta premisa las hipótesis aquí planteadas son:

1. Los candidatos a CDs pueden ser extraídos automáticamente a partir de la búsqueda de las ocurrencias de patrones definitorios.
2. Es posible filtrar automáticamente excepciones en las ocurrencias encontradas.
3. Se pueden identificar automáticamente los elementos constitutivos de los candidatos a CDs. En esta identificación, a su vez, es importante notar lo siguiente:
   3.1 Los elementos constitutivos, es decir términos y definiciones, suelen seguir patrones de formación sintáctica.
   3.2 Los elementos constitutivos suelen ocupar una posición recurrente dependiendo del patrón definitorio que los conecta.
4. Se pueden clasificar automáticamente los resultados obtenidos a partir de reglas lingüísticas para determinar los mejores candidatos a CDs.

1.5 Estructura de la tesis

Esta tesis está organizada de la siguiente manera:

- En el capítulo 2 se presentarán algunos principios básicos en la organización y descripción del conocimiento especializado. Se abordarán los conceptos de lenguaje especializado, término y definición, y se detallarán brevemente algunas diferencias entre terminología y terminografía. Asimismo, se abordará el tema de la metodología de la práctica terminográfica actual, para dar paso a una descripción de los conceptos de contexto definitorio y de relaciones semánticas. Finalmente, en este capítulo se tratará el tema de las definiciones en textos especializados.

- En el capítulo 3 nos centraremos en el estado del arte a partir de la descripción de estudios relacionados con la extracción automática de CDs. Comenzaremos por explicar ciertas particularidades de los conceptos de conocimiento definitorio y extracción de información. A continuación abordaremos algunas consideraciones sobre la diferencia entre extracción de relaciones semánticas y extracción de CDs, para finalmente detallar algunos estudios enfocados en la extracción de información definitoria.
- El capítulo 4 constituye un estudio lingüístico de CDs en textos de especialidad. Aquí describiremos, en primer lugar, las diferencias entre los conceptos de contexto y contexto definitorio. Enseguida detallaremos algunos conceptos que se han tratado en el ámbito de la terminografía y de la extracción de conocimiento definitorio: los enunciados de interés definitorio, los actos performativos definitorios y los contextos ricos en conocimiento. Detallaremos asimismo algunas generalidades a tomar en cuenta para una definición aplicada del concepto de contexto definitorio. Por último, presentaremos un análisis lingüístico de CDs en español, así como una tipología semántica de dichos contextos pertinente para nuestra investigación.
- En el capítulo 5 describiremos el sistema propuesto en esta tesis. Detallaremos en primer lugar el corpus de pruebas utilizado. Enseguida presentaremos una descripción general del algoritmo, para finalmente explicar los diferentes procesos que incluye el sistema. Dichos procesos están relacionados con la extracción de candidatos a partir de la búsqueda de patrones definitorios y su procesamiento automático para filtrar contextos no relevantes, identificar el término y la definición, así como organizarlos de acuerdo con una mayor probabilidad de relevancia.
- En el capítulo 6 detallaremos la evaluación del sistema. Describiremos la obtención de nuestro corpus de evaluación, las métricas utilizadas y la metodología general, para finalmente presentar los resultados obtenidos.
- Por último, en el capítulo 7 haremos una recapitulación de la tesis, presentaremos las principales aportaciones y detallaremos el trabajo futuro relacionado con nuestra investigación.

2. Algunas consideraciones sobre la organización y descripción del conocimiento especializado

En este capítulo abordaremos los conceptos básicos necesarios para entender el planteamiento general de nuestra investigación. Partimos de una descripción general del lenguaje especializado (2.1), para después especificar los conceptos de término y definición (2.2). Enseguida abordaremos algunas consideraciones sobre la diferencia entre terminología y terminografía (2.3). Lo anterior nos dará paso a detallar la metodología de la práctica terminográfica actual (2.4). Explicaremos básicamente los conceptos de CDs y relaciones semánticas (2.5), y por último, detallaremos la forma en que suelen realizarse las definiciones en textos de especialidad (2.6).

2.1 El lenguaje especializado

Una característica implícita en las sociedades humanas es que existen actividades particulares que se demarcan de las demás porque conllevan un grado de conocimiento específico que no poseen todos los individuos. El *conocimiento* puede ser entendido como un conjunto de ideas y experiencias que forman representaciones mentales de la realidad, y dado que en una sociedad ningún individuo posee la estructura total del conocimiento de su comunidad, el conocimiento se divide en áreas temáticas o disciplinas, derivando en la creación de subespacios del espacio del conocimiento (Sager 1993: 39).

Se habla de *conocimiento especializado* para referirse a aquél asociado a las actividades de un área específica, frente al *conocimiento general* que se encuentra presente en una mayor gama de situaciones que comparten todos los individuos de una comunidad. No obstante esta categorización, el conocimiento se modela según las situaciones comunicativas específicas en las que se adquiere o emplea, y se encuentra por lo tanto en un constante proceso de comunicación. Es a partir de la comunicación del conocimiento que se pueden establecer las pautas necesarias para reconocer las relaciones y diferencias entre las áreas de

conocimiento especializado, así como demarcar fronteras entre éstas y aquellas que pertenecen al campo del conocimiento general.

De manera análoga a la distinción que puede hacerse entre el conocimiento general y el conocimiento especializado, existe una diferencia entre el lenguaje empleado en situaciones comunicativas generales y el lenguaje que se utiliza en la comunicación del conocimiento de un área específica. El lenguaje que se emplea en situaciones comunicativas cuya finalidad es la transmisión del conocimiento que comparten los miembros de una colectividad se conoce como *lenguaje especializado*.

Es difícil establecer una diferencia clara entre el lenguaje especializado y el lenguaje general. Sin embargo, existen ciertos factores que pueden ayudar a la comprensión de las diferencias básicas entre ambas entidades. En esta distinción debe tenerse en cuenta tanto el uso lingüístico particular que se hace del lenguaje como el contexto comunicativo en el que se emplea (Pérez 2002: 3.4.1). Los individuos se comunican de manera distinta en diferentes situaciones, y la forma en que expresan su conocimiento depende tanto del contexto y la situación en la que lo hacen, como del tipo de conocimiento que expresan (Pearson 1998: 26).

Para Cabré (1993: 128-129), el lenguaje especializado hace referencia al: "conjunto de subcódigos –parcialmente coincidentes con el subcódigo de la lengua común– caracterizados en virtud de unas peculiaridades 'especiales', esto es, propias y específicas de cada uno de ellos, como pueden ser la temática, el tipo de interlocutores, la situación comunicativa, la intención del hablante, el medio en que se produce un intercambio comunicativo, el tipo de intercambio, etc.". Además, añade esta autora, el lenguaje especializado se puede distinguir del lenguaje general en el tipo de expresiones escritas u orales que producen los expertos de un área determinada.

Los expertos pueden ser considerados como personas que tienen experiencia en un campo específico, aunque dicho campo no tiene que ser necesariamente un área técnica o científica (Bowker y Pearson 2002: 27). Estos últimos autores señalan que las situaciones comunicativas que se transmiten a través de las distintas

expresiones escritas u orales del lenguaje especializado pueden ser clasificadas en 3 tipos:

1. La comunicación entre expertos
2. La comunicación entre expertos y semi-expertos
3. La comunicación entre expertos y no-expertos

En el primer tipo se utiliza un lenguaje altamente especializado debido al bagaje de conocimiento particular que comparten y entienden dichos expertos. En el segundo caso, los semi-expertos son por ejemplo estudiantes o expertos de campos de conocimiento relacionados, y en este tipo de comunicación es probable que se utilice el mismo lenguaje altamente especializado pero acompañado de explicaciones o referencias en lenguaje general. En el tercer caso, los no-expertos son aquellas personas que por diferentes motivos se ven implicadas en las situaciones comunicativas de un determinado campo, por ejemplo estudiantes, y en este caso se empleará el lenguaje general en la medida en que sea posible.

Tomando en cuenta que cada área de conocimiento posee un conjunto léxico que la caracteriza frente a las demás, en cada situación comunicativa se utilizará en mayor o menor medida un vocabulario específico. Existirá además una mayor o menor necesidad de explicar dicho conjunto léxico dependiendo de la situación comunicativa: mientras que la comunicación entre los mismos expertos da por sentado muchos de los conceptos que se incluyen en su campo de conocimiento, la comunicación con los semi-expertos y no-expertos será lo más explicativa posible, utilizando recurrentemente definiciones donde se aclare el significado del conjunto léxico particular.

2.2 Términos y definiciones

Ahora bien, las unidades léxicas que hacen referencia a objetos específicos del ámbito de una disciplina se conocen como *términos*, y el conjunto de términos de un área conforma su *terminología* (Sager 1993: 43). Los términos constituyen unidades cuyos valores son relativos a un conjunto y cuyo propósito es establecer, mediante un proceso de denominación, una relación biunívoca con un concepto (Rey 1995: 136). En este sentido, los *conceptos* se

entienden a su vez como construcciones mentales de los procesos cognoscitivos humanos a través de los cuales se clasifican los objetos mediante la abstracción arbitraria o sistemática y se utilizan con el fin de estructurar el conocimiento y la apreciación del mundo (Sager 1993: 47).

En el proceso denominativo los especialistas asignan formas lingüísticas específicas para referirse a un concepto que pertenece a la estructura nocional de su área (Cabré 1993: 91-92). Las denominaciones pueden usarse para designar objetos materiales o no materiales, y lo que se vincula a la representación mental del término es una abstracción, "una generalización basada en las experiencias que hemos ido acumulando en contacto con el mundo que nos rodea" (Arntz y Picht 1995: 58-59).

Los términos, entendidos como denominaciones especializadas, son entonces los nombres que se utilizan para designar conceptos, tomando en cuenta que "son unidades usadas en la comunicación especializada para designar los 'objetos' de una realidad preexistente" (Cabré 1993: 169).

En esta investigación, el concepto de término lo enmarcamos dentro del contexto específico de la Teoría Comunicativa de la Terminología (TCT), que detallaremos en el siguiente apartado. Por ahora, basta aclarar que en la TCT "los términos son unidades léxicas activadas singularmente por sus condiciones pragmáticas de adecuación a un tipo de comunicación". (Cabré 1999: 132). Es así que en este trabajo consideraremos a un término como una unidad léxica especializada cuyo significado se encuentra relacionado con un área de conocimiento particular. Si seguimos este enfoque, palabras que podrían suponer un grado de especialización completamente diferente, como *mar* o *trombocitopemia*, serán entonces unidades especializadas en tanto su significado se encuentra relacionado con un ámbito de conocimiento específico, en este caso con el ámbito del medio ambiente y la medicina.

Por su parte, la definición es "una descripción lingüística de un concepto, basada en el listado de un número de características que transmiten el significado del concepto" (Sager 1993: 68). Las definiciones pueden ser entendidas como ecuaciones, en las cuales el lado izquierdo corresponde al concepto expresado mediante una

denominación, y el lado derecho corresponde a la descripción de la comprensión del concepto (Arntz y Picht 1995: 68). Las definiciones conforman un vínculo entre los conceptos y los términos a través de una ecuación en la cual el término es la incógnita, y donde mediante el acto de definir se fija entonces la referencia exacta de un término a un concepto (Sager, 1993: 45, 68).

La definición, en un lenguaje especializado, trata precisamente de dar cuenta de las propiedades inherentes a las unidades léxicas de dicho lenguaje, con el fin de describirlas y delimitarlas frente a otras unidades de su estructura conceptual. En este sentido, la definición se contempla como:

> "un recurso textual privilegiado para la representación de los significados de las palabras. Consiste en una redacción simple que puede situar un significado dentro de una categoría más amplia y puede reflejar las características básicas para que, por la experiencia o por el conocimiento adquirido, podamos aprehenderlo, podamos relacionarlo con algún referente, o podamos identificarlo frente a otros significados de la lengua" (Lorente 2001: 104).

De esta forma, en este trabajo consideramos que las definiciones representan los significados que se asignan a unidades léxicas y a través de los cuales se esclarecen los conceptos que representan los términos.

Es importante notar que entre conceptos, términos y definiciones existe una relación ineludible que juega un rol importante en la organización y descripción del conocimiento especializado. Los términos, además de constituir la base de la comunicación especializada escrita y oral, permiten organizar el conocimiento de cada área, ya que a través de su estudio es posible entonces la estructuración y delimitación del conocimiento, en tanto que los términos proyectan las características de organización de cada campo especializado (Cabré 1993: 369).

2.3 Terminología y terminografía

La terminología, señala Cabré (1993: 82), además de constituir el conjunto de términos correspondientes a un área particular puede ser entendida, por un lado, como "el conjunto de principios y de bases conceptuales que rigen el estudio de los términos", y por otro, como "una materia de intersección que se ocupa de la designación de los conceptos de las lenguas de especialidad", cuyo objetivo, a grandes rasgos, es la denominación de los conceptos.

Sager (1993: 22), por su parte, define estas tres concepciones de la terminología de la siguiente forma:

1. "El conjunto de prácticas y métodos utilizados en la recopilación, descripción y presentación de términos;"
2. "Una teoría, es decir, el conjunto de premisas, argumentos y conclusiones necesarias para la explicación de las relaciones entre los conceptos y los términos;"
3. "Un vocabulario de un campo temático especializado."

Como vocabulario de un campo particular, ya hemos hablado del papel que desempeña una terminología en la organización del conocimiento especializado, y de la relación ineludible que el conjunto léxico de un área establece con su sistema de conceptos, al igual que de los procesos de denominación y de significación que representan los términos y las definiciones respectivamente.

Como teoría, la terminología comenzó a considerarse una materia independiente de la lingüística a partir de los argumentos propuestos por Wüster, en los años treinta, sobre la necesidad de tratar los términos de forma diferente a las palabras del lenguaje general (Pearson 1998: 10). Lo anterior derivó en la denominada Teoría General de la Terminología (TGT) desarrollada posteriormente por los seguidores de Wüster de la escuela de Viena.

En este punto, Sager (1993: 35) menciona que se pueden identificar tres dimensiones en una teoría de la terminología: la dimensión cognoscitiva, la dimensión lingüística y la dimensión comunicativa. La dimensión cognoscitiva está en relación con la ordenación del conocimiento y con la explicación del sistema de conceptos de un

campo de especialidad a partir del conjunto léxico de éste, es decir, de su terminología. La dimensión lingüística hace referencia a los procesos de denominación seguidos para asignar formas lingüísticas específicas a cada concepto de un campo. Por último, la dimensión comunicativa se refiere al proceso continuo de transmisión de conocimiento inherente a todas las actividades especializadas y en la cual se justifica la práctica de la compilación y el procesamiento de la terminología.

Es a partir de la importancia de la dimensión comunicativa que Cabré propone la Teoría Comunicativa de la Terminología (TCT), la cual:

> "se perfila como una propuesta concebida dentro de una teoría amplia del lenguaje, y está incluida en una teoría de la comunicación que contiene los fundamentos necesarios de una teoría del conocimiento. Esta propuesta integra, teórica y metodológicamente, la variación lingüística, tanto formal como conceptual, y asume que los términos están asociados a características gramaticales […] y pragmáticas […]. La TCT pretende también dar cuenta de los términos como unidades al mismo tiempo singulares y similares a otras unidades de comunicación, dentro de un esquema global de representación de la realidad, admitiendo la variación conceptual y denominativa, y teniendo en cuenta la dimensión textual y discursiva de los términos." (Cabré 1999: 136)

Por otro lado, como conjunto de prácticas y métodos, la terminología está influenciada por la materia y el campo de conocimiento a la que hace referencia, lo cual la lleva a ser reconocida como una actividad interdisciplinaria (Sager 1993: 22). La terminología como una disciplina científica se encuentra relacionada con la lingüística, la filosofía, la información y documentación, el conjunto de las ciencias y sus aplicaciones, la lingüística computacional y la ingeniería del conocimiento (Arntz y Picht 1995: 22)[1]. Sin embargo, a diferencia de algunas áreas con las que se interrelaciona, la terminología no recurre únicamente a los

[1] Sobre la interdisciplinariedad actual de la terminología, específicamente en la denominada *sociedad de la información*, véase el apartado 3.1.1 en Pérez (2002).

datos de una sola área de conocimiento, sino que recurre al conjunto de datos que conforman su campo de trabajo (Cabré 1993: 82).

Una característica importante que transmite el concepto de interdisciplinariedad a la terminología es el hecho de que su práctica metodológica se reformula constantemente debido a los cambios naturales que se presentan en las áreas con las que se interconecta. El cambio en la concepción de su metodología a partir de los avances tecnológicos de las últimas décadas, por ejemplo, es una de las reformulaciones más importantes que ha sufrido el concepto de terminología como disciplina. Sager apunta al respecto:

> "La terminología hoy en día se asocia, por lo general, con el suministro de los servicios de información que requieren la recopilación de la información sobre los términos para compilar diccionarios y glosarios […]. Ahora la recopilación y el procesamiento terminológico son procesos semiautomáticos, que responden constantemente a las innovaciones tomadas de la informática, las ciencias de la información y la lingüística automatizada." (Sager 1993: 23-24)

Ahora bien, el trabajo terminológico al que hace mención Sager encargado de recopilar información sobre términos se conoce como *terminografía*. Esta disciplina, que ha surgido a partir de la necesidad de estructurar sistemas conceptuales especializados y de la necesidad de compilar términos, y que además ha progresado recientemente como producto de los avances y las necesidades de comunicación (García de Quesada 2001: 1.4.2), se refiere a las distintas actividades que suponen la búsqueda, organización y clasificación de unidades léxicas especializadas con el fin de desarrollar recursos donde se registren sus usos, descripciones y significados.

Desde esta perspectiva, resulta fácil confundir las actividades terminográficas con aquellas relacionadas a la práctica lexicográfica, la cual tiene igualmente como objeto de estudio las unidades léxicas, pero, a diferencia de la terminografía, se encarga del léxico de las áreas de conocimiento general. A pesar de las

relaciones y semejanzas que guardan ambas prácticas, existen ciertas consideraciones básicas para su distinción[2].

En primer lugar, el concepto de terminografía es relativamente joven comparado con la larga y establecida tradición de la lexicografía (Pérez 2002: 3.3). En segundo lugar, a pesar de que en la práctica comparten métodos y recursos para desarrollar su trabajo, el objetivo que persigue la terminología es el de definir unidades léxicas especializadas y no de conocimiento general. Por último, la aproximación que realizan a su objeto de estudio es distinta: la lexicografía parte de una concepción semasiológica, es decir, parte de la palabra para describir su significado, mientras que la terminografía parte de una concepción onomasiológica, es decir, parte del análisis de los conceptos para denominar los términos.

En este sentido, respecto a la relación entre terminología y terminografía, y a la diferencia entre estas y los conceptos de lexicología y lexicografía, García de Quesada señala lo siguiente:

> "el sufijo *-logía* se usa para denominar las disciplinas que se ocupan de la construcción de un marco teórico tanto del léxico (*lexicología*) como de los términos (*terminología*), y el sufijo *-grafía* para las disciplinas que se ocupan de la puesta en práctica de dichos marcos, lexicografía y terminografía, respectivamente." (García de Quesada 2001: 1.1)

En resumen, en el marco de los objetivos generales y particulares de esta tesis coincidimos con García de Quesada (2001: 1.1) en el sentido de que la distinción particular entre terminología y terminografía denota una diferencia entre la vertiente teórica y la vertiente aplicada. En esta investigación consideramos a la terminografía como la rama aplicada de la terminología, y consideramos además que en su práctica se encuentra embebida la finalidad de extraer información definitoria sobre términos, lo cual

[2] Para un estudio detallado sobre la distinción entre terminografía y lexicografía véase por ejemplo: Cabré (1993) apartado 1.6.2; Cabré (1999) capítulo "La terminología hoy: concepciones, tendencias y aplicaciones"; García de Quesada (2001) apartado 1.4; Montero (2002) apartado 1.3; Pérez (2002) apartado 3.3; Rey (1995) capítulo 7.

surge como motivación para el desarrollo de una herramienta que ayude en este proceso particular.

2.4 Metodología de la práctica terminográfica actual

La terminología, actualmente, es de interés para distintos tipos de investigadores como traductores o lexicógrafos, o bien especialistas como ingenieros, científicos o técnicos de diversas áreas, quienes han notado la necesidad de mejorar los aspectos comunicativos de sus prácticas al igual que el acceso a la información producida en sus áreas de conocimiento (Bourigault *et al*. 2001: viii). El procesamiento de la terminología se realiza hoy en día casi exclusivamente mediante sistemas computacionales (Sager 1993: 187), y podría decirse que la *terminografía computacional* es una práctica que ha surgido de la necesidad de procesar grandes cantidades de datos léxicos y terminológicos de una manera más eficiente[3].

A partir del avance tecnológico de las últimas décadas han surgido diferentes líneas de investigación en torno a la necesidad de satisfacer las demandas actuales de los distintos procesos terminológicos. Diversas técnicas del procesamiento del lenguaje natural y de la inteligencia artificial han permitido que la extracción y la representación del conocimiento especializado se realicen de forma eficiente y permitan la organización de estructuras sofisticadas que reflejen el conocimiento de un área especializada (Bourigault *et al*. 2001: viii).

El uso de sistemas computacionales ha permitido procesar la terminología de manera práctica gracias a su capacidad de almacenamiento, flexibilidad y rapidez. La flexibilidad que permiten los textos en formato electrónico para su tratamiento

[3] "Al principio se crearon sistemas de procesamiento de datos léxicos y terminológicos basados en metodologías clásicas de procesamiento de datos, pero el resultado quedó lejos de ser satisfactorio. Esto condujo a reconocer que el almacenamiento de datos léxicos no podía considerarse simplemente como otra aplicación de las técnicas tradicionales de procesamiento de información sino que había que desarrollar nuevas técnicas para representar adecuadamente el tipo de información contenida en los diccionarios especializados." Sager (1993: 187).

automático ha dado pie, por ejemplo, a la elaboración de bastas colecciones de documentos especializados: los denominados corpus lingüísticos. Por su parte, el desarrollo acelerado de medios de almacenamiento de información digital ha permitido a su vez un creciente desarrollo en el tamaño de dichos corpus.

Es importante señalar que los problemas y necesidades que surgen en la terminografía computacional requieren una nueva visión desde el punto de vista teórico y práctico (Bourigault *et al*. 2001: ix). En el caso específico de la compilación terminológica Sager apunta lo siguiente:

> "Puesto que la automatización afecta fundamentalmente a la naturaleza y métodos de la compilación terminológica es indispensable elaborar un conjunto de principios completamente nuevos para la compilación, que difiere en casi todos los aspectos de los principios establecidos en la época precedente a la automatización. Las posibilidades que ofrecen el análisis automático de textos y el procesamiento de grandes cantidades de datos han modificado la base misma de la compilación terminológica, la opinión sobre la relevancia de los datos, y el grado de intervención humana en el proceso." Sager (1993: 188)

De esta manera, los recursos desarrollados por la terminografía computacional constituyen uno de los avances más importantes relacionados con la práctica terminográfica actual. Algunos de los principales recursos que facilitan en gran medida las diversas etapas en la labor terminográfica son: a) diccionarios en formato electrónico, b) ontologías, c) corpus lingüísticos y d) bases de datos terminográficas. A continuación explicamos brevemente cada uno de estos recursos.

a) Los diccionarios en formato electrónico (Machine Readable Dictionaries: MRDs) y tesauros son recursos que contienen información léxica y conceptual. Representan una de las fuentes a las que el terminógrafo puede acudir en busca del conocimiento del área que estudia.
b) Las ontologías son también fuentes de conocimiento y pueden ser vistas como mapas conceptuales donde se representan los sistemas de conceptos de un área particular. Integran

información semántica referente a los tipos de relaciones que existen entre los diversos términos que contienen.

c) Los corpus lingüísticos son uno de los recursos que constituyen en la actualidad una base indispensable en la práctica terminográfica, sobre todo en la compilación terminológica (Sager 1993: 189). Los corpus son colecciones de textos en formato electrónico que incluyen documentos representativos de un área de conocimiento, tales como informes de investigación, publicaciones académicas, tesis, etc. Suelen estar estructurados mediante etiquetas para representar diferentes tipos de información de los textos, por ejemplo información morfológica, información sintáctica o información semántica.

d) Las bases de datos terminográficas[4] son una colección de registros donde se almacena diferente tipo de información relevante sobre un término específico. Las bases de datos terminográficas pueden servir en distintas tareas relacionadas con el procesamiento del conocimiento especializado, y bien pueden constituir instrumentos de trabajo para traductores o expertos en normalización, como fuentes para usuarios de distintos campos de actividades técnicas, estudiantes de un campo de conocimiento particular, o de manera específica para aplicaciones del procesamiento del lenguaje natural (Sager 1993: 194-195).

Así, en la compilación terminológica los registros de las bases de datos terminográficas, es decir las fichas o entradas, juegan un papel esencial. Dependiendo del tipo de trabajo que se realice, las entradas contendrán tanta información relevante sobre los términos como sea necesario. Cada entrada constituye un conjunto de datos

[4] Consideramos importante notar que existe una variación denominativa sobre el concepto de banco de datos terminográficos a lo largo de su historia: banco de términos, banco de datos terminológicos, banco de datos electrónicos, etc. En esta investigación decidimos adoptar el término de banco de datos terminográficos en consenso con la idea de que representan un recurso que sirve en el trabajo de la vertiente aplicada de la terminología, es decir la terminografía. Coincidimos en este aspecto con García de Quesada (2001): "refiriéndose a base de datos, aún conscientes del uso extendido dentro de la literatura del sufijo *-lógica* en detrimento de *-gráfica*, entendemos que el sufijo que se ha de utilizar es *-gráfica*, ya que se trata de una aplicación de un método de manipulación, de un producto final y en ningún caso de una disciplina."

donde se incluye información léxica de los términos, como definiciones, denominaciones o contextos, e información específica sobre su registro, como clave de clasificación, código de gestión, fecha de registro de datos, etc. (Arntz y Picht 1995: 268).

La norma ISO 10241 (1992: 7) señala que los contenidos mínimos que debe incluir una entrada terminológica son:

- Número de entrada
- Término preferido que representa el concepto
- Definición del concepto

Adicionalmente, y dependiendo del tipo de trabajo que se realice, cada ficha puede incluir la siguiente información:

- Pronunciación
- Forma abreviada
- Forma completa, cuando el término preferido es una abreviación
- Símbolo
- Campo de conocimiento
- Gramática
- Referencia a fuentes
- Términos no preferidos (por ejemplo obsoletos)
- Otra representación o representaciones del concepto (por ejemplo fórmulas o figuras)
- Referencias a términos o entradas relacionadas
- Ejemplos de uso del término
- Notas
- Equivalencias del término en otras lenguas

Sager, a partir de la idea de que la información de un registro terminológico "es compleja y consiste en un número de subconjuntos que se pueden compilar y procesar de manera bastante independiente" (Sager 1993: 205), señala que el contenido de cada entrada terminológica de un término se puede dividir en distintas especificaciones: conceptual, lingüística, pragmática y equivalente. El conjunto de estas especificaciones y los datos particulares que giran en torno al registro conforman un modelo complejo y moldeable según lo requiera la situación específica.

En la tabla 2.1 presentamos un modelo de registro terminológico propuesto por Sager, donde podemos comparar los contenidos señalados en la norma ISO antes descrita y observar que los elementos mínimos coinciden en una misma línea: *definición – término – contexto – término equivalente*, alrededor de los cuales giran los demás contenidos que se especificarán de acuerdo con la necesidad y finalidad pertinente del trabajo terminográfico.

INFORMACIÓN FUENTE							
Origen No.	Tipo Página	Origen No.	Tipo Página	Origen No.	Tipo Página	Origen No.	Tipo Página
Especificación Conceptual		**Especificación Lingüística**		**Especificación Pragmática**		**Especificación Equivalente**	
		Lengua		Lengua		Lengua	
Definición		Término		Contexto		Término equivalente	
		Información gramatical				Información gramatical	
Enlaces con otros conceptos		Sinónimos		Notas de uso o ejemplo		Sinónimos	
Notas de alcance		Abreviaturas		Uso		Abreviaturas	
Campo temático		Variantes		Uso		Variantes	
Fecha	Tipo	Fecha	Tipo	Fecha	Tipo	Fecha	Tipo
No. de registro		No. del área de aplicación		Terminólogos			
INFORMACIÓN DE MANTENIMIENTO							

Tabla 2.1. Modelo de registro terminológico (tomado de Sager (1993: 206))

Las categorías de este modelo no son estrictas en el sentido de que cada ficha deba contener cada uno de los elementos propuestos para cada especificación; por el contrario, son flexibles en tanto la información de una especificación puede tener el mismo valor en otra, la información se puede actualizar independientemente y puede estar interrelacionada con otras entradas, y la cantidad de datos se puede adaptar según las necesidades de información requerida.

En relación con los objetivos de nuestro trabajo, una herramienta como la que aquí proponemos tendría una incidencia directa en el proceso de elaboración de entradas terminográficas. Los datos provistos por un extractor de información definitoria estarían relacionados con los datos básicos que permiten la elaboración de fichas para el trabajo terminológico. En la búsqueda de CDs, debe tenerse en cuenta que éstos aportan diferente tipo de información que puede incluir tanto una definición como sinónimos de los términos, abreviaturas, indicaciones de uso, alcance, etc., la cual sería un tipo de información que quedaría registrada en una ficha terminográfica.

En resumen, la práctica terminográfica actual está claramente influenciada por el constante avance en el desarrollo de recursos por parte de la terminografía computacional. En la delimitación del tema y la definición del trabajo terminológico entran en juego diccionarios en formato electrónico u ontologías, como fuentes de consulta de información léxica estructurada. Por otra parte, durante el desarrollo del trabajo terminográfico los corpus lingüísticos son una fuente de conocimiento donde los terminógrafos pueden partir para compilar la terminología y analizarla en situaciones reales. Finalmente, en la presentación de los resultados los bancos de datos terminográficos constituyen actualmente uno de los recursos más flexibles, moldeables y manipulables de acuerdo con el tipo de trabajo que se realice.

2.5 Contextos definitorios y relaciones semánticas

Por otra parte, Pérez señala que en la práctica terminográfica se ha contado tradicionalmente con tres formas para representar el conocimiento especializado:

> "La descripción de los conceptos por medio de sus características, la estructuración de las relaciones que existen entre los conceptos […] y la formulación de definiciones que describan al concepto en el marco de una estructura conceptual." Pérez (2002: 5.1).

Para describir los conceptos y estructurar las relaciones que existen entre ellos se debe tener en cuenta que las características conceptuales pueden ser de dos tipos generales: *atributos* y *relaciones semánticas*.

Los *atributos* hacen referencia a cada concepto particular sin involucrar a otros conceptos del mismo domino de conocimiento. Entre los atributos se establecen las características y cualidades específicas de los términos.

Por su parte, las *relaciones semánticas* unen a cada concepto con otros conceptos del mismo dominio de conocimiento al cual pertenecen. En este tipo de relaciones se señalan, por ejemplo, las relaciones jerárquicas de los términos con respecto a una red de conceptos donde se representan igualmente otros términos relacionados.

Por lo general, señala Meyer (2001: 280), en bases de datos terminográficas, al igual que en diccionarios especializados, las características principales de los términos, es decir los atributos, se representan de manera explícita en las entradas terminográficas, mientras que las relaciones semánticas están representadas de manera implícita a través de las definiciones y los ejemplos provistos para aclarar el significado de cada término.

Meyer agrega que existen tres procedimientos básicos que se deben tener en cuenta antes de que los terminógrafos puedan formular entradas donde se definan los atributos y las relaciones semánticas de los términos (Meyer 2001: 279):

1. Identificación de los términos: ¿cuáles son los términos que deberían ser descritos para el dominio en cuestión?
2. Análisis conceptual de los términos: ¿qué significan estos términos?
3. Análisis lingüístico de los términos: ¿cómo son usados los términos en su contexto?

El procedimiento inicial está relacionado con el hecho de compilar la terminología que se representará en el trabajo terminográfico en cuestión. Enseguida, es necesario encontrar información conceptual sobre los términos que permita al terminógrafo formular

definiciones de acuerdo con el tipo de trabajo que realiza, lo cual está estrechamente ligado a su vez con el tercer procedimiento de analizar los términos en su contexto real de uso.

Independientemente del tipo de fuentes en las que esté basado el trabajo, el terminógrafo cuenta con dos estrategias básicas para adquirir conocimiento conceptual sobre los términos: consultar especialistas del área de conocimiento en cuestión, o bien consultar textos relacionados con el área de estudio (Meyer 2001: 279).

Los *textos especializados*, es decir aquellos que pertenecen a un área de conocimiento particular, pueden ser vistos como registros que vehiculan el conocimiento de un campo específico y en los cuales se recurre en mayor o menor medida a la descripción o explicación de los conceptos de acuerdo con el tipo de situación comunicativa a la cual pertenecen. Como señalamos en 2.1., en la comunicación entre expertos y semi-expertos, o bien entre expertos y no-expertos se utilizan con mayor frecuencia contextos a través de los cuales se aporta información sobre los atributos y relaciones de los términos.

En relación con lo anterior, los fragmentos textuales donde se aporta información pertinente para conocer el significado de un término se conocen como *contextos definitorios*. Pueden considerarse como unidades discursivas que vehiculan información predicativa sobre un término de un dominio de conocimiento específico. Además, en los contextos definitorios no sólo se aporta información que permite entender el significado de los términos, sino también conocer las relaciones semánticas que éstos presentan con otros términos y establecer así una red conceptual del campo de conocimiento al que pertenecen. A continuación presentamos un ejemplo de un CD.

Contexto Definitorio
"Los compuestos que no derivan de la adormidera, pero que ejercen efectos directos uniéndose a los receptores específicos para opiáceos se denominan opioides."

Tabla 2.2. Ejemplo de contexto definitorio (tomado del Corpus Técnico del IULA (Bach *et al*. 1997))

En el ejemplo anterior observamos que se introduce información definitoria sobre el término *opioides*, especificándose que pertenece a una clase general de *compuestos*, cuya distinción de otros tipos de compuestos es *que no derivan de la adormidera, pero que ejercen efectos directos*… etc.

En resumen, la importancia de los contextos definitorios y las relaciones semánticas en la práctica terminográfica radica precisamente en la necesidad de comprender un término y situarlo frente a los demás términos de su campo, a través del análisis de las situaciones reales en las que éstos se definen dentro de la comunicación del lenguaje especializado.

2.6 Las definiciones en textos especializados

El estudio de la realización de las definiciones en los textos especializados juega entonces un papel esencial en la práctica terminográfica, donde se suelen considerar dos clases generales de definiciones: *definiciones intensionales* y *definiciones extensionales* (ISO 704 2000(E): 15; Wright y Budin 1997: 325).

Las *definiciones intensionales* son aquellas donde se hace mención explícita del concepto superordinado del término seguido de las características que lo distinguen de otros términos, mientras que las *definiciones extensionales* señalan la totalidad de los objetos a los que refiere un concepto. El primer tipo de definiciones es lo que se considera también como definiciones *analíticas* o definiciones de *genus y diferencia*. A continuación señalamos dos ejemplos que provee la norma ISO 704 (2000(E): 15-16) de definición intensional y extensional:

a) **Lead pencil** ~Pencil whose graphite core is fixed in a wooden casing that is removed for usage by sharpening.
b) **Threatened species** ~Critically endangered species, endangered species or vulnerable species.

El primer ejemplo es una definición intensional donde se provee la clase general a la cual pertenece el término *lead pencil*: *pencil*, así como las claves que lo caracterizan: *pencil whose grapgite core is fixed in a wooden casing*... etc. El segundo caso corresponde a un

ejemplo de definición extensional donde se enumeran los tipos que conforman al término *threatened species*.

Por otro lado, también en la línea de describir las definiciones en el ámbito de la práctica terminográfica, Trimble (1985) las clasifica en *definiciones simples* y *definiciones complejas*. En esta clasificación estructural, el primer tipo corresponde a definiciones que en los textos especializados se expresan en una sola oración, mientras que las definiciones complejas se encuentran en grupos de párrafos y por regla general su núcleo es una definición simple. Las definiciones simples, a su vez las clasifica en:

1. *Definiciones formales*, las cuales arrojan la información más completa en tanto siguen la fórmula aristotélica de *Término = Género próximo + Diferencia específica*.
2. *Definiciones semi-formales*, las que sólo incluyen dos de estos tres elementos, es decir, el término que se define y la diferencia específica.
3. *Definiciones no-formales*, las cuales no proporcionan información tan precisa como lo hacen las dos clases anteriores, sino su función es la de definir alguna característica en un sentido muy general para que el lector pueda reconocer un elemento familiar del término (Trimble 1985: 77-78).

Tipo definición	**Elementos constitutivos**
Definiciones formales	1. El nombre del término que se define 2. La clase a la cual pertenece el término 3. Las diferencias entre el término y todos los demás términos de su clase
Definiciones semi-formales	1. El nombre del término que se define 2. Las diferencias entre el término y todos los demás términos de su clase
Definiciones no-formales	1. El nombre del término que se define 2. Otras palabras o frases que tengan un significado aproximado o den una idea general del término

Tabla 2.3. Tipología de definiciones simples según Trimble (tomado de Trimble (1985: 77-78)

En la tabla anterior podemos observar algunas características de las definiciones simples propuestas por Trimble. Algunos ejemplos que propone este mismo autor son los siguientes:

a) Definición formal: "An arachnid is an invertebrate animal having (or, which has) eight legs extending at equal intervals from a central body."
b) Definición semi-formal: "An arachnid has eight legs extending at equal intervals from a central body."
c) Definición no-formal: An arachnid is a spider.

El primero corresponde a una definición formal en la cual se especifica que el término *arachnid* es un *invertebrate animal* cuya característica es *having (or, which has) eight legs extending at equal intervals from a central body*. En el segundo ejemplo, si eliminamos el hiperónimo *invertebrate animal* tenemos entonces una definición semi-formal. Finalmente, en el tercer caso se especifica únicamente un sinónimo del término, es decir *spider*, que pueda resultar familiar para la comprensión del concepto que representa el término. En este último caso de las definiciones no-formales, Trimble señala que pueden ser definiciones por sinonimia, definiciones por enunciados negativos y definiciones por antonimia.

En síntesis, las clases generales en las que se clasifican los distintos tipos de definición en el trabajo terminográfico están relacionadas con la cantidad y precisión de la información que proveen. Se considera a las definiciones como una información completa cuando se señala tanto la clase a la que pertenece el término como sus características distintivas. Cuando esta información carece de la clase general se considera a la definición como un tipo de información incompleta, pero no por ello deja de ser útil para comprender el significado del término. Igualmente, las definiciones que eliden la clase general del término suelen ser casos donde mayoritariamente se transmite información sobre las relaciones semánticas de los términos.

Cabe señalar que para nuestro estudio consideramos pertinente una tipología como la de Trimble, con el fin de describir la diversidad de formas en que las definiciones pueden encontrarse en los textos de especialidad, al igual que los distintos tipos de contenido semántico que pueden aportar. Más adelante, en el capítulo 4,

presentaremos una tipología semántica de contextos definitorios basada en el tipo de información semántica que aportan los distintos tipos de definiciones.

3. Avances en la extracción automática de contextos definitorios

En el capítulo anterior abordamos algunos conceptos básicos necesarios en nuestra investigación. Hemos hecho mención a la metodología de la práctica terminográfica actual, y hemos visto el rol que desempeñan los contextos definitorios y las relaciones semánticas en la organización y descripción del conocimiento especializado. Asimismo, nos hemos referido al estudio de la forma en que suelen presentarse las definiciones en textos especializados y hemos hecho mención de una tipología propuesta por Trimble para clasificar distintos tipos de definiciones.

Ahora bien, en este capítulo revisamos algunas metodologías y/o sistemas que han tratado directamente el tema de la extracción automática de contextos definitorios. En primer lugar abordaremos algunas nociones básicas en torno al concepto de conocimiento definitorio y la extracción de información (3.1). Enseguida trataremos ciertas consideraciones en torno a la diferencia entre la extracción de relaciones semánticas y la extracción de contextos definitorios, principalmente para delimitar un estado del arte pertinente a nuestra investigación (3.2). Posteriormente analizamos algunas metodologías y algunos sistemas enfocados en la extracción de contextos definitorios (3.3), para finalmente hacer una análisis contrastivo sobre dichas metodologías o sistemas, así como de sus procesos evaluativos y de sus resultados (3.4).

3.1 Conocimiento definitorio y extracción de información

La cantidad de información disponible hoy en día en formato electrónico ha propiciado el desarrollo de diversas herramientas computacionales para facilitar su procesamiento a distintos niveles. Distintas áreas en el ámbito de la informática se han enfocado en la elaboración de sistemas que faciliten el procesamiento de la información. Una de estas áreas es el *Procesamiento del Lenguaje Natural* (PLN), un campo de la inteligencia artificial que estudia los problemas referentes al desarrollo de sistemas para la generación y entendimiento del lenguaje humano.

A su vez, algunas áreas dentro del PLN se han dirigido específicamente al procesamiento de información textual. Tal es el caso de la *Recuperación de Información* (RI), cuya finalidad es la elaboración de sistemas para la búsqueda y selección de documentos que cumplan ciertos criterios señalados por un usuario (por ejemplo, los motores de búsqueda en Internet), o bien la *Extracción de Información* (EI), que se encarga de desarrollar sistemas para la búsqueda y selección de datos específicos sobre eventos, entidades o relaciones a partir de un conjunto de documentos.

En el ámbito de la terminología, la EI se ha orientado a elaborar sistemas para la extracción de información terminológica, dentro de los cuales se encuentran los sistemas para la extracción de términos. Este tipo de EI ha provisto de sistemas basados en reglas lingüísticas, en reglas estadísticas o híbridos (una combinación de los dos anteriores) para el análisis de textos con el fin de extraer una lista de candidatos a términos (Vivaldi 2004; Vivaldi y Rodríguez 2007).

La EI se ha enfocado además en la extracción de otro tipo de información terminológica que suele denominarse *conocimiento definitorio*, y que de manera análoga a la noción de contexto definitorio que explicamos en el capítulo anterior, es un tipo de información que permite inferir el significado de los términos a partir de la descripción de sus atributos, características o relaciones semánticas. De esta forma, la extracción de conocimiento definitorio es un proceso común en distintos trabajos enfocados principalmente a:

a) La extracción de relaciones semánticas (RSs)
b) La extracción de contextos definitorios (CDs)

Este tipo de estudios, que abordan el problema de extraer automáticamente información relevante sobre términos, constituyen un tipo de EI que ha supuesto un avance en la elaboración de recursos léxicos y terminológicos para la organización conceptual de unidades de conocimiento especializadas, así como para la descripción de sus significados.

Por un lado, ciertas investigaciones se han concentrado en la extracción automática de RSs utilizando tanto diccionarios en formato electrónico como textos especializados, principalmente como soporte en la construcción y organización de lexicones, terminologías, taxonomías y ontologías (Auger y Barrière 2008: 1).

Por otro lado, algunos estudios se han enfocado en la extracción de CDs sobre textos especializados para la elaboración de recursos como diccionarios, bases de datos léxicas, bancos de conocimiento terminológico, o bien como un proceso anterior a la misma extracción de RSs.

Además, en los últimos años la extracción de conocimiento definitorio ha comenzado a influir en el desarrollo de herramientas para mejorar la estructuración de la información presente en la Web. Tal es el caso del peso que se le ha otorgado al uso de ontologías en el desarrollo de la *Web Semántica*.

Ahora bien, las metodologías para la extracción de conocimiento definitorio tienden a coincidir en ciertos aspectos, por lo que la frontera que delimita los estudios relacionados con la extracción de RSs frente a la extracción de CDs puede prestarse a confusión, principalmente por la similitud entre sus metodologías.

Es por ello, y por la necesidad de establecer un estado del arte concreto y acorde con nuestra investigación, que consideramos necesario hacer una breve mención a las características específicas y distintivas de estos tipos de extracción de conocimiento definitorio.

3.2 Consideraciones sobre la diferencia entre extracción de relaciones semánticas y extracción de contextos definitorios

La extracción de RSs ha despertado un creciente interés en diversas disciplinas científicas como la inteligencia artificial, la lingüística computacional, la lingüística generativa o las ciencias cognitivas (Auger y Barrière 2008: 1). Uno de los primeros ámbitos donde se comenzó a extraer RSs de manera automática fue sobre diccionarios en formato electrónico (MRDs, *Machine Readable Dictionaries*). Existe un acuerdo general de que los MRDs pueden ser usados

como un repositorio de información léxica y taxonómica que puede ser extraída usando técnicas computacionales a partir del análisis de la definición en una entrada lexicográfica o terminográfica (Calzolari y Picchi 1988; Pustejovsky *et al*. 1993; Boguraev y Pustejovsky 1996).

La información contenida en las definiciones analíticas, es decir, aquellas que esclarecen el género próximo y la diferencia específica de un término, es uno de los tipos de información que ha despertado mayor interés para su extracción automática a partir de MRDs. El trabajo de Amsler (1981) fue uno de los primeros en asumir la utilidad del género próximo y la diferencia específica en las definiciones analíticas. En dicho trabajo se presentó una metodología para la elaboración de taxonomías a partir de la identificación de las definiciones presentes en las entradas de un diccionario. Alsawhi (1987) propuso, por su parte, un método basado principalmente en la búsqueda del patrón IS-A para identificar el género próximo en una definición, al igual que extraer y categorizar información léxica utilizando las entradas del *Longman Dictionary of Contemporary English* (*LODCE*). Algunos de los tipos de información léxica procesada en las entradas de dicho diccionario eran la *clase*, *colectivo*, *propiedades* o *materiales* del término.

El mismo patrón IS-A se ha usado, además, con la finalidad de delimitar una serie de relaciones semánticas a partir del modelo de las definiciones analíticas. Por ejemplo, en el marco del proyecto multilingüe ACQUILEX[1], Vossen y Copestake (1993) recurrieron al patrón IS-A con la finalidad de delimitar una serie de relaciones semánticas a partir del modelo de las definiciones analíticas. En este estudio se contemplaban tres tipos de relaciones derivadas de este patrón: *hiponimia – hiperonimia*, *sinonimia*, e *individuación*. De igual forma, Dolan *et al*. (1993) trabajaron en una metodología para la extracción no sólo de hiperónimos a partir del LODCE, sino además de relaciones como *parte_de*, *locación* y *propósito*, entre otras.

En la siguiente tabla podemos observar algunos ejemplos de la extracción de RSs a partir de las entradas de un diccionario.

[1] http://www.cl.cam.ac.uk/research/nl/acquilex/

1	Entrada	**Nail** "A thin piece of metal with a point at one end and a fiat head at the other for hammering into a piece of wood, usu. to fasten the wood to something else"
	Resultado	Class – piece Material – metal Properties – thin Has-part - class point
2	Entrada	**Authority** (n, 7) "A person, book, etc., mentioned as the place where one found certain information"
	Resultado	Hypernym - person (= [+human]) Hypernym - book Location - find Typical_subject - [+human] Typical_object - information

Tabla 3.1. Ejemplo de relaciones semánticas extraídas del LODCE (tomado de 1: Alsawhi (1987: 198) y 2: Dolan *et al*. (1993: 7))

Podemos ver que el tipo de información extraída era no sólo el hiperónimo del término, sino otras relaciones específicas como *clase*, *materiales*, *propiedades* o *partes_de*. Observamos también que el resultado esperado son pares de términos unidos mediante una relación específica, por ejemplo:

- Nail *class* piece
- Nail *has_part* point
- Authority *hypernym* person | book

Por otro lado, algunas investigaciones observaron que el modelo de definición analítica presente en los diccionarios no era suficiente para describir el contenido conceptual de los términos, y que era necesario considerar además otro tipo de definiciones y otro tipo de fuentes. Estudios como el de Sager y Ndi-Kimbi (1995) sobre la realización de las definiciones en textos especializados cobraron entonces una mayor relevancia, a la vez que la extracción de RSs se concentró en la explotación de corpus lingüísticos y en la búsqueda de otros patrones diferentes al patrón canónico IS-A.

De esta forma, Hearst (1992) se enfocó en el uso de corpus lingüísticos para la búsqueda de hipónimos a través de patrones léxico-sintácticos. En la siguiente tabla se presentan dos ejemplos de la formalización de dichos patrones.

El estudio de Hearst incluía un proceso que permitía adquirir patrones nuevos por medio de la extracción de las ocurrencias de dos términos sobre los cuales se conociera previamente el tipo de relación semántica que los unía. Se establecía una ventana de coocurrencia entre estos dos términos, es decir un número determinado de palabras, y a partir de ésta se inferían nuevos patrones léxico-sintácticos.

1	Patrón	*such NP as* {*NP ,*}* {(*or* [*and*)} *NP*
	Ejemplo	…works by such authors as Herrick, Goldsmith, and Shakespeare
	Resultado	→ hyponym: author, Herrick → hyponym: author, Goldsmith → hyponym: author, Shakespeare
2	Patrón	*NP* {,} *including* {*NP* ,} * {*or* | *and*} *NP*
	Ejemplo	…all common-law countries, including Canada and England
	Resultado	→ hyponym: Canada, common-law country → hyponym: England, common-law country

Tabla 3.2. Ejemplos de patrones para la búsqueda de hipónimos (tomado de Hearst (1992: 541))

En la tabla anterior podemos observar el resultado que derivaba en pares de términos unidos mediante una relación de hiponimia: *author – Shakespeare*; *Canada – common-law country*, etc.

Otras investigaciones recurrieron a esta misma idea de utilizar pares de términos que comparten una relación semántica con el fin de adquirir nuevos patrones léxico-sintácticos. Morin (1998), por ejemplo, utilizó esta metodología para extraer hipónimos sobre un corpus de textos científicos en el área de agronomía, mientras que Berland y Charniak (1999) la utilizaron para buscar las *partes* de un

concepto sobre el *North American News Corpus*[2]. Por otro lado, Condamines y Rebeyrolle (2001) presentaron una metodología para la elaboración de una *base de conocimiento terminológico basada en corpus*[3], donde se partía del análisis de las ocurrencias de dos términos que compartieran una relación para poder identificar *patrones de relaciones conceptuales*[4]. Como resultado, para una relación del tipo *has the responsability for*, se podían identificar los siguientes patrones: a) *is entrusted with*, b) *is responsable for*, c) *to be in charge of*, d) *to be within the competence of*.

A la par de estos trabajos que explotaban la idea de buscar ocurrencias de términos junto con ciertos patrones léxico-sintácticos, Pearson (1998) presentó un estudio donde describía la posibilidad de usar esta metodología para adquirir no sólo RSs específicas, sino enunciados con descripciones completas sobre el significado, modo de operación y condiciones de uso de los términos, es decir CDs. En la misma línea, Meyer (2001) señaló que la ocurrencia de términos en conjunto con patrones léxico-sintácticos puede aportar información útil para situar al término dentro de una red conceptual específica y también para describir sus atributos, con lo cual se provee información básica sobre su significado.

Tanto en los estudios de Pearson como de Meyer, los enunciados donde se presentan los patrones léxico-sintácticos pueden contener cierto tipo de RSs, siendo la más evidente la relación de hiponimia, es decir, la que se establece entre el término y su género próximo (expresada por la ecuación: *X = Y + características distintivas*). El género próximo corresponde al hiperónimo y puede estar introducido por patrones canónicos como IS-A o *se entiende por*, *se define como, se conoce a*, etc. Asimismo, otro tipo de patrones léxico-sintácticos que no establecen el género próximo al cuál pertenece el término, pero que pueden ser usados como punto de

[2] En ambos estudios, los resultados obtenidos eran comparados con la base de datos léxica WordNet (Fellbaum 1998), una red semántica con una basta cantidad de información conceptual sobre sustantivos, verbos, adjetivos y adverbios.

[3] *Corpus-based Terminological Knowledge Base – CTKB.*

[4] "Conceptual relation pattern: a discursive structure used as an indication of the possible transition from the discourse to a model, allowing, the more or less direct construction of a model in the form of a semantic relation depending on its relation with the context." (Condamines 2002: 6).

partida en la formulación de definiciones a través de la información que expresan sobre una determinada relación como funcionalidad o extensión, son patrones del tipo *contiene*, *está constituido por*, *funciona como*, *se usa en*, etc.

Pearson y Meyer coincidieron igualmente en la idea de que existen otro tipo de características en los textos especializados, como los signos de puntuación o la tipografía textual, que por sí mismas o en conjunto con los patrones léxico-sintácticos pueden servir como puntos clave en la extracción de información definitoria sobre términos, lo cual dio como resultado la inclusión de *patrones tipográficos* en el grupo de patrones definitorios.

Tomando en cuenta lo anterior, se comenzaron a desarrollar algunos sistemas que integraban la búsqueda de patrones léxico-sintácticos y patrones tipográficos, por ejemplo el sistema de Muresan y Klavans (2002) denominado DEFINDER. Este sistema partía de la búsqueda de frases como *is called*, *is the term used to describe*, *is defined as*, y tipografía textual como paréntesis o guiones, con el fin de extraer definiciones de textos especializados en el área de medicina. Como podemos observar en la siguiente tabla, el tipo de información extraída podía ser una relación semántica o una definición.

Myocardial Infarction	
1	Heart tissue death
2	The most extreme state of oxygen deprivation, in which whole regions of heart muscle cells begin to die for lack of oxygen
3	Heart attack

Tabla 3.3. Ejemplos de definiciones extraídas automáticamente por DEFINDER (tomado de Muresan y Klavans (2002: 232))

En el ejemplo anterior, a) y c) presentan dos posibles sinónimos del término *myocardial infarction*, mientras que b) muestra una definición completa a partir de la descripción del género próximo del término y sus características distintivas. En este caso, la extracción de conocimiento definitorio implica no sólo una relación semántica (sinonimia), sino un conjunto de información útil para la comprensión del significado del término.

Como veremos en el siguiente apartado, los estudios sobre la extracción de CDs siguieron en general la misma línea de buscar *patrones definitorios* (léxico-sintácticos o tipográficos), tomando en cuenta que éstos pueden recuperar relaciones semánticas específicas, al igual que descripciones generales acerca del significado de los términos, y que pueden servir en la elaboración de diversos tipos de recursos terminológicos o bien como punto de partida para la misma extracción de relaciones semánticas.

En síntesis, de las metodologías para la extracción de relaciones semánticas y la extracción de contextos definitorios podemos resaltar tres puntos de encuentro generales.

El primero y más evidente es el tipo de patrones utilizados para la búsqueda de relaciones semánticas o contextos definitorios. Un patrón como IS-A se podrá usar para recuperar hiperónimos, aunque también podrá usarse para extraer descripciones del significado de un término, sobre todo en los casos en los que el hiperónimo va acompañado de las características distintivas que lo distinguen de otros miembros de su clase.

El segundo punto de encuentro es la metodología para la identificación o descubrimiento de patrones. La identificación inicial de patrones se realiza de diversas formas, siendo por lo general un proceso manual o como mucho un proceso semiautomático. En ocasiones se parte de la identificación manual de contextos definitorios en un corpus, y a partir de ellos se identifica el paradigma inicial de patrones. También se puede partir de las ocurrencias de un término específico y sobre ellas delimitar manualmente el grupo de patrones a utilizar. Otra posibilidad es iniciar con la extracción de ocurrencias de dos términos sobre los cuales se tiene certeza que comparten una relación semántica específica, y establecer una ventana de palabras entre ellos, la cual puede servir asimismo para la obtención del inventario de patrones.

En tercer lugar, el tipo de resultado esperado constituye otro punto en común, en tanto lo que se extrae en ambos casos son un término específico ligado a otro término o grupos de términos, y unidos mediante una relación semántica. En el caso de la extracción de CDs, el resultado conforma además un conjunto de información pertinente para proveer definiciones, constituir puntos de inicio para

formular definiciones, o aportar información terminológica sobre un término específico (Meyer 2001: 281-282).

Así, podemos sintetizar que la metodología de extracción de RSs incluye lo siguiente:

Extracción de relaciones semánticas	
1	Delimitación de las relaciones semánticas de interés.
2	Descubrimiento de patrones que expresen explícitamente las relaciones delimitadas, así como las condiciones sintácticas dentro de las cuales se realiza la relación semántica.
3	Búsqueda de ocurrencias de relaciones a partir de los patrones descubiertos.
4	Implementación de las relaciones semánticas extraídas como nuevas instancias en ontologías o bases de datos terminológicas.

Tabla 3.4. Metodología de la extracción de relaciones semánticas (tomado de Auger y Barrière (2008: 3))

Mientras que, por su parte, la extracción de CDs recurre a los siguientes puntos:

Extracción de contextos definitorios	
1	Identificación de un paradigma inicial de patrones definitorios.
2	Extracción automática de los patrones definitorios y análisis de los resultados, con el fin de añadir otros patrones al paradigma inicial e incluir restricciones en los patrones.
3	Aplicación de los cambios necesarios.
4	Repetición de 2 y 3 tantas veces sea necesario para mejorar el funcionamiento del sistema.

Tabla 3.5. Metodología de la extracción de contextos definitorios (tomado de Meyer (2001: 292))

De las tablas anteriores podríamos sintetizar a grandes rasgos las siguientes diferencias:

a) La extracción de RSs se ha realizado tanto en documentos estructurados (MRDs, enciclopedias electrónicas) como en documentos no-estructurados (textos especializados). Por su

parte, la extracción de CDs se ha basado principalmente en documentos no-estructurados.

b) La búsqueda de RSs puede realizarse a partir de patrones tan específicos como lo requiera el ámbito de aplicación[5]; por su parte, la búsqueda de CDs tiende a realizarse con patrones más genéricos que puedan ser aplicados a cualquier área de conocimiento.
c) Los patrones utilizados en la búsqueda de CDs incluyen patrones léxico-sintácticos y patrones tipográficos, mientras que en la extracción de RSs el uso de los últimos es casi nulo[6].
d) El resultado esperado en la extracción de RSs es un conjunto delimitado y conciso de información entre dos términos o grupos de términos, entre los cuales existe una relación específica. En la extracción de CDs el resultado es un conjunto de información que incluye además descripciones, condiciones de uso o información pragmática que ayuda a la comprensión del significado del término.
e) Aunado a lo anterior, el resultado de la extracción de relaciones semánticas, estructuralmente hablando, será un grupo de palabras o frases, mientras que en el caso de la extracción de contextos definitorios el resultado tenderá a ser una oración o un grupo de oraciones.
f) La extracción de contextos definitorios puede considerarse como un fin en sí mismo, o bien un paso previo que puede servir en la delimitación de relaciones semánticas.

Tomando en cuenta las distinciones que hemos señalado hasta ahora, en el siguiente apartado nos enfocaremos a realizar una revisión del estado del arte de estudios cuya finalidad es la extracción de CDs y que de una manera u otra han incidido en la concepción de nuestra propia metodología.

---

[5] Un inventario amplio se puede encontrar en el estudio de Feliu (2004), quien incluye patrones para detectar, entre otras, relaciones más específicas como: *semejanza*, *inclusión*, *secuencia*, *causa*, *instrumento*, *asociación*.

[6] El uso de paréntesis como patrones tipográficos es uno de los pocos a los que se ha recurrido en la búsqueda de hipónimos (véase por ejemplo el trabajo de Davidson (1997)). El uso más frecuente de la tipología textual para la extracción de relaciones semánticas se da en combinación con patrones léxico-sintácticos, por ejemplo cuando el término se encuentra marcado entre comillas, o resaltado en itálicas.

3.3 Análisis de metodologías y sistemas para la extracción de contextos definitorios

En este apartado analizamos algunas investigaciones que han abordado el tema de la extracción automática de CDs. Nos referiremos a ellos con el nombre del sistema (en el caso de que lo tenga), con el nombre del proyecto en el que se enmarca el estudio, o bien con el título de una publicación al respecto. De cada uno de ellos reseñaremos los siguientes puntos:

- Lengua de aplicación
- Descripción general y finalidad del estudio
- Corpus de estudio o ámbito de aplicación
- Metodología usada
- Resultados y evaluación

Cabe aclarar de antemano dos conceptos básicos que suelen emplearse en la evaluación de las metodologías de extracción de CDs: los índices de Precisión y Cobertura (*Precision & Recall*). A grandes rasgos, la precisión es una medida para determinar cuánta información extraída automáticamente corresponde a información *relevante*, mientras que la cobertura es una medida para indicar cuánta de la información *relevante* en el input (corpus de entrada) se extrajo automáticamente. Pensando en el escenario de la extracción de CDs, estas medidas se obtendrían de la siguiente manera:

Precisión = CDs extraídos automáticamente / Total de posibles CDs extraídos automáticamente.

Cobertura = CDs extraídos automáticamente / Total de CDs en el corpus de entrada.

Es importante señalar dos puntos: 1) para poder determinar la cobertura es necesario conocer de antemano el número total de información relevante (CDs) en el corpus, y 2) un número cercano al 1 indica mejores resultados tanto en precisión como en cobertura.

3.3.1 Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes

En los trabajos de Rebeyrolle (2000) y Rebeyrolle y Tanguy (2000) se describe una metodología para la extracción de CDs a partir de patrones *morfo-sintácticos*. Se presentan, asimismo, algunas consideraciones sobre la introducción de definiciones en textos de especialidad y el diseño de patrones para su extracción automática.

Corpus de estudio o ámbito de aplicación

En este trabajo se recurrió a un corpus de análisis constituido por:

1. 1 manual de Geomorfología: 275,000 palabras
2. 34 artículos científicos en Ingeniería del Conocimiento: 230,000 palabras
3. 2 documentos de una empresa francesa: 205,000 palabras
4. 1 conjunto de artículos extraídos de la Enciclopedia Universal: 215,000 palabras

Metodología

A partir de un estudio sobre la realización de definiciones en textos especializados, los autores proponen una tipología de CDs junto con una representación formal lingüística:

Tipo de CD	Representación formal
De designación	[SNa Vdésigner SNx-X] [N0 Vdésigner SNa SNx-X]
De denominación	[SNx-X Vs'appeler SNa]
De significación	[A Vsignifier B-X]
Introducidos por *es decir*	[A, c'est-à-dire B-X] [B-X, c'est-à-dire A]
De clasificación	[SNa est un Nx-X]
Parentéticos	[SNa (SNx-X)] [SNx-X (SNa)]

Tabla 3.6. Contextos definitorios y su representación en el estudio de Rebeyrolle y Tanguy (tomado de Rebeyrolle y Tanguy (2000: 155))

En la tabla anterior podemos observar algunos ejemplos de los tipos de CDs considerados y los patrones que se utilizaron para su extracción automática. En este caso, *SNa* representa el término a definir; *SNx–X* es el sintagma que ocupa el lugar de la definición; cuando los elementos de la definición no son sintagmas nominales se denotan entonces con *A* y *B–X*. Por ejemplo, en esta formalización la estructura de una definición analítica que incluya un término, un género próximo y una diferencia específica sería: *SNa es un NX–X*.

De esta forma, para extraer CDs se buscó automáticamente el modelo anterior a partir de conformar patrones con 3 verbos o estructuras verbales distintas: patrones con el verbo *definir*, patrones con verbos como *significar* y patrones con el verbo *ser*.

	Patrones definitorios
1	défini$
2	définir * comme
3	définir 6 comme
4	définir (Non Vbe) * comme
5	(signifier | vouloir | dire | entendre)
6	(signifier | vouloir | dire | entendre) * par
7	(est | sont) (un | une | le | la | les | l' | des)
8	être * (det | num)

Tabla 3.7. Ejemplos de patrones definitorios en el estudio de Rebeyrolle y Tanguy (tomado de Rebeyrolle y Tanguy (2000))

En la tabla anterior, en el patrón 1 el símbolo $ significa cualquier número de letras hasta el siguiente espacio en blanco, por lo que este patrón podía recuperar verbos (*define*) o sustantivos (*definición*). En 2, el símbolo * significa cualquier palabra hasta el siguiente adverbio *como*, mientras que en 3 el número 6 indica una ventana específica de seis palabras posibles entre el verbo y el adverbio *como*. En el patrón 4, esta ventana se amplia a cualquier palabra hasta el siguiente *como*, siempre y cuando no sea un verbo. El patrón 5 incluye los verbos *significar*, *querer*, *decir* y *entender* de manera aislada. Por su parte, el patrón 6 añade una ventana de cualquier palabra, enseguida de los verbos mencionados

anteriormente, hasta la siguiente preposición *para*. El patrón 7 es la combinación del verbo ser en 3ª persona singular o plural en presente, seguido de un artículo o determinante. Por último, el patrón 8 es cualquier ocurrencia del verbo *ser* en cualquier persona y tiempo gramatical, seguido de una ventana de palabras hasta el siguiente determinante o adjetivo numeral.

Resultados y evaluación

Rebeyrolle y Tanguy señalan que para analizar la calidad de los resultados obtenidos se basaron en su competencia lingüística. Con ello se pudo conformar manualmente una lista exhaustiva de CDs presentes en el corpus de estudio. Dicha lista contenía 1,574 CDs y se utilizó para medir estadísticamente la efectividad de los patrones definitorios en la extracción de información definitoria.

La evaluación estaba basada en el siguiente principio: se medía la relevancia de los CDs extraídos automáticamente a partir de los patrones definitorios, mediante la comparación con los CDs reales identificados manualmente. Los resultados de precisión (P) y cobertura (C) de los patrones de la tabla 3.7 fueron los siguientes:

	Patrones definitorios	**P**	**C**
1	défini$	5.15%	100%
2	définir * comme	73.33%	100%
3	définir 6 comme	92.59%	90.91%
4	définir (Non Vbe) * comme	91.67%	100%
5	(signifier | vouloir | dire | entendre)	41.94%	100%
6	(signifier | vouloir | dire | entendre) * par	48.60%	100%
7	(est | sont) (un | une | le | la | les | l' | des)	16.83%	93.53%
8	être * (det | num)	3.94%	99.73%

Tabla 3.8. Resultados de precisión y cobertura en el estudio de Rebeyrolle y Tanguy (tomado de Rebeyrolle y Tanguy (2000))

Observamos que los índices están expresados porcentualmente y en este caso la mejor calificación es determinada por un número cercano al 100%. En los patrones 1, 2, 4, 5 y 6 se obtuvo el mejor índice de cobertura, es decir, que todos los CDs que incluían estos

patrones y que estaban presentes en el corpus fueron extraídos automáticamente. En el caso de la precisión, 3 y 4 obtuvieron los índices más altos, lo cual significa que la mayoría de candidatos recuperados eran efectivamente CDs. Por el contrario, con patrones como 1 y 8 se extrajo una cantidad mucho mayor de ruido, ya que menos del 6% de todos los candidatos recuperados automáticamente constituían CDs. Es interesante cómo en el caso del verbo *definir* se incrementó notablemente la precisión con la inclusión de una ventana de palabras hasta el siguiente adverbio *como*, con lo cual se podían recuperar aposiciones o locuciones adverbiales, por ejemplo *se define (generalmente | en términos generales) como*. Cuando la misma ventana se redujo a seis palabras aumentó un poco más la precisión pero bajó la cobertura, lo cual indica que no todos los CDs del corpus se recuperaron automáticamente. En el caso de 7, la restricción del tiempo y persona gramatical en el verbo *ser*, al igual que la no inclusión de una ventana de ocurrencias hasta el siguiente determinante o artículo, aumentó la calidad de la precisión, aunque en ambos casos se obtuvieron índices bajos comparados con los obtenidos por otros patrones definitorios.

En resumen, el trabajo de Rebeyrolle describe las propiedades lingüísticas de patrones definitorios a la hora de extraer automáticamente CDs. La metodología empleada consistió en formalizar patrones con algunos tipos de verbos o estructuras verbales, y en extraer estos patrones sobre un corpus de textos especializados. La evaluación demostró la calidad de ciertos patrones, mientras que otros son propensos a recuperar una gran cantidad de ruido.

3.3.2 DEFINDER

Este sistema, aplicado al inglés, fue desarrollado por Muresan y Klavans (2002) en el *Computer Science Department* de la Universidad de Columbia. La finalidad de esta aplicación era extraer definiciones de textos en-línea orientados a consumidores en el dominio especializado de medicina, y presentar a los pacientes información sobre términos técnicos en un lenguaje sencillo y de fácil entendimiento, sobre todo en definiciones de una especialización intermedia.

Corpus de estudio o ámbito de aplicación

Ya que el ámbito de aplicación de este sistema eran artículos de medicina orientados a pacientes, se seleccionaron artículos de revistas y periódicos, capítulos de libros y manuales escritos por especialistas en un lenguaje común. El corpus de estudio estaba constituido a partir de 5 fuentes: *The Merck Manual of Medical Information*, *Columbia University College of Physician & Surgeons Complete Home Medical Guide*, *Cardiovascular Institute of the South*, *Reuters Health Newspaper for Consumers* y *Medical Industry Today*.

Metodología

La metodología para extraer definiciones consistió en dos módulos principales: un módulo de búsqueda y un módulo de análisis. El primero estaba dedicado a extraer definiciones mediante la búsqueda de patrones léxicos y tipográficos en conjunto con una gramática de estados finitos. Los patrones buscados eran, por ejemplo, *is the term for* o *is called*, así como marcas tipográficas como paréntesis o guiones. En este módulo era necesario además un proceso de filtrado, tomando en cuenta que los patrones podían constituir una fuente de error, ya que se usaban no sólo para definir, sino también para introducir explicaciones y enumeraciones.

El segundo módulo usaba una combinación de estrategias gramaticales y un analizador sintáctico estadístico para encontrar fenómenos lingüísticos usados comúnmente en la redacción de definiciones en textos especializados. Dichos fenómenos incluían aposiciones, cláusulas relativas o anáforas. En este punto, surgía el problema de diferenciar el término y su definición cuando ambos elementos estaban expresados, por ejemplo, en una sola frase nominal (una estructura común en definiciones sinonímicas). Para resolver este problema se utilizaba un método estadístico basado en las frecuencias de los candidatos a término o definición, teniendo en cuenta la hipótesis de que el término se usa un número mayor de veces en el texto, mientras que la definición se suele mencionar una sola vez.

Término	Textos menos especializados	Textos más especializados
Foam cells	White blood cells that have ingested fat	Lipid laden macrophages originating from mono-cytes or from smooth muscle cells

Tabla 3.9. Ejemplos de definiciones en textos orientados a pacientes frente a textos dirigidos a especialistas (tomado de Muresan y Klavans (2002: 231))

En la tabla anterior observamos un ejemplo de los tipos de CDs que eran extraídos automáticamente a partir de textos médicos dirigidos a pacientes, frente a definiciones de otros medios más especializados como UMLS (*Unified Medical Language System*).

Resultados y evaluación

Para analizar los resultados obtenidos, la metodología de evaluación se dividió en 3 partes, cuyos objetivos eran: evaluar el desempeño del sistema para extraer definiciones en términos de precisión y cobertura, evaluar la calidad del diccionario generado automáticamente juzgado por especialistas y no especialistas, y comparar los resultados con diccionarios en línea.

Con respecto al primer punto, señalan las autoras que una finalidad común en cualquier evaluación de un sistema de PLN es la comparación de los resultados obtenidos automáticamente con los resultados que podrían obtenerse de forma manual. Por lo tanto, conformaron un corpus de evaluación con 9 textos que cumplían los mismos criterios del corpus de estudio, y se pidió a 4 sujetos que seleccionaran manualmente todas las definiciones en dichos textos. Se determinó un *gold-estándar* cuando las definiciones encontradas manualmente eran marcadas por al menos 3 de los 4 sujetos. Los resultados de precisión y cobertura fueron los siguientes:

P	C
86.95%	75.47%

Tabla 3.10. Resultados de precisión y cobertura en DEFINDER (tomado de Muresan y Klavans (2002: 232))

Estos resultados indican que un gran número de definiciones se recuperaron automáticamente (75% del total) y en éstas no se extrajo demasiado ruido (sólo el 13% aproximadamente no fueron definiciones).

En cuanto a la calidad del diccionario generado automáticamente juzgado por especialistas y no especialistas, se usó UMLS y OMD (*Online Medial Dictionary*) como recursos de comparación. Se escogieron 8 sujetos no especialistas a los cuales se les asignaron 15 términos médicos aleatorios junto con sus respectivas definiciones de UMLS, OMD y el sistema DEFINDER, sin citar la fuente. La finalidad era asignar una calificación de calidad en cuanto a usabilidad y claridad de la definición, en una escala del 1 al 7, donde 1 era muy malo y 7 excelente. En estos casos, señalan las autoras, *usabilidad* significaba que la definición era útil para entender el término, mientras que *claridad* correspondía al grado de especialización de la definición. Los resultados de este proceso se encuentran en la siguiente tabla:

Usabilidad	DEFINDER UMLS OMD	5.17 2.94 3.9
Claridad	DEFINDER UMLS OMD	5.65 3.18 4.3

Tabla 3.11. Resultados del índice de calidad promedio en DEFINDER (tomado de Muresan y Klavans (2002: 233))

Los resultados demuestran que DEFINDER era mejor calificado por no especialistas, tanto en usabilidad como en claridad de las definiciones, ya que los sujetos de la evaluación consideraban que las definiciones extraídas automáticamente por el sistema eran más fáciles de leer y por tanto más útiles para entender el significado de los términos.

Para contrastar estos resultados, se siguió la misma metodología con 15 sujetos especialistas en medicina, pero en este caso se pidió calificar (igualmente del 1 al 7) la exactitud y la entereza o integridad de las definiciones. Como resultado se obtuvo un promedio de 5.87 para la exactitud, y 5.38 para la entereza,

demostrando que los textos dirigidos a pacientes o no especialistas pueden ser una buena fuente para la extracción de definiciones.

Finalmente, se compararon los resultados de DEFINDER con otras fuentes en línea. La finalidad de esta comparación era resaltar el hecho de que el sistema desarrollado podía servir como complemento a los actuales diccionarios. En este caso, se seleccionaron dos diccionarios especializados, UMLS y OMD, y un glosario llamado *Glossary of Popular and Technical Medical Terms* (GPTMT). Se seleccionaron 93 términos extraídos automáticamente por el sistema junto con sus respectivas definiciones, y se encontraron 3 casos a partir de los resultados:

1. El término estaba listado y definido con la misma definición de DEFINDER en alguno de los diccionarios.
2. El término estaba listado en alguno de los diccionarios, pero la definición no era la del sistema.
3. El término no estaba listado en ninguno de los diccionarios.

Caso	UMLS	OMD	GPTMT
1	60% (56)	76% (71)	21.5% (20)
2	24% (22)	-	-
3	16% (15)	24% (22)	78.5% (73)

Tabla 3.12. Resultados porcentuales de la cobertura en diccionarios en DEFINDER (tomado de Klavans y Muresan (2001: 202))

Observamos que sólo en UMLS se da el caso de que la definición del término no es la misma que la extraída por DEFIDNER. En los demás casos, la definición era la misma o bien el término no se encontraba en el diccionario. El caso particular del glosario GPTMT indica que casi el 80% de las definiciones encontradas por el sistema no habían sido consideradas previamente en el recurso en línea.

En resumen, la metodología de Klavans y Muresan constituye un sistema para la extracción de definiciones, a partir de patrones definitorios, sobre textos dirigidos a no especialistas. Una de las aportaciones más interesantes de este trabajo es la metodología de evaluación, donde se pone en evidencia la dificultad de definir un

gold-estándar a la hora de evaluar lo que se considera como una *buena* definición. Asimismo, resalta la importancia de un sistema de este tipo para el mejoramiento y extensión de recursos léxicos y terminológicos ya existentes, en tanto la extracción automática de definiciones puede complementar dichos recursos con datos que no hayan sido contemplados con anterioridad.

3.3.3 Mining defining contexts to help structuring differential ontologies

El trabajo de Malaisé (2005), una investigación a nivel de tesis de doctorado realizada en la Universitè Paris 7-Denis Diderot, tiene como principal finalidad extraer CDs en francés como un paso previo a la elaboración de ontologías *diferenciales*, las cuales se entienden como estructuras terminológicas jerárquicas normalizadas que constituyen el primer paso en la construcción de cualquier ontología. En este tipo de ontologías, cada término está conectado con su *definición sistemática*, es decir, cada término se asocia a sus principios diferenciales. Estos principios consisten de: a) la similaridad con sus parientes (las características semánticas que el término comparte con su hiperónimo, es decir, el género próximo en las definiciones analíticas); b) la diferencia con sus parientes (las características que distinguen al término de otros que comparten el mismo género próximo); c) la similaridad con los hermanos (las características que el término comparte con sus co-hipónimos); d) la diferencia con los hermanos (las características que lo diferencian de otros co-hipónimos) (Malaisé *et al*. 2005: 22).

Corpus de estudio o ámbito de aplicación

El desarrollo de la metodología se basó en dos corpus, uno de pruebas y otro de evaluación. El primero consistía de documentos pertenecientes a diversos géneros, como descripciones de documentales, extractos de reportes de tesis o documentos Web; estos documentos fueron recuperados en su mayoría de Internet y constaban de 76,000 palabras. El segundo estaba formado por documentos Web en el dominio de dietética y nutrición en diferentes contextos médicos, y constaba de un total de 480,000 palabras en 44,000 oraciones.

Metodología

La metodología de este trabajo, señalan Malaisé *et al*. (2004), se basó principalmente en la extracción de definiciones no especializadas con el fin de compilar información pertinente para la elaboración de una ontología, específicamente sobre el eje vertical (hipónimos), el eje horizontal (co-hipónimos) y el eje transversal (relaciones cruzadas del mismo dominio).

El primer paso en la metodología de Malaisé fue delimitar qué tipo de patrones léxicos utilizaría para la búsqueda automática de contextos definitorios. Tomando como base el trabajo de Auger (1997), determinó que buscaría patrones relacionados con definiciones formales, semi-formales e informales. Este tipo de definiciones podían estar introducidas a partir de:

a) Verbos metalingüísticos (*dénnomer*, *definir*).
b) Marcadores metalingüísticos (*définition*, *nom*) en combinación con verbos (*référer*, *utiliser*).
c) Marcadores discursivos (*c'est-à-dire*, *en d'autres terms*).
d) Signos de puntuación (específicamente *paréntesis*).

Una vez delimitados los patrones, éstos fueron buscados en el corpus de prueba, y una vez obtenidas las ocurrencias se identificaba automáticamente en ellas el término principal de cada contexto (el término que se define), y los términos secundarios (aquellos que se encuentran en la definición, como el caso del género próximo). Esta identificación de los términos constitutivos de los CDs se realizaba mediante dos criterios: uno contextual, que tomaba como referente la posición que podían ocupar los términos de acuerdo con el patrón definitorio, y otro que tomaba como referencia la categoría morfosintáctica de dicho patrón.

Finalmente, una vez identificado el término sobre el cual se aportaba información definitoria, se identificaba automáticamente el tipo de relación conceptual. Para ello, se delimitó previamente el tipo de relación que podía aportar cada uno de los patrones definitorios. Las relaciones podían ser de cuatro tipos: lingüísticas (sinónimos y antónimos), jerárquicas (hiperónimos), transversales (meronimia o relaciones de causa), y horizontales (relaciones que representan a dos o más términos en un mismo nivel).

Resultados y evaluación

Para implementar los procesos de la metodología, el corpus era convertido en XML y los patrones se describían en plantillas XSLT que se aplicaban a dichos corpus. Con estas plantillas era posible extraer los CDs y sus términos constitutivos, así como proponer una relación semántica concreta. Los resultados finales eran presentados para su validación manual a través de un formulario HTML, donde el usuario podía corregir los términos propuestos en las diferentes posiciones (el que se define y los que se encuentran en la definición), o seleccionar una relación semántica distinta a la propuesta por el sistema. En la siguiente figura podemos ver un ejemplo de la interfaz de validación.

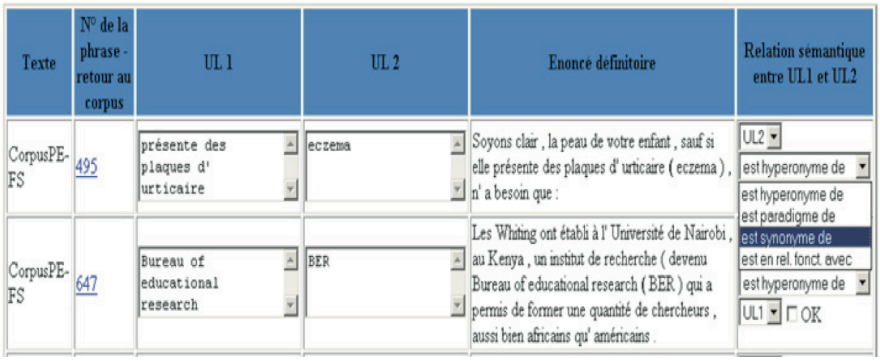

Figura 3.1. Ejemplo de resultados en el estudio de Malisé (tomado de Malaisé *et al*. (2005: 38))

Esta metodología fue evaluada en diferentes partes correspondientes a cada proceso: la extracción automática de CDs, la identificación automática de los términos y la identificación automática de la relación conceptual.

En cuanto a la extracción automática de CDs, se reportan los siguientes resultados obtenidos a partir de un corpus de evaluación, donde el índice de precisión corresponde a una estimación *global*:

P	C
55%	39.3%

Tabla 3.13. Resultados de precisión y cobertura para la extracción de CDs en el estudio de Malaisé (tomado de Malaisé *et al*. (2005: 39-41))

En este caso, señalan los autores que los mejores resultados eran obtenidos por aquellos CDs compuestos por más de un patrón definitorio. La calidad estaba determinada por el número de patrones presentes en el contexto: entre más patrones contuviera, mayor sería la probabilidad de que el contexto fuera un CD. Además, dependía del tipo de patrón que se utilizara, ya que existen algunos que recuperan mejores índices de precisión. Señala, para ejemplificar lo anterior, el caso del patrón *por ejemplo*, que produce más ruido que un patrón como *definido como*.

En la identificación de términos se obtuvo un porcentaje para el índice de precisión, el cual osciló entre el 31% y el 56%:

P
31% - 56%

Tabla 3.14. Resultados de precisión para la identificación de términos en el estudio de Malaisé (tomado de Malaisé *et al*. (2004))

Finalmente, para la identificación automática del tipo de relación conceptual, se tuvo en cuenta si la relación identificada equivalía a la del CD (RS esperada), si tenía otra relación conceptual (RS no esperada), o si la relación no podía determinarse por problemas en la extracción previa del CD (RS indefinida).

Tipo de CD	P
RS esperada	(341) 49.3%
RS no esperada	(351) 50.7%
RS indefinida	(62)

Tabla 3.15. Resultados de precisión para la identificación de RSs en el estudio de Malaisé (tomado de Malaisé *et al*. (2005: 43))

En la tabla anterior, observamos que el número de relaciones identificadas era de 692 (341 *RS esperada* y 351 *RS no esperada*), mientras que el número de relaciones indefinidas (*RS indefinida*) fue de 62 casos. Esto quiere decir que un mayor porcentaje de las relaciones presentes en el texto fue identificado automáticamente, aunque casi la mitad de estas fueron clasificadas como una relación

no esperada. Al respecto, se hace hincapié en la necesidad de un refinamiento de las reglas a partir de la reasignación del tipo de relación que se puede recuperar con cada patrón. Tal es el caso de los *paréntesis* como patrón tipográfico que puede introducir contextos léxico-sintácticos idénticos, pero en los cuales la relación puede ser interpretada como hiperonimia, sinonimia, tratamiento médico o relaciones más dependientes del contexto, como la conexión entre una enfermedad y el lugar geográfico donde ocurre.

En resumen, el trabajo de Malaisé introduce una metodología para la extracción de CDs como paso previo a la construcción de ontologías diferenciales. Por ende, en su estudio se presta interés no sólo a las descripciones acerca del significado de un término que pueden extraerse mediante la búsqueda automática de CDs, sino a las relaciones semánticas específicas que se pueden dar entre el término que se define y los términos de la definición. Este estudio resalta la idea de utilizar patrones contextuales para la identificación del término que se define en el contexto, es decir, las posibilidades de que el término aparezca a izquierda o derecha respecto al patrón definitorio que lo liga con su definición.

3.3.4 Hacia un sistema de extracción de definiciones en textos jurídicos

El trabajo de Sánchez y Márquez (2005), aplicado al español, conforma un primer acercamiento para establecer un mecanismo de extracción de definiciones en textos jurídicos. La finalidad de este estudio es extraer definiciones mediante patrones recurrentes y constituir una base de datos donde puedan manipularse y consultarse las definiciones mediante una interfaz de usuario.

Corpus de estudio o ámbito de aplicación

Este trabajo se realizó específicamente para el dominio jurídico de la Ley Orgánica del Trabajo de Venezuela.

Metodología

La metodología de estos autores consistió, en primer lugar, en realizar un análisis lingüístico de textos jurídicos para determinar

qué tipo de patrones suelen introducir definiciones. Se identificaron tres grupos para las formas verbales *entenderse*, *ser* y *ser considerado*. Para sus experimentos decidieron trabajar con patrones recurrentes de las formas en presente de *entenderse*, entre las que se incluyen *se entiende por* y *se entiende como*. El sistema propuesto seguía, a grandes rasgos, los siguientes pasos:

1. Identificar oraciones dónde estuviera presente un marcador de definiciones.
2. Etiquetar las partes de la oración.
3. Extraer el *definens*[7].
4. Extraer su definición.
5. Almacenar el *definens* con su definición.

Es interesante señalar que el etiquetado se realizaba de la siguiente forma: en primer lugar se etiquetaban todas las clases de palabras pertenecientes a un inventario cerrado, por ejemplo artículos, adjetivos, pronombres, conectores discursivos y reformuladores (los patrones definitorios), y también signos de puntuación, como dos puntos y comas. En segundo lugar, todo lo que no pertenecía a las categorías definidas para el inventario anterior era etiquetado como *vacío*.

Después del etiquetado se reconocía automáticamente el *definens* y su definición. Para ello se consideraba la estructura del patrón definitorio y se seguían algunas reglas que ayudaran a distinguir distintos casos en que el *definen* pudiera estar separado de su definición. Esto último lo ejemplifican mediante el patrón *se entiende por*, que normalmente presenta al término después de dicho patrón y está separado de su definición por un signo de puntuación, por ejemplo *se entiende por X, Y*.

De esta forma, algunas reglas para detectar los casos en que el *definens* estaba separado de su definición, una vez etiquetada la oración, eran:

1. Un artículo después de un *vacío*.
2. Una coma después de un vacío y un artículo después de una coma.

[7] Lo que en nuestro estudio denominamos *término*.

3. Dos puntos después de un vacío.
4. La preposición *a* después de un vacío.

Resultados y evaluación

Para evaluar su propuesta, Sánchez y Márquez tomaron como referencia los índices de precisión y cobertura. Para ello, previamente identificaron todas las definiciones que presentaran el patrón *entenderse* en los textos jurídicos que analizaron, donde encontraron un total de 38 definiciones para dicho patrón.

P	C
97.44%	100%

Tabla 3.16. Resultados de precisión y cobertura en el estudio de Sánchez y Márquez (tomado de Sánchez y Márquez (2005))

Los resultados señalan que todas las definiciones identificadas manualmente para los patrones derivados de la forma *entenderse* eran identificadas automáticamente (cobertura); sin embargo, identificaron un contexto *no definitorio* de más, lo que supuso que la precisión no fue del 100%.

Es importante destacar, como bien señalan los autores, que esta evaluación se realizó sin tomar en cuenta que en algunos casos existían oraciones que completaban las definiciones y que no se extraían automáticamente, lo cual reduciría los porcentajes de precisión y cobertura si se tomaran en cuenta para la evaluación.

En síntesis, en este primer acercamiento se presenta una metodología para la extracción de definiciones en textos jurídicos en español. Se pone de manifiesto que, con un mínimo coste de etiquetado (en términos de procesamiento automático), se podía identificar el término y su definición.

3.3.5 Automated detection and annotation of term definitions in German text corpora

El estudio de Storrer y Wellinghoff (2006), desarrollado para el alemán, muestra una metodología para detectar y anotar

automáticamente definiciones en textos técnicos a partir de verbos definitorios y patrones basados en la valencia de dichos verbos. La finalidad consistía en detectar contextos que tuvieran definiciones y en anotar sus componentes principales: el término que se define y su significado. Además se pretendía extraer relaciones semánticas (tipo Word-Net) que se encontraran entre el término que se define y los términos presentes en la definición.

Corpus de estudio o ámbito de aplicación

El corpus de estudio consistió de 20 documentos técnicos en el dominio de tecnologías del texto con un total de 103,805 palabras.

Metodología

El primer paso consistió en la anotación manual de definiciones en el corpus de estudio. El modelo de anotación comprendía tres elementos: el término que se define, el significado postulado para dicho término y el verbo que relaciona a los dos elementos anteriores. Asimismo, este modelo de anotación se concentraba en definiciones con patrones correspondientes a definiciones aristotélicas o analíticas. Como resultado, se anotaron 174 definiciones en el corpus, las cuales fueron usadas como gold-estándar y como base empírica en la factibilidad del estudio de extracción de relaciones semánticas a partir de definiciones.

Acto seguido se definió un paradigma de 19 patrones definitorios con verbos y formas verbales como *ser*, *conocer como*, *llamar*, *definir como*, *conocido como* o *hablar de* (Storrer y Wellinghoff 2006: 2375). Este paradigma se implementó a partir de un sistema denominado *Insight Discoverer Extractor*[8], el cual permitía definir conceptos generales para los componentes principales del modelo de anotación de definiciones, y especificar para cada patrón la valencia o lugar correspondiente al término y la definición.

En la extracción de RSs se utilizaron las definiciones extraídas y anotadas automáticamente, sobre las que se aplicaron reglas para extraer las relaciones que ocurrían entre el término y los términos de la definición. Si el patrón definitorio era analítico, entonces la

[8] http://www.temis-group.com/

regla indicaba que el término definido era el hipónimo (subclase) del término en la posición de género próximo en la definición.

Resultados y evaluación

En la siguiente tabla presentamos los resultados de la etapa correspondiente a la extracción de CDs:

Patrón	Ocurrencias	P	C
sein	80	31%	83%
bezeichen als	16	43%	75%
verstehen unter	13	100%	85%
nennen	10	100%	20%
bestehen aus	7	41%	100%
spezifizieren als	4	100%	100%
heißen	3	50%	100%
verwenden als	3	9%	100%
bedeuten	2	11%	100%
beschreiben	2	33%	100%
begreifen als	1	100%	100%
benennen	1	100%	100%
charakterisieren als	1	100%	100%
definieren als	1	100%	100%
gebrauchen	1	50%	100%
sprechen von	1	50%	100%
Terminus einführen	1	100%	100%
vorstellen als	1	100%	100%
bekant als	1	50%	100%
total	149	34%	70%

Tabla 3.17. Resultados de precisión y cobertura en el estudio de Storrer y Wellinghoff (tomado de Storrer y Wellinghoff (2006: 2375))

Los resultados obtenidos, señalan Storrer y Wellinghoff, eran dependientes del patrón definitorio. La cobertura era significativamente mayor cuando la definición aparecía precedida

de una preposición que formaba parte del patrón definitorio, como en el caso de *entender por* o *especificar como*.

Mencionan los autores que los resultados de los patrones que ocurrían una sola vez no eran significantes y tendrían que ser evaluados en un corpus de mayor tamaño. Por último, resaltan el caso problemático del patrón *ser*, que puede ser utilizado en una gran variedad de contextos que no aportan información definitoria.

En resumen, con la metodología expuesta en este trabajo era posible extraer y anotar automáticamente definiciones a partir de patrones verbales definitorios, y posteriormente buscar algún tipo de relación semántica específica. Cabe resaltar la idea de recurrir a las valencias o posiciones que pueden ocupar el término y la definición dependiendo del verbo que los conecta, una idea parecida a la utilización de patrones contextuales por parte de Malaisé.

3.3.6 MOP

Otro estudio en la línea de extraer información definitoria es el descrito en Rodríguez (2005). Este autor describe un sistema denominado *MOP* (*Metalinguistic Operator Processor*), desarrollado para el inglés, el cual tiene la finalidad de extraer unidades de conocimiento especializadas (conocidas como *OMEs* (*Operaciones Metalingüístias Explícitas*) a partir de la detección de fragmentos metalingüísticos en textos de especialidad. Las OMEs se consideran *operaciones explícitas*, ya que el autor introduce indicaciones que aportan información sobre la forma en que debe entenderse el término, y son consideradas *metalingüísticas* porque se utiliza el lenguaje para hablar del propio lenguaje. De esta manera, las OMEs son en cierta medida contextos definitorios donde se puede encontrar no sólo información que sirva para entender el significado de un término, sino también información pragmática, por ejemplo sobre su origen, direcciones o condiciones de uso[9].

[9] Tómese como ejemplo el siguiente contexto donde se aporta información sobre la evolución histórica de un término: "In 1965 the term soliton was coined to describe waves with this remarkable Behaviour". (Tomado de Rodríguez 2004: 20).

Corpus de estudio o ámbito de aplicación

El corpus para obtener patrones metalingüísticos consistió en 19 artículos en el área de sociología (5,581 oraciones) de revistas académicas. Con el fin de extender el paradigma inicial de patrones se recurrió a las áreas de enseñanza y ciencia del *British National Corpus*. Para la evaluación se utilizó además un libro de texto en línea en el área de histología (5,146 oraciones) y una muestra de resúmenes tomados de la base de datos MedLine (1,403 oraciones).

Metodología

Para identificar automáticamente este tipo de enunciados definitorios, Rodríguez (1999) delimitó en primera instancia los elementos constitutivos de las OMEs: un autónimo[10], información semántica-pragmática y marcadores-operadores. El primero es el sujeto lógico de la oración. Mediante el segundo se proporciona información semántica, instrucciones de uso e interpretación del término. Y finalmente, la función del tercero es la de relacionar los dos elementos anteriores a partir de verbos metalingüísticos como *to define*, *to refer*, *to denominate*, y marcadores tipográficos que cumplen una función discursiva, por ejemplo a través de fórmulas como *comillas* + *término*, que pueden proveer información sobre el término de manera económica.

Las OMEs se clasifican de manera general en dos tipos:

1. Informativas o Directivas, las cuales aportan información intencional y extensional sobre el término. Suelen conformar estructuras del tipo: *X is calledY*, *X implies Y in the context Z*, *X = Y*.
2. Instruccionales, en las cuales se hace explícita la connotación que el lector debe comprender para un término determinado.

A partir de esta tipología, Rodríguez buscó manualmente en una primera instancia todas las ocurrencias de OMEs en el corpus de sociología. Se encontró así enunciados que podían contener patrones con verbos específicos (*called*, *coined*), frases verbales (*defined as*, *known as*), descriptores (*word*, *term*) o marcadores no

[10] Lo que aquí se considera como *término*.

léxicos (*comillas*). Posteriormente buscó estos patrones en el *British Nacional Corpus* y obtuvo un total de 10,937 fragmentos con alguna ocurrencia de dichos patrones, de las cuales el 50% aproximadamente eran OMEs.

Posteriormente, con el total de las ocurrencias (válidas y no válidas) constituyó un corpus que después analizó automáticamente mediante algoritmos de aprendizaje para identificar combinaciones recurrentes que presentaran algún fragmento no metalingüístico, con la idea de conformar una base para realizar un filtrado automático. En los resultados de esta etapa no se recuperaban automáticamente una gran cantidad de OMEs, ya que faltaban aproximadamente un 30%. Esto podía deberse al hecho de utilizar una lista no exhaustiva de patrones metalingüísticos.

Por tanto, el siguiente paso fue especificar los patrones (p. ej. el patrón *called* lo delimitó a *calls*, *called*, *call*) con lo que pudo aumentar los resultados notablemente. Una vez realizado este proceso, las OMEs eran etiquetadas morfosintácticamente, y mediante reglas sintácticas, pragmáticas y de estructura de argumentos se identificaban automáticamente cuáles eran los elementos constitutivos con el fin de llenar los campos de una base de datos de información metalingüística (Rodríguez 2004).

Resultados y evaluación

Para realizar la evaluación, el autor clasificó manualmente las OMEs en dos subcorpus en las áreas de sociología e histología. En el proceso de identificación de las OMEs mediante la búsqueda de patrones metalingüísticos, se obtuvieron los siguientes índices:

Corpus	P	C
Sociología	0.94	0.68
Histología	0.9	0.5

Tabla 3.18. Primeros resultados de precisión y cobertura para identificar OMEs en el estudio de Rodríguez (tomado de Rodríguez (2004: 18))

En este caso, se observan índices bajos de cobertura, lo cual podía estar relacionado con la lista de patrones no exhaustiva que se empleó en esta primera etapa. Por ello, se evaluó la extracción sobre

un gold-estándar donde se eliminaron oraciones que no tuvieran uno de los patrones utilizados, con la idea de tener un panorama más realista del funcionamiento del sistema con la lista actual de patrones. Además, se añadió el índice *F1 score* (también conocido como *F-score* o *F-measure*), que a grandes rasgos es una medida utilizada para dar una perspectiva en balance entre los índices de precisión y cobertura. De esta manera, los índices fueron los siguientes, donde se observa un incremento sobre todo en el índice de cobertura:

Corpus	P	C	F1
Sociología	0.97	0.79	0.87
Histología	0.94	0.81	0.87

Tabla 3.19. Segundos resultados de precisión y cobertura para identificar OMEs en el estudio de Rodríguez (tomado de Rodríguez 2004: 18))

En resumen, Rodríguez propone las OMEs como operaciones comunicativas especializadas donde se puede encontrar información definitoria y pragmática sobre un término. La extracción automática de OMEs tiene como finalidad poblar *bases de datos metalingüísticas* (MIDs: Metalinguistic Information Databases), las cuales deben ser vistas no como un producto final, sino como un recurso semiestructurado que puede ser usado como ayuda en las distintas facetas del trabajo lexicográfico y terminográfico.

3.3.7 Mining online sources for definitional knowledge

En Saggion (2004) se presenta un estudio para el inglés enfocado a la extracción de definiciones para sistemas de pregunta-respuesta. Este trabajo parte de la necesidad de encontrar respuestas de manera automática a preguntas del tipo *¿Qué es X?* o *¿Quién es X?*, y se enfoca no sólo a dominios especializados sino también a lengua general.

Corpus de estudio o ámbito de aplicación

El ámbito de aplicación de este trabajo fue principalmente la web, de donde se obtuvieron ocurrencias de términos para buscar patrones definitorios utilizando Google, WordNet y la Enciclopedia

Británica. La evaluación se llevó a cabo en una colección de textos llamada AQUAINT, conformada por cerca de 1 millón de textos del *New York Times*, la red de noticias *AP*, y la parte en inglés de la red de noticias *Xinhua*.

Metodología

En este caso, la metodología para la extracción de definiciones partía de una consulta específica con la forma de una pregunta, por lo que el primer paso consistía en identificar cuál era el término sobre el que se tenía que extraer información definitoria. Una vez identificado el término, se buscaban y extraían automáticamente las ocurrencias de éste en dos recursos lexicográficos: WordNet y la Enciclopedia Británica. Por otro lado, previamente se había conformado manualmente una lista de 50 patrones definitorios, entre los que se incluían formas como *TERM is a*, *such as TERM* o *like TERM*. Estos patrones se combinaban con el término y se formulaban expresiones que eran buscadas directamente en la web.

El proceso de adquirir ocurrencias del término de búsqueda en los recursos lexicográficos y en la web tenía la finalidad de adquirir términos secundarios que suelen aparecer en las definiciones del término de búsqueda. Entre más ocurrencias tuvieran los términos secundarios en las definiciones, era más probable que dichos términos formaran parte de la definición del término de búsqueda.

De esta manera, se conformaban grupos donde el término de búsqueda *X* estaba relacionado con un grupo de términos *Y* comunes en su definición. Finalmente, para poder encontrar información definitoria del término de búsqueda, se buscaban en el corpus todos aquellos enunciados que contuvieran X y las ocurrencias más relevantes del grupo Y.

Resultados y evaluación

El autor reporta los resultados de aplicar esta metodología en el ámbito de una competencia denominada *TREC QA 2003*. La evaluación consistía en resolver 30 preguntas *Quién* y 20 preguntas *Qué*. Para cada respuesta se tenía una lista previa de fragmentos textuales considerados como relevantes. Estos fragmentos se comparaban con la información obtenida automáticamente y se

evaluaban con la medida *F1 score*. De esta forma se obtuvo como resultado el siguiente índice:

F1
0.236

Tabla 3.20. Resultados de F-score en el estudio de Saggion (tomado de Saggion (2004: 29))

En el escenario de la competencia, 0.555 correspondía al mejor resultado, 0.192 al resultado intermedio, y 0.000 era el peor. Esto quiere decir que el resultado sobrepasaba la media de 0.192 y se podría deducir que en poco más de la mitad de los casos se pudo obtener automáticamente una definición para preguntas del tipo "What is X" y "Who is X".

En resumen, en el trabajo de Saggion se presenta una metodología para responder a preguntas a partir de la identificación del término de búsqueda y su posterior combinación con patrones definitorios. El conjunto de término + patrón definitorio era buscado tanto en recursos léxicos y enciclopédicos, como en documentos en línea, con el fin de encontrar términos secundarios que suelen ocurrir en las definiciones del término de búsqueda. Cabe resaltar la idea de que este trabajo constituye un esfuerzo por integrar diferentes recursos léxicos para la misma búsqueda de información definitoria.

3.3.8 LT4eL

Language Technology for eLearning (LT4eL) fue un proyecto coordinado por la Universidad de Utrecht, Holanda, en conjunto con 11 instituciones educativas[11] y patrocinado por la unión europea. En este proyecto se trabajó en alemán, búlgaro, checo, holandés, inglés, maltés, polaco, portugués y rumano.

[11] Universidad de Utrecht, Holanda; Universidad de Hamburgo, Alemania; Universidad "Al.I.Cuza" de Iasi, Rumania; Universidad de Lisboa, Portugal; Universidad Charles de Praga, República Checa; IPP, Academia Búlgara de Ciencias, Bulgaria; Universidad de Tübingen, Alemania; ICS, Academia Polaca de Ciencias, Polonia; Universidad de Ciencias Aplicadas de Zurich, Suiza; Universidad de Malta, Malta; Eidgenössische Hochschule, Suiza; Open University, Reino Unido.

En los 30 meses de duración de este proyecto, entre 2005 y 2008, se integraron herramientas multilingües y técnicas de Web semántica para la recuperación automática en Internet de material de apoyo en la enseñanza. La finalidad de este proyecto consistió en facilitar un acceso personalizado a contenidos educativos a través de sistemas para el manejo de la enseñanza y la cooperación en el manejo de contenidos en línea. Una parte central del proyecto se enfocó en desarrollar metodologías para la extracción automática de definiciones con el fin de proporcionar herramientas de ayuda en la elaboración de glosarios. (Monachesi 2007).

Corpus de estudio o ámbito de aplicación

En alemán se utilizó un corpus de estudio en el ámbito de derecho, compuesto por más de 6,000 veredictos en leyes ambientales. Para las demás lenguas se conformó un corpus de textos especializados principalmente en el dominio de ciencias computacionales y aprendizaje electrónico (*eLearning*). Algunas cifras específicas de algunos corpus de estudio y/o evaluación son las siguientes:

1. Alemán: 237,935 oraciones.
2. Búlgaro: 76,800 palabras.
3. Checo: 90,000 palabras.
4. Holandés: 77 textos con un promedio de 6,568 palabras cada uno.
5. Portugués: textos de 3 áreas distintas con 274,000 palabras en total.
6. Rumano: 56 documentos con aproximadamente 700,000 palabras .

Metodología

Las metodologías para la extracción de definiciones en las lenguas de este proyecto, excepto en alemán, partieron de la base común de definir manualmente gramáticas para representar distintos tipos de patrones que pudieran servir como claves en la identificación automática de definiciones (Borg 2007). A partir del corpus en los ámbitos de ciencias computacionales y de aprendizaje electrónico, se obtuvo un grupo de definiciones que fue anotado en XML. Las definiciones de este grupo se categorizaron en seis tipos:

1. Definiciones con el patrón IS-A.
2. Definiciones con patrones como *significa*, *es definido*, *es llamado*.
3. Definiciones con signos de puntuación como conectores entre el término y la definición.
4. Definiciones resaltadas por algún tipo de tipografía textual, por ejemplo cuando se encuentran en *viñetas*, *listas* o *tablas*.
5. Definiciones que contienen un pronombre que funciona como referencia anafórica; por ejemplo, en casos donde se presenta el término en un párrafo y se define en el siguiente, a partir de una referencia como *este concepto se define como*.
6. Otro tipo de definiciones que no encajan en ninguno de los tipos anteriores, por ejemplo aquellas donde el conector es un marcador reformulativo como *es decir*.

De estos tipos de definiciones se pudieron desarrollar gramáticas específicas para las distintas lenguas del proyecto. Cada gramática era un documento XML que presentaba una estructura similar de reglas con expresiones regulares o referencias a otras reglas (Del Gaudio y Branco 2007: 662). Dichas reglas podían ser de cuatro tipos:

1. Reglas simples para capturar sustantivos, adjetivos y preposiciones, etc.
2. Reglas para identificar verbos y patrones verbales.
3. Reglas para identificar secuencias sintácticas, como frases nominales o frases preposicionales.
4. Reglas complejas que combinan las anteriores, con el fin de identificar términos y definiciones.

La idea detrás de las gramáticas era conformar una secuencia de reglas que pudiera ser aplicada en los textos para extraer candidatos a definiciones. Además, las gramáticas iniciales se utilizaron para aprender automáticamente la importancia de cada patrón definitorio mediante un algoritmo genético (Borg 2007; Borg *et al*. 2007), al igual que implementar reglas de aprendizaje automático para mejorar el desempeño de la extracción (Westerhout y Monachesi 2008; Degórski *et al*. 2008; Kobyliński y Przepiórkowski 2008).

Por su parte, en el proyecto en alemán, Walter y Pinkal (2006) comenzaron por analizar manualmente una muestra aleatoria de 40

veredictos en leyes ambientales, donde se encontraron 130 definiciones. A partir de este análisis manual detectaron que las definiciones en el dominio de análisis suelen contener cinco elementos constitutivos:

1. El término que se define.
2. La información definitoria.
3. Un conector que indica el tipo de relación entre el término y la información sobre su significado.
4. Información sobre el área de dominio donde se aplica la definición, por ejemplo mediante frases como *en el área de*.
5. Palabras a las cuales no se les puede asignar ninguna función específica, pero que pueden ayudar en la identificación de la definición; por ejemplo, la frase *el término*.

Una vez identificadas las estructuras recurrentes de las definiciones, éstas fueron procesadas con un analizador que genera una representación semántica mediante un análisis en cascada de los componentes de una oración, y da como resultado un árbol XML con información lingüística completa. A partir de estos resultados declararon reglas para la extracción de definiciones basadas en sus elementos constitutivos, así como reglas de filtrado para excluir, por ejemplo, casos donde el término equivale a una anáfora.

Resultados y evaluación

La metodología para la evaluación de los resultados en todas las lenguas, excepto alemán, partió de la anotación manual de las definiciones en un corpus de prueba. El número total de definiciones era contrastado con el número de definiciones extraídas automáticamente. Para ello, utilizaron los índices de precisión y cobertura junto con una medida denominada *F2*, la cual es una variante de la medida *F1* o *F-score* que se utiliza en aquellos casos donde es más importante tener un panorama del desempeño de la cobertura frente al de precisión.

En portugués, por su parte, se empleó el mismo corpus de pruebas para la evaluación. En este caso, el corpus estaba compuesto por documentos en las áreas de *Information Society (IS)*, *Information Technology (IT)* y *eLearning*.

Los resultados globales se muestran en la siguiente tabla, donde se puede observar que los índices obtenidos en los corpus *IS* e *IT* superan a los índices del corpus *eLearning*. Señalan los autores que esto se puede deber al estilo y el propósito de cada tipo de corpus: los primeros contienen un enfoque educativo y por lo tanto una estructura más formal, mientras que el tercero presenta una estructura donde se enmarcan definiciones menos explícitas.

Corpus	**P**	**C**	**F2**
IS	0.14	0.86	0.32
IT	0.33	0.69	0.51
eLearning	0.11	0.59	0.24
Total	0.14	0.86	0.66

Tabla 3.21. Resultados de precisión y cobertura en LT4eL (portugués) (tomado de Del Gaudio y Branco (2007: 666))

En este caso, los autores reportan errores como resultado de una mala anotación morfosintáctica del corpus de prueba y señalan que este tipo de evaluación cuantitativa debería complementarse con una evaluación cualitativa para tener un panorama más completo del desempeño del sistema.

La gramática, en el caso del rumano, se evaluó con los 4 primeros tipos de definiciones señalados: definiciones con el patrón IS-A: *is_def*; definiciones con patrones como "significa": *verb_def*; definiciones con patrones tipográficos: *punct_def*; y definiciones con patrones de tipografía textual: *layout_def*.

Patrón	**P**	**C**	**F2**
is_def	0.53	1	0.77
verb_def	0.75	1	0.90
punct_def	0.14	1	0.33
layout_def	0.04	1	0.002

Tabla 3.22. Resultados de precisión y cobertura en LT4eL (rumano) (tomado de Iftene *et al*. (2007: 22))

En la tabla anterior observamos que los mejores resultados de precisión fueron obtenidos para los patrones con verbos definitorios

distintos al patrón IS-A. Este último, señalan los autores, es más conflictivo por la frecuencia de aparición del verbo *ser*, por lo cual arroja como resultado una mayor cantidad de enunciados no definitorios. En el caso de la cobertura, resulta interesante notar que todos los CDs del corpus fueron extraídos automáticamente.

En el caso de las lenguas eslavas (búlgaro, checo y polaco), se llevó a cabo una evaluación cuantitativa, igualmente partiendo de un corpus donde fueron anotadas manualmente las definiciones. Una definición extraída automáticamente era considerada buena si ésta coincidía o se solapaba con una definición detectada manualmente.

Lengua	**P**	**C**	**F2**
Búlgaro	22.5%	8.9%	11.1
Checo	22.3%	46%	33.9
Polaco	23.3%	32%	28.4

Tabla 3.23. Resultados de precisión y cobertura en LT4eL (lenguas eslavas) (tomado de Przepiórkowski *et al*. (2007: 47))

Los resultados en estas lenguas conllevan un problema específico a la metodología multilingüe seguida en el proyecto, ya que las gramáticas contemplan que las definiciones suelen estar constituidas, estructuralmente, por o dentro de una oración; mientras que, en las lenguas eslavas, las definiciones suelen conformar estructuras multi-oracionales. Por ejemplo, en búlgaro el 36% de las definiciones identificadas manualmente estaban formadas por más de una oración.

Por su parte, en holandés se obtuvieron los siguientes índices:

Patrón	**P**	**C**	**F1**	**F2**
is_def	20.97	91.80	34.15	43.19
verb_def	25.76	41.46	31.78	34.46
punct_def	2.58	76.92	4.99	7.25
layout_def	6.15	40.74	10.68	14.16

Tabla 3.24. Resultados de precisión y cobertura en LT4eL (holandés) (tomado de Westerhout y Monachesi (2007))

Los autores incluyeron en estos resultados ambos índices *F1* y *F2*. Igualmente, se hace hincapié en la cantidad de ruido que se puede recuperar con el patrón IS-A. No obstante, en este caso los patrones con verbos del tipo *significar*, *definir* o *entender* (verb_def) recuperaron igualmente una cantidad notable de ruido, a la vez que la cobertura fue menor al 50% de los casos. Según los autores, esto se debe al tipo de verbos que se incluyeron en este grupo. Se incluyó, por ejemplo, el verbo *prevenir*[12], que al igual que el verbo *ser* puede emplearse en una gran cantidad de oraciones..

Como resultado de la metodología en alemán, se generaron 33 reglas de extracción, que aplicadas al corpus de 6,000 veredictos arrojaron como resultado un total de 5,461 candidatos. Después de aplicar las reglas de filtrado, el número de candidatos se redujo a 1,486 casos, de los cuales se seleccionaron 473 para la evaluación. Estos candidatos debían ser analizados manualmente por 2 individuos para decidir si eran o no definiciones, tomando en cuenta un criterio *amplio* respecto al concepto de definición. Lo resultados de precisión fueron los siguientes:

Individuo	P
1	44.6%
2	48.6%

Tabla 3.25. Resultados de precisión en LT4eL (alemán) (tomado de Walter y Pinkal (2006: 24))

De los resultados observamos que se analizó únicamente el índice de precisión partiendo del análisis manual hecho por dos evaluadores. Este índice obtuvo un resultado menor al 50%, en ambos casos, lo cual señala que poco menos de la mitad de los candidatos resultaron ser definiciones.

A pesar de no recurrir al índice de cobertura, los autores señalan que la cantidad de definiciones extraídas sobre el total de oraciones

[12] Respecto a la inclusión de este tipo de verbos, los autores señalan el siguiente ejemplo y la siguiente explicación: "'A non-breaking space prevents a line from being splitted between two words'. Whereas not everybody will consider this as a definition, they probably will consider the next sentence, which contains the same information, as a definition: 'A non-breaking space is a space that prevents a line from being splitted between two words'". (Westerhout y Monachesi 2007)

del corpus de evaluación (130 definiciones en 3,500 oraciones), es un indicio de que el número de reglas usadas parece estar lejos de tener un buen resultado en términos de cobertura, por lo cual debe reconsiderarse el número y tipo de patrones definitorios empleados.

En resumen, *LT4eL* fue un proyecto multilingüe donde se enmarcaba la extracción automática de definiciones como un proceso de ayuda en la elaboración de contenidos para el aprendizaje electrónico. La metodología general consistió en desarrollar gramáticas particulares para cada lengua a partir de patrones tanto verbales como tipográficos. La evaluación fue principalmente de tipo cualitativa, recurriendo a la anotación manual de las definiciones en los corpus de prueba para comprobar el funcionamiento de la metodología.

3.3.9 GlossExtractor

En Navigli y Velardi (2007) se describe una aplicación web desarrollada para el inglés y denominada GlossExtractor[13], cuya función es extraer una lista de candidatos a definiciones sobre varios tipos de documentos en Internet (glosarios en línea, documentos web o páginas especificadas por un usuario), a partir de una lista de términos proporcionada previamente. Este sistema permite al usuario definir el ámbito de búsqueda y los términos sobre los cuales se pretende extraer información definitoria.

Corpus de estudio o ámbito de aplicación

El desarrollo de la metodología expuesta en el trabajo de Navigli y Velardi contempla a la web como corpus sin delimitar un ámbito de estudio particular. Sin embargo, en una etapa de evaluación del sistema se utilizó un corpus de prueba que constaba de 1,000 definiciones en el área de economía, y de 250 definiciones para el área de medicina extraídas de documentos y glosarios en línea.

[13] http://lcl.uniroma1.it/glossextractor

Metodología

Este sistema tiene como antecedente inmediato el sistema *OntoLearn* (Cucchiarelli *et al*. 2004), que sirve para obtener automáticamente una ontología a partir de documentos compartidos por miembros de una comunidad en línea. Esta técnica incluía 3 pasos: 1) la extracción automática de términos, 2) la elaboración automática de glosarios, y 3) el enriquecimiento de una ontología. Con el fin de hacer accesibles los procesos automáticos que forman parte de OntoLearn, los autores comenzaron a desarrollar aplicaciones web, la primera de ellas fue TermExtractor, una aplicación en línea para la extracción automática de términos.

GlossExtractor está compuesto por tres módulos: 1) extracción de candidatos a definiciones, 2) identificación de los mejores candidatos basado en filtros estilísticos y 3) un filtro basado en el dominio. En la siguiente figura se puede observar una representación de la arquitectura general.

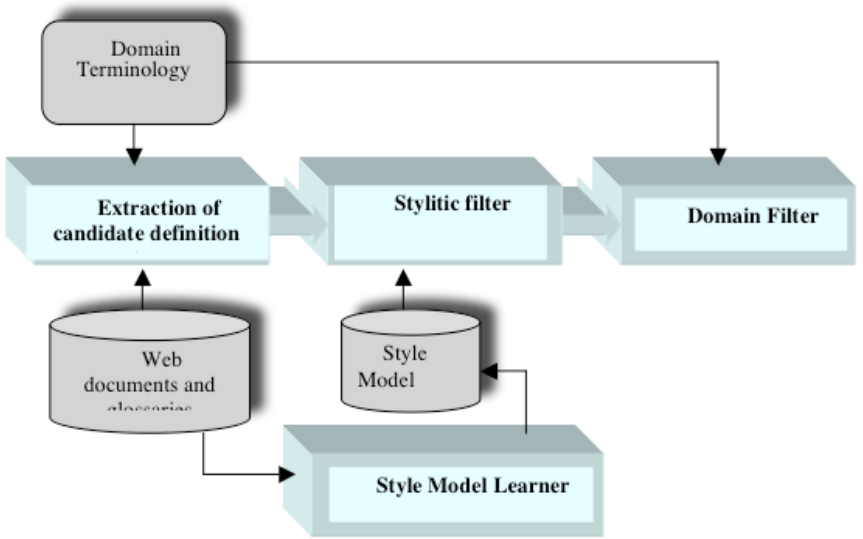

Figura 3.2. Arquitectura general de GlossExtractor (Tomado de Navigli y Velardi (2007: 340))

El sistema parte de una lista de términos introducida por un usuario. Esta lista se busca en primer lugar en glosarios en línea indexados previamente. En seguida, por cada término de la lista se generan una serie de patrones de búsqueda, por ejemplo "*<term> is a*",

"*<term> defines*", "*<term> refers to*", "*<term> is a kind of*". Cada secuencia conforma una expresión de búsqueda que se lanza a Google, y por cada una de ellas se bajan las primeras cinco páginas y se convierten a texto plano. Una vez obtenidos estos documentos, en ellos se busca cada oración que contenga el término de entrada. Esto último, como señala Velardi *et al*. (2008), resulta en un aumento considerable de ruido, principalmente en oraciones *no definitorias* u oraciones que contienen *definiciones fuera de contexto*, es decir, oraciones que si bien contienen información definitoria, ésta no pertenece al ámbito del término que se busca[14].

Los siguientes dos pasos consisten en tratar de eliminar el ruido. El primero de ellos es un filtro estilístico que busca las definiciones más canónicas, es decir, aquellas que presentan un modelo analítico donde se hace explícito el género próximo del término y sus características distintivas. La justificación de lo anterior radica en que, con ello, se obtienen definiciones que contienen un estilo uniforme comúnmente adoptado por lexicógrafos; se distingue entre definiciones y no definiciones, especialmente cuando los candidatos a definición son extraídos de texto irrestricto, en lugar de glosarios en línea; y se extrae de las definiciones información que permite reorganizar los resultados de manera taxonómica. Cabe señalar que los filtros para extraer definiciones canónicas se obtuvieron a partir de un algoritmo de aprendizaje automático[15] que permitió generar un árbol de decisión para llevar a cabo este proceso de filtrado.

El último paso consiste en un filtro que elimina definiciones que no son pertinentes al dominio de interés. Esto se logra obteniendo un modelo probabilístico a partir de analizar el dominio de la terminología de entrada y de asignar una medida de pertenencia a cada uno de los candidatos a definiciones.

Resultados y evaluación

Como resultado, GlossExtractor provee una lista de definiciones de una terminología provista por un usuario. Dicha terminología puede

[14] "For example, 'A model is a person who poses for a photographer, painter, or sculptor' is pertinent to the fashion domain but not to the software engineering domain." (Velardi *et al*. 2008: 21).
[15] El algoritmo J48 de WEKA: http://www.cs.waikato.ac.nz/ml/weka

ser una lista de términos o un término específico. Los candidatos a definiciones se presentan para la evaluación del usuario, quien tiene la opción de poder filtrar aquellas definiciones que no sean de su interés. Por último, se da la posibilidad de guardar los resultados en formato TXT o XML.

La evaluación del sistema se llevó a cabo, por un lado, con el fin de valorar la efectividad de los filtros estilísticos y de dominio utilizando un gold-estándar, y por otro lado, para valorar el proceso de elaboración automática de un glosario a partir de la web. En primer lugar se conformó un conjunto de definiciones de glosarios y documentos en línea. Se extrajeron definiciones de glosarios profesionales para obtener ejemplos de definiciones buenas y se conformó un primer grupo *G1*. Enseguida se usó GlossExtractor para buscar definiciones de los términos de *G1* en los documentos en línea, con lo cual se obtuvo el grupo *G2*. Se comparó manualmente el set de definiciones extraídas en *G1* y *G2* para separar buenos y malos candidatos. De esta forma se obtuvo un gold-estándar que se usó en la evaluación de la capacidad del sistema para:

a) Clasificar correctamente buenas y malas definiciones, cuando éstas son extraídas únicamente de glosarios.
b) Clasificar correctamente buenas y malas definiciones, cuando éstas son extraídas únicamente de documentos en línea.
c) Eliminar definiciones que son buenas, pero que no pertenecen al dominio en cuestión.

Tipo documento	P	C	F1
Glosarios Web	0.88	0.91	0.89
Documentos en línea	0.87	0.86	0.86

Tabla 3.26. Resultados de precisión y cobertura en GlossExtractor (tomado de Navigli y Velardi (2007: 46))

Los resultados anteriores son para el corpus de entrenamiento y el corpus de evaluación. En el caso de a) y b) se obtuvieron índices de precisión y cobertura separados para las definiciones extraídas de glosarios Web y para aquellas extraídas de documentos en línea. De manera general, se pueden observar índices altos tanto en precisión

como en cobertura, en ambos casos arriba del 0.85.

Para evaluar el desempeño del sistema en el caso de c), se recopiló un conjunto de 1,000 definiciones en el área de economía, junto con 250 en el área de medicina. Se aplicó el sistema al total de definiciones y se organizó de acuerdo con el valor de pertenencia al dominio de economía. El sistema generaba una lista ordenada de definiciones en la cual la primera definición *no pertinente* (una definición de medicina) aparece en la posición 571 del total, mientras que 75 definiciones de economía aparecen en las posiciones 571 a 1,000.

En síntesis, este sistema conjunta una serie de procesos para extraer información relevante en Internet a partir de una lista de términos especificada por un usuario. Una ventaja es precisamente la disponibilidad en línea, que aunado al proceso de filtrado de definiciones no pertinentes al dominio en cuestión, conforma una buena herramienta para la extracción de conocimiento definitorio independiente del contexto.

3.3.10 Corpógrafo

En última instancia queremos hacer mención de una herramienta cuya finalidad no es directamente la extracción de definiciones, pero ofrece esta posibilidad entre otras opciones. Esta herramienta se denomina *CORPÓGRAFO*[16] y se ha implementado para alemán, español, inglés, italiano, francés y portugués. Es un entorno web que sirve de ayuda en diferentes aspectos del trabajo con corpus, como son su compilación y la extracción terminológica, o el desarrollo de ontologías a partir de éstos. Sarmento *et al*. (2006: 1502) mencionan que esta herramienta permite:

a) Compilar y manejar un corpus (textos que el usuario ingresa al entorno).
b) Crear y manejar bases de datos terminológicas.
c) Extraer semiautomáticamente términos.
d) Extraer semiautomáticamente definiciones y relaciones semánticas.

[16] http://poloclup.linguateca.pt/ferramentas/gc/

e) Exportar los resultados en XML.

Corpus de estudio o ámbito de aplicación

Esta herramienta no parte de un área temática en concreto, sino que cada usuario es quien define el ámbito de trabajo. En Pinto y Oliveira (2004) se reportan algunos experimentos con un subcorpus de medicina en el área de fibromialgia: 10 textos en portugués europeo (21,667 palabras) y 23 textos en inglés (80,295 palabras).

Metodología

El módulo de extracción semiautomática de definiciones funciona a partir del módulo anterior de extracción terminológica. Es decir, una vez que el usuario ingresa los documentos con los que desea trabajar, el módulo de detección de términos obtiene un grupo de candidatos que el usuario valida manualmente. Una vez validado este grupo, cada término se combina con una serie de patrones definitorios típicos y se formulan así expresiones regulares de búsqueda. Algunos ejemplos son los siguientes:

a) "TERM (is | are) * that"
b) "We understand by TERM a * that"
c) "* are (named | known as) TERM"

Las oraciones donde ocurren estas expresiones regulares de búsqueda son presentadas al usuario para su validación final. En total se utilizan alrededor de 135 patrones para portugués, 120 para inglés, y una docena más para lenguas como español, italiano y francés. Se tiene planeado, señalan los autores, dar la posibilidad al usuario de que agregue manualmente otras expresiones regulares de búsqueda propias.

Resultados y evaluación

En Pinto y Oliveira (2004) se explica someramente una evaluación que partió de una lista de términos en inglés y portugués. Para cada término se seleccionaron manualmente las definiciones en el subcorpus de fibromialgia y se compararon con las definiciones extraídas automáticamente.

A continuación presentamos una selección de resultados para algunos términos en inglés[17].

Término	**Total definiciones**	**Total definiciones *correctas***	**Total definiciones *extraídas***
Fibromialgia	31	29 (94%)	33 (88%)
Tender point	6	2 (33%)	2 (100%)
Trigger point	4	4 (100%)	4 (100%)
Substance P	4	1 (25%)	1 (100%)
Secondary fibromialgia	3	2 (67%)	2 (100%)
Sympathetic nervous system	3	0 (0%)	0 (0%)
Fibromialgia pain	2	1 (50%)	1 (100%)
Primary fibromyalgia	2	1 (50%)	1 (100%)

Tabla 3.27. Resultados porcentuales de definiciones extraídas a partir de CORPÓGRAFO (tomado de Pinto y Oliveira (2004))

En la tabla anterior se presentan algunos términos y sus resultados porcentuales, en los cuales la primera columna indica el total de definiciones presentes para cada término, la segunda contiene el número de definiciones correctas extraídas automáticamente, y la tercer columna presenta el número total de definiciones propuestas por CORPÓGRAFO. Observamos por ejemplo que para el término *fibromialgia* existen 31 definiciones en el corpus de evaluación. De este número el sistema propone 33, de las cuales 29 casos eran definiciones correctas.

Con estos resultados se pretende mostrar un panorama básico del desempeño del sistema para la extracción de definiciones. Señalan los autores que la lista de patrones compilada es una lista muy básica que no dará cuenta de la variedad de fórmulas que pueden emplearse para introducir definiciones en los textos. A pesar de ello, señalan que con esta lista básica se puede extraer un primer grupo de candidatos a definiciones sobre el cual el usuario realiza una validación manual para comprobar la utilidad de cada resultado.

[17] Cabe señalar que en el trabajo de Pinto y Oliveira (2004) no se detalla una evaluación global de la metodología.

En resumen, CORPÓGRAFO constituye una herramienta completa para la compilación y procesamiento de corpus. Permite desde la generación de un corpus a partir de documentos provistos por un usuario, hasta la explotación de éste pasando por la extracción terminológica, la extracción de definiciones y la elaboración de ontologías. Resulta interesante la idea de permitir al usuario formular sus propias expresiones de búsqueda basadas en otros patrones definitorios. Además, cabe señalar que es uno de los pocos estudios que reportan metodologías aplicadas a la lengua española.

3.4 Análisis contrastivo

En los siguientes apartados sintetizamos algunas ideas básicas de los estudios que hemos abordado. En primer lugar se expondrán algunos conceptos generales sobre las metodologías de extracción automática (3.4.1). En segundo lugar se detallarán algunas características comunes en las metodologías de evaluación (3.4.2). En tercer lugar se realizará un análisis comparativo de los resultados obtenidos en cada caso (3.4.3). Y por último, en cuarto lugar se presentarán a modo de conclusión algunas consideraciones importantes en el tema (3.4.4).

3.4.1 Metodologías de extracción

Como hemos visto a lo largo de este capítulo, el concepto más evidente en la extracción automática de CDs es partir de la búsqueda de patrones definitorios, una idea que surgió del análisis de las ocurrencias de definiciones en textos especializados. A partir de dichos estudios, se observó que en los contextos donde se aporta información definitoria sobre un término se suelen emplear una serie de claves sintácticas y tipográficas que podrían servir como punto de inicio en la extracción automática de CDs. Estas claves se pueden englobar, de manera general, en patrones tipográficos y patrones sintácticos.

Los patrones tipográficos hacen referencia a signos de puntuación, que suelen funcionar como conectores entre el término y su definición, y a la tipografía del texto, la cual se utiliza para resaltar visualmente alguna información importante. Por su parte, los patrones sintácticos pueden estar constituidos por verbos que

introducen una predicación definitoria acerca de un término particular, pueden estar formados por marcadores discursivos, o bien estar conformados por frases más complejas dónde se combinan algunos marcadores y verbos. A continuación presentamos un resumen de los diferentes tipos de patrones utilizados en los estudios reseñados en este capítulo:

	Referencia[18]	Idioma[19]	Tipo Patrón	Ejemplos
1	Rebeyrolle y Tanguy	fr	verbales	défin$ *définir * comme*
2	DEFINDER	in	signos verbales frases	*paréntesis*, *guiones* *is called* *is the term for*
3	Malaisé	fr	signos verbales marcadores	*paréntesis* *dénnomer* *nom , c'est-à-dire*
4	Sánchez y Márquez	es	verbales	*entenderse*
5	Storrer y Wellinghoff	al	verbales	*definieren als*
6	MOP	in	signos verbales marcadores	*comillas* *called, known as* *word, term*
7	Saggion	in	verbales	*TERM is a, such as TERM*
8	LT4eL	al, bu, ch, ho, in, ma, pol, por, ru	signos tipografía verbales	*paréntesis* *viñetas* *IS-A, defined as, Called*
9	GlossExtractor	in	verbales	*TERM refers to* *TERM is a kind of*
10	Corpógrafo	al, es, in, it, fr, por	verbales	*TERM is a * that* ** are known as TERM*

Tabla 3.28. Ejemplos de patrones definitorios empleados en la extracción de CDs

De la tabla anterior resaltamos lo siguiente:

[18] En este caso la referencia la hacemos a los nombres de los autores si el desarrollo no tiene un nombre específico.

[19] Donde: **al**: alemán, **bu**: búlgaro, **ch**: checo, **es**: español, **fr**: francés, **ho**: holandés, **in**: inglés, **it**: italiano, **ma**: maltés, **pol**: polaco, **por**: portugués y **ru**: rumano.

1. La mayoría de los estudios parte de la inclusión de patrones específicamente verbales. Es común el uso de verbos aislados, en formas personales como *define* o *denomina* y el uso de formas compuestas como *conocido como*, *es llamado*, *designar como*, etc.
2. También se utilizan en menor medida otros patrones sintácticos *no verbales*, como el caso de DEFINDER donde se menciona el uso del patrón *is the term for*, o en MOP donde se utilizan palabras simples como *word* o *term*.
3. En el caso de los patrones tipográficos, el uso más extendido es el de los signos de puntuación, de los cuales se emplean en mayor medida los *paréntesis*.
4. Por su parte, el uso de la tipografía textual es referido sobre todo como un peso extra que se les otorga a los candidatos que están resaltados por ejemplo en itálicas, aunque en los estudios del proyecto LT4eL se hace referencia al uso de la tipografía como un patrón independiente (por ejemplo cuando los CDs aparecen en viñetas).
5. En último lugar, cabe mencionar que el desarrollo de las metodologías para la extracción de CDs ha recibido mayor atención en francés e inglés. Es reciente la implementación en otras lenguas como alemán, portugués, lenguas eslavas o el mismo español, del cual sólo tenemos constancia de los estudios de Sánchez y Márquez y CORPÓGRAFO.

Ahora bien, la extracción automática de CDs no termina en la identificación de contextos con algún patrón definitorio. Como hemos visto, los patrones que suelen utilizarse pueden recuperar *ruido* en mayor o menor medida: algunos patrones utilizan verbos como *definir* o *entender*, cuyo carácter metalingüístico los hace más probables a ocurrir en construcciones donde se aporta información definitoria; sin embargo, patrones como *IS-A* son bastante frecuentes en todo tipo de oraciones y tienden a producir una mayor cantidad de contextos no relevantes.

Es por ello que en algunas metodologías se hace mención a la necesidad de especificar los patrones definitorios, tratando de eliminar, por ejemplo, algunos tiempos verbales en los patrones que suelen recuperar más ruido. También se menciona la necesidad de desarrollar filtros que permitan eliminar aquellos contextos que se

recuperaron por tener algún patrón definitorio, pero que no corresponden a un CD. Algunas consideraciones son las siguientes:

1. Rebeyrolle y Tanguy especifican el uso del verbo *definir* a patrones donde se incluya una ventana más el adverbio *como*: *definir * como*. En sus experimentos formulan diferentes opciones: la ventana puede ser cualquier palabra, puede estar formada por un total de seis palabras o puede contener cualquier palabra excepto un verbo. Los resultados más precisos se obtuvieron con esta última opción.
2. Storrer y Wellinghoff reportan igualmente que los resultados con mejor precisión fueron obtenidos con patrones más específicos que incluían la preposición *por* (*se entiende por*) o el adverbio *como* (*se define como*).
3. Para eliminar contextos donde aparece un patrón definitorio pero que no aportan información relevante, Rodríguez conformó un corpus con CDs y no CDs que analizó con algoritmos de aprendizaje automático. La idea era encontrar patrones recurrentes en los contextos no relevantes y así tener una primera base para implementar su filtrado automático.
4. Por otra parte, aunque en esta misma línea de aprendizaje automático, en algunas lenguas del proyecto LT4eL se implementaron algoritmos genéticos para asignar un mayor peso a aquellos patrones definitorios que tenían mejores resultados. Más que diseñar un filtro de *malos candidatos*, se diseñó un proceso para reconocer a los *mejores*.
5. GlossExtractor implementa igualmente un proceso para reconocer los mejores candidatos mediante expresiones regulares que fueron identificadas con un algoritmo de aprendizaje automático. En este caso lo que se aprendió fue el *estilo* recurrente en las definiciones analíticas, por ser consideradas las que proporcionan información más completa.
6. De la misma forma, Saggion expone una metodología para reconocer los mejores candidatos buscando automáticamente los términos secundarios en las definiciones del término de búsqueda, y *filtrando* los mejores candidatos donde dichos términos aparecen en mayor número y frecuencia.

La idea de filtrar malos candidatos responde a la necesidad de extraer menos ruido y por ende tener una mayor certeza de que lo extraído automáticamente corresponde efectivamente a contextos

definitorios. No obstante, como hemos señalado en los puntos 4, 5 y 6, también se han desarrollado métodos a la inversa, es decir, tratando de filtrar o reconocer los mejores candidatos. Cabe señalar el trabajo de Xu *et al*. (2006) donde se presenta un método de *ranking* o clasificación jerárquica de candidatos a definiciones.

Una vez filtrados los malos candidatos, algunos estudios coinciden en la necesidad de identificar automáticamente el término y la definición en los CDs. En este punto es necesario diferenciar dos tipos de aproximaciones que tienen que ver con el enfoque de trabajo. El primero es un enfoque que parte de la búsqueda directa de patrones definitorios para identificar CDs. El segundo parte de un término o lista de términos con la finalidad de buscar información pertinente sobre ellos. Este último enfoque, por ejemplo, es el reportado en GlossExtractor y en Corpógrafo.

Por último, algunas consideraciones sobre la identificación de los elementos constitutivos de CDs son las siguientes:

1. Malaisé señala la idea de utilizar patrones contextuales para identificar al término y la definición dependiendo del verbo que se utilice. En un patrón como *IS-A*, el término sólo podrá aparecer a la izquierda del verbo: *X es un*; mientras que verbos como *definir* permiten una mayor posibilidad de estructuras: *X se define como*, *se define a X como*, *se define como X a…*, etc.
2. Storrer y Wellinghoff mencionan la posibilidad de recurrir a las valencias de cada verbo definitorio para determinar el lugar que podría ocupar el término con respecto al verbo que lo conecta con su definición. Por ejemplo, X sería un sujeto en: *X se define como*; pero sería un objeto en: *se define a X como.* En estos casos se debería recurrir a un analizador sintáctico para reconocer los elementos oracionales.
3. El hecho de recurrir a las valencias de cada verbo para identificar el término se sintetiza en los trabajos de LT4eL, a partir del uso de gramáticas que representan las posibilidades estructurales de cada patrón.

3.4.2 Metodologías de evaluación

El análisis contrastivo de metodologías de evaluación en el ámbito de la extracción automática de CDs dista mucho de ser una tarea trivial, principalmente porque existen algunos problemas que hacen de este proceso evaluativo uno de los menos estandarizados en la bibliografía al respecto. El primer problema, y a nuestro parecer el más grande, es decidir *¿qué es?* y *¿qué no es?* una definición. Unido a esto se presenta el problema de decidir *¿qué es una buena definición?*, lo cual nos lleva al siguiente problema de qué criterios se deben tener en cuenta para compilar un *gold-estándar* con el fin de tener un repositorio de *buenas definiciones* que sirva como punto de comparación para la evaluación de la extracción automática.

Como veremos en seguida, estos problemas se han tratado de solventar de distintas maneras en las metodologías que hemos detallado:

1. Rebeyrolle y Tanguy recurren a su competencia lingüística para identificar manualmente definiciones sobre un corpus en las áreas de geomorfología, ingeniería del conocimiento y documentos empresariales. Las definiciones identificadas manualmente conformaron su gold-estándar sobre el cual se compararon los resultados obtenidos automáticamente.
2. Muresan y Klavans idearon una metodología para evaluar el sistema DEFINDER de forma cuantitativa y cualitativa. En cuanto a la evaluación cuantitativa, la primera comparación de los resultados automáticos la realizaron contra un gold-estándar de definiciones compilado manualmente por 4 sujetos, y en el cual se habían seleccionado aquellas definiciones que hubieran sido aprobadas por lo menos por 3 de ellos. La segunda comparación fue realizada contra diccionarios y glosarios para ver si las definiciones automáticas se encontraban en estos recursos. Por su parte, la evaluación cualitativa consistió en pedir a especialistas y no especialistas su opinión sobre la usabilidad y claridad de las definiciones extraídas automáticamente.
3. Malaisé evaluó la extracción de CDs recurriendo a su competencia lingüística para identificar manualmente las definiciones y compararlas con los resultados automáticos.

4. En el caso de Saggion, la evaluación se realizó contra un gold-estándar compilado manualmente por los organizadores de la competencia.
5. Velardi *et al*. utilizaron como gold-estándar un conjunto de definiciones extraídas de diccionarios y glosarios, así como clasificadas manualmente. La finalidad era analizar el desempeño del sistema para extraer definiciones a partir de patrones y el desempeño de los filtros estilísticos para identificar los mejores candidatos a definiciones. También compilaron manualmente un conjunto de definiciones en dos áreas distintas, las mezclaron y a partir de este conjunto evaluaron la capacidad del sistema en cuanto al filtro de dominio se refiere.
6. En los demás estudios, la base para la evaluación de los candidatos extraídos automáticamente fue la propia identificación manual de definiciones.

De lo anterior podemos deducir algunas conclusiones en cuanto a las metodologías de evaluación. En primer lugar, la extracción de CDs se ha considerado típicamente un problema binario donde se tiene que clasificar, de un grupo de candidatos extraídos a partir de un conjunto de patrones, aquellos que corresponden o no a lo esperado, en este caso a definiciones o a contextos definitorios. En segundo lugar, para analizar el desempeño del sistema en esta clasificación binaria es indispensable tener entonces un punto de comparación, es decir, un grupo de *buenas definiciones*. Ahora bien, la conformación de este grupo de buenas definiciones se ha llevado a cabo tanto por *especialistas* como por *no especialistas* en el dominio de estudio, lo cual presenta un problema básico: analizar manualmente un texto en busca de definiciones supone la influencia de los antecedentes y del conocimiento del individuo que realiza esta tarea. El *especialista* en el área de conocimiento no suele estar entrenado para decidir si un objeto es o no una definición, lo que aporta es su punto de vista para decidir si el objeto ayuda o no a esclarecer el significado de un término. Por otro lado, el *no especialista* suele ser el que desarrolla la herramienta o un experto en el área de terminografía y no una persona involucrada en el dominio de estudio, por lo que suele aportar su punto de vista para decidir si lo evaluado tiene o no la forma de una definición.

En el trabajo de Klavans y Muresan vemos como se trata de solventar este problema con la inclusión de una evaluación cuantitativa y cualitativa. Se pone en evidencia la necesidad de evaluar no sólo la cantidad de definiciones extraídas automáticamente, sino también la calidad de las mismas. De esta forma, se recurre entonces a un gold-estándar de definiciones adquirido a partir del acuerdo entre 4 individuos, a la revisión manual de los resultados automáticos por parte de especialistas y no especialistas, así como a la comparación de las definiciones extraídas automáticamente contra recursos ya estructurados, como diccionarios o glosarios. El caso de GlossExtractor es parecido, ya que se evalúa no sólo la extracción de definiciones a partir de patrones, sino la calidad de las mismas comparándolas con definiciones de diccionarios.

Cabe señalar que, en el caso de conformar un gold-estándar de definiciones por parte de especialistas o no especialistas, surge además el problema de cómo establecer un acuerdo común para lo que será considerado una definición. Klavans y Muresan lo resuelven recurriendo a la decisión de 3 de 4 individuos, una especie de *revisión por pares*. Por su parte, Przepiórkowski *et al*. mencionan un *Interannotator Agreement*, que es un índice para representar el acuerdo en la decisión de clasificar manualmente una definición y el cual se perfila como una opción en este difícil proceso de clasificación.

A pesar de las diferencias en las distintas metodologías de evaluación, existe un común acuerdo de utilizar los índices de precisión y cobertura para representar cuánto ruido se extrajo y cuántos objetos de estudio fueron correctamente extraídos. Se ha utilizado además el índice *f-score* en distintas modalidades: el índice normal *F1*, para dar una idea más general del balance entre los resultados de precisión y cobertura, y el índice *F2*, donde se da más peso a la cobertura, principalmente cuando es más importante saber cuántos objetos de estudio se extrajeron automáticamente.

3.4.3 Resultados de evaluación

En la tabla 3.29 se sintetizan los resultados de los índices de precisión (P), cobertura (C) y f-score (F1 y F2).

Rf.	Id.	Área	Tamaño	P	C	F1 / F2
1	fr	Geomorfología	275,000 p	-	-	-
		Ing. conocimiento	230,000 p			
		Empresariales	205,000 p			
		Enciclopedia	215,000 p			
2	in	Medicina	-	86.95%	75.47%	-
3	fr	Dietética y nutrición	480,000 p	55%	39.3%	-
4	es	Derecho	-	97.44%	100%	-
5	al	Tecnologías texto	103,805 p	34%	70%	-
6	in	Sociología	5,581 o	0.97	0.79	F1 0.87
		Histología	5,146 o	0.94	0.81	F1 0.87
7	in	Noticias	1,000,000 t	-	-	F1 0.23
8	al	Derecho	237,935 o	48.6%	-	-
	bu	Cómputo	76,800 p	22.5%	8.9%	F2 11.1%
	ch	" "	90,000 p	22.3%	46%	F2 33.9%
	ho	" "	14 t	-	-	-
	pol	" "	83,200 p	23.3%	32%	F2 28.4%
	por	" "	274,000 p	0.14	0.86	F2 0.66%
	ru	" "	700,000 p	-	-	-
9	in	Economía	1,000 o	0.87	0.86	F1 0.86
		Medicina	250 o			
10	in	Medicina	80,295 p	-	-	-
	por	" "	21,667 p			

Tabla 3.29. Resultados de evaluación en las metodologías y/o sistemas para la extracción de contextos definitorios

Para elaborar esta tabla, se tomó como principal criterio el hecho de que lo evaluado fuera el proceso más general en las metodologías descritas y correspondiera a la extracción de CDs mediante patrones definitorios. Sobre el tamaño del corpus de evaluación, *p* equivale a palabra, *o* equivale a oración y *t* a texto. De estos datos podemos concluir lo siguiente:

1. Las áreas de estudio incluyen diferentes dominios de conocimiento especializado, siendo el que más se acerca a la lengua general el corpus de Noticias en el estudio de Saggion
2. La mayoría de corpus en el proyecto LT4eL pertenecen al dominio general de cómputo, con subdisciplinas en temas

como *sociedades de la información*, *tecnologías de la información* o *eLearning*.

3. El tamaño del corpus de evaluación es completamente variable, va de casos con poco más de 20,000 palabras hasta un conjunto de 1 millón de textos. En DEFINDER y Sánchez y Márquez no se reporta el tamaño del corpus.
4. Por otro lado, algunos estudios, como el de Rebeyrolle y Tanguy y CORPÓGRAFO, no presentan una evaluación global de la metodología, sino índices aislados para cada tipo de patrón definitorio. Además, en el segundo caso sólo se presentan resultados para el inglés.
5. Medir el índice de cobertura supone conocer previamente el número total de definiciones en el corpus de evaluación, es decir, un laborioso proceso manual. En el estudio en alemán de LT4eL se empleó un corpus de 237,935 oraciones, razón por la cual se excluyó el índice de cobertura.
6. En cuanto a los datos numéricos, se observa un equilibrio de índices altos de precisión y cobertura sobre todo en DEFINDER, Sánchez y Márquez, MOP y GlossExtractor. En otros casos alguno de los dos índices recuperaba mejores resultados, como la precisión en Malaisé o la cobertura en Storrer y Wellinghoff y en el estudio en portugués de LT4eL.
7. No obstante, es bastante difícil establecer un análisis contrastivo entre los resultados. Si bien algunos estudios presentan índices altos tanto para precisión como cobertura, se debe tomar en cuenta las condiciones generales en las cuales se encontraron estos índices. Tal es el caso del trabajo de Sánchez y Márquez que reportan una precisión de 97.44% y una cobertura de 100%, pero a partir de la extracción automática de 38 ocurrencias para un solo verbo definitorio.
8. Saggion reporta un f-score de 0.23, aunque en el contexto de sus resultados el sistema que obtuvo el índice más alto era de 0.55.
9. Rodríguez (2004: 21) señala que sus índices obtenidos podrían ser comparados con los de DEFINDER, sin embargo aquí debería tomarse en cuenta que el primero realizó una identificación manual de las definiciones a comparar en la evaluación, mientras que los segundos recurrieron al acuerdo de cuatro individuos en dicha selección manual.
10. El caso de los resultados en el proyecto LT4eL podría ser el que mejor se preste a comparación, tomando en cuenta la

similitud de las metodologías empleadas. Sin embargo, aquí debería tomarse en cuenta el caso de las lenguas eslavas, por ejemplo el búlgaro, donde las definiciones suelen expresarse en más de una oración y la metodología general en este proyecto contemplaba una oración como la extensión general de un CD.

11. Debido a lo anterior, los índices de precisión y cobertura más bajos se observan para las lenguas eslavas.
12. Algunos datos importantes de los estudios que no presentaron una evaluación global: en rumano, el índice de precisión para los patrones *verb_def* fue de 0.75, frente al 0.04 para el caso del patrón *layout_def*, lo cual indica la gran cantidad de ruido que supone buscar la tipografía como un patrón independiente; en holandés el patrón *is_def* obtuvo un 91.8% de cobertura con un 20.97% de precisión, donde se muestra asimismo la cantidad de ruido que suele recuperarse con esta forma; otra cobertura baja de precisión la obtuvo el patrón *défin$*, en el trabajo en francés presentado por Rebeyrolle y Tanguy, el cual incluía únicamente el lema, mientras que si se especificaba el patrón a una secuencia de *définir*, más cualquier palabra excepto un verbo, más el adverbio como, entonces la precisión aumentaba de 5.15% a 91.67%.

3.4.4 Conclusiones

De lo expuesto hasta ahora podemos concluir que la similitud más evidente en las metodologías radica en el enfoque basado en patrones como puntos clave en el reconocimiento automático de fragmentos con información definitoria. Existe además un acuerdo general de que los patrones sintácticos son los más productivos para la extracción de CDs, y de éstos, los patrones verbales presentan una mayor preferencia frente a construcciones sintácticas que incluyen palabras metalingüísticas pero no verbos. La frecuencia de verbos definitorios para la constitución de patrones es mayor que el uso de otras palabras o frases metalingüísticas, ya que patrones como *la definición de* o *el concepto de* suelen emplearse para introducir CDs en una menor medida que patrones verbales como *se llama* o *se describe como*.

En las metodologías de extracción automática se hace evidente además la necesidad no sólo de formular patrones concisos y poco ambiguos, sino también el hecho de que invariablemente los patrones recuperarán contextos no relevantes, por lo cual es necesario desarrollar filtros para tratar de excluir esos enunciados no definitorios. Asimismo, si tomamos en cuenta los pasos en la metodología de extracción de contextos definitorios propuestos por Meyer[20], añadir nuevos patrones definitorios supone un proceso inherente a la metodología de extracción, sin embargo, mientras no se añadan restricciones a estos patrones también supondrá una mayor recuperación de ruido.

También es clara la necesidad de reconocer los elementos constitutivos de los CDs cuando el enfoque parte de la búsqueda de patrones sin conocer previamente el término sobre el cual se busca una definición. En este punto entramos en la necesidad de delimitar, para cada patrón, el tipo de información sintáctica y semántica que puede introducir. No obstante, se corre el riesgo de desarrollar reglas específicas para el área de estudio en cuestión, reglas cuya adaptación a otros dominios constituirá una labor intensiva.

Sobre la aplicación de aprendizaje automático, su uso más evidente es en el reconocimiento de patrones para filtrar candidatos no relevantes. Este punto resulta de gran interés, ya que entramos en el terreno de la riqueza del lenguaje y en la gran variedad de escenarios donde se pueden utilizar los verbos más comunes en los patrones definitorios. El problema para aprender reglas automáticamente es que éstas suelen basarse en el nivel sintáctico, y a este nivel en algunos casos es casi imposible reconocer si un candidato es bueno o no, ya que sintácticamente se puede dar el caso de un contexto no definitorio que siga todas las pautas o reglas que se presentan en los elementos constitutivos de un CD (el término es una frase nominal, la definición suele comenzar con un determinante, etc.), pero que semánticamente no aporte información definitoria sobre un término.

Por último, en cuanto a los resultados esperados, el punto clave en la extracción de CDs, al igual que en otros sistemas de extracción de información, es lograr un equilibrio entre el total de objetos de

[20] Ver la tabla 3.5 de este capítulo.

estudio recuperados automáticamente del corpus, frente al total de ruido también recuperado de manera automática. En algunos escenarios puede ser más importante que la mayoría de objetos extraídos automáticamente sean objetos válidos, es decir, que el ruido sea menor. En otros casos será más relevante que todos, o una gran parte de los objetos de estudio, hayan sido recuperados automáticamente. Como señala Meyer, será más fácil limpiar manualmente el ruido en una lista de candidatos, que ir al corpus para comprobar manualmente que todos los objetos de estudio hayan sido correctamente extraídos. Pero esta importancia recae finalmente en el escenario donde se proyecte la aplicación.

4. Contextos definitorios en textos de especialidad

A lo largo del capítulo anterior revisamos el estado del arte en la extracción automática de información definitoria y realizamos un análisis comparativo entre las diferentes metodologías tanto de extracción como de evaluación. Una idea general en los estudios revisados es que la extracción de información definitoria y conceptual sobre términos puede realizarse a partir de la búsqueda de CDs.

Tomando en cuenta lo anterior, en este capítulo realizamos un análisis lingüístico de CDs en textos especializados en español, con el fin de sentar las bases metodológicas para su extracción automática.

Comenzaremos con una revisión de los conceptos de *contexto* y *contexto definitorio* desde el punto de vista del trabajo terminográfico y con miras a su extracción automática (4.1). Después describiremos un análisis lingüístico de CDs en español, donde especificaremos una tipología de patrones definitorios (4.2). Por último, presentaremos una tipología semántica de CDs que nos ayudará en la clasificación del tipo de información definitoria extraída (4.3).

4.1 Análisis del concepto de contexto definitorio en el ámbito terminográfico

Como señalamos en el apartado 2.5, un CD, a grandes rasgos, es un fragmento textual donde se aporta información sobre los atributos y/o relaciones semánticas de un término y que puede ser útil para comprender su significado. En el marco de esta investigación, creemos conveniente delimitar el concepto de CD a partir de una revisión de la literatura al respecto en el ámbito de la terminografía y en el ámbito de estudios relacionados con su extracción automática.

De esta forma, a lo largo de este apartado veremos distintas descripciones referentes a la realización, forma y alcance de los

CDs. Lo anterior tiene la finalidad de delimitar una definición aplicada del concepto de contexto definitorio que se adapte a nuestra necesidad de extracción automática.

4.1.1 Contexto y contexto definitorio

Para De Bessé (1991: 112, 115), el *contexto* se entiende como el entorno lingüístico de un término constituido por un enunciado, es decir, las palabras o frases alrededor de dicho término, y que condiciona su forma, funcionamiento, sentido, valor y uso. El contexto, añade De Bessé, conforma el punto de inicio de cualquier trabajo terminográfico y tiene dos funciones: aclarar el significado de un término e ilustrar su funcionamiento.

En este mismo sentido, en las normas ISO un contexto se define como un texto o una parte del texto donde ocurre un término y que ejemplifica un concepto o el uso de una designación (ISO 12620 1999: 25).

Específicamente, los contextos son indispensables en la adquisición de datos que permitan al terminógrafo redactar la entrada a partir de la cual se hará la definición, y constituyen por ende un elemento escencial para la descripción de un concepto (De Bessé 1991: 115-116).

Ahora bien, en las normas ISO podemos encontrar una clasificación general de los diferentes tipos de contextos, los cuales pueden dividirse en:

- Contextos lingüísticos
- Contextos metalingüísticos
- Contextos explicativos
- Contextos asociativos
- Contextos definitorios

La explicación de cada uno de estos tipos de contextos y un ejemplo se puede observar en la tabla 4.1.

Tipología de Contextos
Contextos lingüísticos. Ilustran la función del término en el discurso del dominio especializado pero no aportan información conceptual: ▪ *Cylindrical grinders* consume relatively little power.
Contextos metalingüísticos. Consisten en información sobre los términos como símbolos en la forma en que éstos son usados de forma autónoma: ▪ The term *expertise* in French when it is used to mean *compétence d'expert* (*expert competence*) is a borrowing from English.
Contextos explicativos. Aportan explicaciones generales o sintetizadas sobre los términos: ▪ The "reed", which keeps the warp yarns separated, helps to determine cloth width.
Contextos asociativos. Contienen la cantidad mínima de información conceptual que se necesita para asociar un concepto al dominio de otro concepto específico: ▪ Machine tool operations such as blanking, piercing, lancing, shearing, *beading*, and flanging can also be performed in a press brake.
Contextos definitorios. Contienen información sustancial acerca de un concepto pero no contienen el rigor formal de una definición: ▪ Weaving is a method of producing cloth by interlacing two or more sets or yarns, at least one warp and one filling set, at right angles to each other.

Tabla 4.1. Tipología de contextos según la norma ISO 12620 (tomado de ISO 12620 (1999: 25-26))

Observamos que los contextos lingüísticos son todos aquellos contextos que no pueden ser categorizados dentro de las demás clases porque, a diferencia de éstas, no aportan información conceptual sobre el término. Estos contextos son la base para el análisis lingüístico descrito en los procedimientos básicos para la elaboración de diccionarios especializados o bases de datos terminográficas, a partir de los cuales se pueden estudiar los contextos reales de uso de los términos. En el ejemplo provisto por las normas ISO vemos un contexto que sólo hace mención del uso del término *cylindrical grinders* sin aportar información conceptual.

Por su parte, los contextos metalingüísticos hacen mención al tipo de contextos donde se habla del término en sí mismo como una unidad del lenguaje. Difieren de los demás en el sentido de que constituyen un tipo de discurso acerca del mismo término, mientras que los otros tipos de contextos conforman un discurso sobre el objeto o la noción a la que los términos hacen referencia.

Los contextos explicativos contienen información sintetizada o muy general sobre el término en cuestión. En el contexto presentado como ejemplo se provee información sobre los atributos del término *reed*, los cuales pueden ser vistos como una función: *which keeps the warp yarns separated*, y como una utilidad: *helps to determine cloth width*.

Los contextos asociativos incluyen información relevante que sirve para conocer el dominio del concepto al cual pertenece un término. En el ejemplo provisto, los términos *blanking*, *piercing*, *shearing* y *beading* son acciones o situaciones que pertenecen al dominio conceptual del término *machine tool operations*.

Por último, los contextos definitorios son contextos que incluyen información básica relevante sobre el término, sin tener la forma estricta de una definición; cuando están compuestos de información explicativa acerca del objeto del concepto, se habla entonces de contextos enciclopédicos.

De Bessé (1991), por su parte, propone una tipología de contextos más específica basada en dos clases generales: 1) *contextos ligados a un término* y 2) *contextos que refieren a un concepto*.

Los contextos ligados a un término sólo proveen una cantidad limitada de información sobre el concepto relacionada con el uso de dicho término, e incluyen:

- Contextos del lenguaje
- Contextos fraseológicos
- Contextos lingüísticos
- Contextos de uso

Los contextos que refieren a un concepto incluyen información de dicho concepto pero aportan poca información sobre el uso del término. En este último grupo se incluyen:

- Contextos definitorios
- Contextos conceptuales
- Contextos enciclopédicos
- Contextos materiales

En la tabla 4.2 presentamos una breve descripción de cada uno de los contextos antes señalados.

Contextos ligados a un término
Contextos del lenguaje. Ocurrencias del término que no contienen indicaciones de su funcionamiento ni elementos de la definición del concepto.
Contextos fraseológicos. Contextos paralelos en dos o más lenguas cuya función es presentar una equivalencia semántica con fines comparativos.
Contextos lingüísticos. Ilustran el uso normal de un término y su comportamiento dentro del lenguaje.
Contextos de uso.- Proporcionan indicaciones sobre el funcionamiento sociolingüístico del término, fundamentalmente indicaciones metalingüísticas.
Contextos que refieren a un concepto
Contextos definitorios. Contienen un cierto número de elementos útiles y necesarios para la descripción de un concepto, pero insuficientes para la redacción de una definición.
Contextos conceptuales. Asocian un término con otros términos que pertenecen al mismo dominio de conocimiento.
Contextos enciclopédicos. Completan una definición mediante información suplementaria sobre el concepto.
Contextos materiales. Proporcionan indicaciones que permiten vincular al término con un dominio de conocimiento particular.

Tabla 4.2. Tipología de contextos según De Bessé (tomado de De Bessé (1991: 112-115))

A diferencia de la tipología propuesta en las normas ISO, en esta clasificación se especifican con más detalle algunas diferencias

entre los tipos de contextos ligados a un término. Los contextos lingüísticos coinciden con los de las normas ISO en ser ocurrencias que ilustran el uso y comportamiento normal de los términos, a través de los cuales pueden analizarse sus construcciones sintácticas y coocurrencias más frecuentes.

Para De Bessé, cualquier ocurrencia simple de los términos que ofrezca información nula o insignificante sobre su funcionamiento sintáctico se considera como un contexto del lenguaje. Los contextos fraseológicos, por su parte, no contienen información sobre el concepto ni ilustran el funcionamiento real de la lengua, ya que tienden a ser creaciones artificiales que no corresponden a una producción lingüística espontánea, y suelen ser útiles en la comparación de diferentes versiones lingüísticas de un texto, por ejemplo en el estudio lingüístico de textos legislativos de países multilingües.

Los contextos de uso podrían equipararse a los contextos metalingüísticos de las normas ISO, ya que en ambas tipologías este tipo de contextos hacen referencia a la información que se provee sobre el uso del término como una unidad del lenguaje. En este sentido, los contextos materiales podrían representar una variación sutil de la información provista en los contextos metalingüísticos, si consideramos a las indicaciones que permiten vincular al término con un dominio específico como información metalingüística sobre el ámbito en el que operan los términos.

Los contextos conceptuales, a su vez, podrían compararse con los contextos asociativos, en tanto ambos incluyen información que permite relacionar un término con otras unidades terminológicas del mismo campo de estudio.

Finalmente, De Bessé expresa la idea de que los contextos definitorios contienen información esencial que permite describir un término pero que es insuficiente para la redacción de una definición, siendo el cometido de los contextos enciclopédicos añadir información suplementaria cuando la presencia de la definición no es suficiente.

4.1.2 Enunciados de interés definitorio

Auger (1997) presenta otra categorización de los distintos tipos de contextos que se pueden encontrar en los textos especializados, específicamente sobre aquellos que contienen información definitoria y/o conceptual. Auger define como *enunciados de interés definitorio* a los contextos que se caracterizan por el hecho de incluir, implícita o explícitamente, elementos de información semántico-referencial que permiten el análisis semántico de las unidades del léxico.

Un enunciado es de interés definitorio cuando ilustra diversos aspectos de una palabra, como su sentido, nivel de lengua, contexto de uso, grafías, etc., o cuando describe los elementos, atributos o las propiedades de un objeto del mundo (Auger 1997: 27). Los enunciados de interés definitorio se pueden clasificar en *enunciados definitorios* y *enunciados sémicos*.

Enunciados de interés definitorio
Enunciados definitorios. Todo discurso que determina de forma *explícita*, total o parcialmente, las características que pertenecen a un concepto, los elementos que lo caracterizan, o bien un aspecto propio de un signo lingüístico, mediante el uso de un vocabulario particular (lingüístico o metalingüístico) que explique dicha determinación.
Enunciados sémicos. Todo discurso que determina de forma *implícita*, total o parcialmente, las características que pertenecen a un concepto, mediante el uso de formas léxicas diferentes a un vocabulario definitorio.

Tabla 4.3. Tipología de enunciados de interés definitorio según Auger (tomado de Auger (1997: 64))

Observamos que en los enunciados definitorios, a diferencia de los enunciados sémicos, la operación de proveer información sobre las características de un concepto se lleva a cabo de manera explícita a través del uso de un vocabulario o de formas léxico-sintácticas específicas, como puede ser el caso de ciertos verbos copulativos (*ser*) o metalingüísticos (*definir, denominar*, etc.). Auger utiliza el siguiente contexto como un ejemplo donde se define el término *oftalmólogos* a partir de un verbo copulativo:

- Les ophtalmologistes sont des médecins spécialistes des yeux.

Por el contrario, desde un punto de vista de la forma, los enunciados sémicos se presentan como un simple discurso narrativo que contiene información semántica sin recurrir al vocabulario o formas léxico-sintácticas que se emplean en los enunciados definitorios. En el siguiente ejemplo, Auger menciona que se provee información sobre el contenido nocional del término *iglú*, sin utilizar algún verbo copulativo o metalingüístico:

- L'eau ne coulait déjà plus sur les murs, à l'intérieur de l'igloo. Et une fois qu'Agaguk mit le bras à travers le trou d'aération, au sommet de l'igloo, il sentit que les blocs étaient fondus de moitié. Il faisat maintenant trop chaud dans l'habitation.

En el caso particular de los enunciados definitorios, éstos siguen un patrón del tipo: [N_0] → [verbo o forma sintáctico-lingüística] → [N_1 – X], en donde [N_0] representa el término que se define, [N_1 – X] es la secuencia definitoria que refiere a [N_0], y el verbo o forma sintáctico-lingüística es una predicación definitoria que une el término con la secuencia de la definición. La tipología de enunciados definitorios la encontramos en la siguiente tabla:

Enunciados definitorios metalingüísticos (EDM)	
EDM de designación	*designar*, *significar*, *querer decir*
EDM de denominación	*denominar*, *llamarse*
EDM sistemáticos	*escribirse*, *pronunciarse*, *el sustantivo*, *el verbo*, *el adjetivo*, *el sintagma*,
Enunciados definitorios lingüísticos (EDL)	
EDL copulativos	*Un X es un Y que…*
EDL de equivalencia	*equivaler a*, *también llamado*, *es decir*
EDL de caracterización	*constituir, características*, *atributos*
EDL de análisis	*compuesto de*, *comprender*
EDL de función	*permitir*, *servir* (*a*, *de*, *para*), *utilizar* (*como*, *dentro*), *emplearse para*
EDL de causalidad	*provocar* (*por*), *obtener por*

Tabla 4.4. Tipología de enunciados definitorios según Auger (tomado de Auger (1997: 64))

Podemos ver que Auger clasifica los enunciados definitorios a partir del tipo de verbo o de las formas léxico-sintácticas que incluyen. Cuando dichos verbos o formas refieren a un discurso sobre el propio lenguaje, se habla de *enunciados definitorios metalingüísticos* (EDM). Cuando los enunciados no incluyen una predicación definitoria metalingüística, sino por ejemplo verbos copulativos o de equivalencia, se habla entonces de *enunciados definitorios lingüísticos* (EDL).

Según Auger, los EDM se dividen en EDM de designación, EDM de denominación y EDM sistemáticos; mientras que los EDL incluyen EDL copulativos, EDL de equivalencia, EDL de caracterización, EDL de análisis, EDL de función y EDL de causalidad.

En el caso de los EDM, la información que se aporta sobre [N_0] está introducida por un tipo de predicación definitoria que cumple una función metalingüística de referir la forma en que opera el término o los atributos de su campo conceptual. Los EDM incluyen verbos o formas que describen, expresan o fijan una convención lingüística, ya que aportan información sobre un signo lingüístico, sobre su forma gráfica o fonética, o bien sobre su contenido nocional.

Por el contrario, la clasificación de los EDL atañe a los atributos de un término, así como a los distintos tipos de relaciones conceptuales que éste puede establecer con otros términos de su mismo campo conceptual. Así, los EDL copulativos y de caracterización harían referencia a los atributos de un término, mientras que los EDL de equivalencia, análisis, función y causalidad estarían enfocados en describir las relaciones conceptuales del término, por ejemplo las relaciones de equivalencia a través de las cuales se podrían encontrar los sinónimos de un término específico, o las relaciones de análisis que hacen explícitas las partes que incluye el término.

En comparación con la tipología de las normas ISO y la tipología de De Bessé, la clasificación de Auger, más que tratar de describir los distintos tipos de contextos en los que los términos aparecen, lo que propone es delimitar aquellos casos que contienen un tipo de información definitoria o conceptual. Deja a un lado la clasificación

de contextos del lenguaje o enciclopédicos y se centra sobre todo en la división de *lingüístico* frente a *metalingüístico*. El metalenguaje, señala Auger (1997: 28), puede ser considerado como un lenguaje natural que se utiliza para describir otro lenguaje natural y constituye una base para la transmisión del conocimiento. Su tipología implica por sí misma las particularidades de lingüístico y metalingüístico, siendo la principal característica que distingue a los EDL de los EDM una diferencia estructural de acuerdo con el tipo de verbo o forma lingüístico-sintáctica que en ellos se incluye.

Así, la tipología de Auger tiene como finalidad el estudio lingüístico de predicaciones definitorias recurrentes que permitan la recuperación de los distintos tipos de enunciados definitorios sobre bases de datos textuales. Más adelante veremos el rol que juegan dichas predicaciones definitorias en la elaboración de patrones para la extracción automática de CDs.

4.1.3 Actos performativos definitorios

En Pearson (1998) se puede encontrar otro estudio en la misma línea de describir los contextos de textos especializados donde aparece información definitoria o conceptual sobre los términos. Este tipo de contextos, en consenso con los distintos tipos de verbos performativos propuestos por Austin (1962), Pearson los denomina *actos performativos definitorios*[1].

El acto de definir dentro de las situaciones comunicativas del lenguaje especializado puede ser considerado como un acto performativo que se usa para describir enunciados del discurso que refieren a un concepto representado mediante un término. Dependiendo de la forma en que dichos enunciados expresan una definición, al igual de quien la expresa en la situación comunicativa, los actos performativos definitorios pueden ser divididos en dos clases generales: *ejercitivos definitorios* y *expositivos definitorios*.

De manera general, los ejercitivos definitorios se refieren a aquellos contextos donde se formula y expresa una definición por primera

[1] En palabras de Pearson (1998: 106): "Austin […] uses the term 'performative' to describe verbs which, when used, invokes some form of conventional procedure and which in themselves constitute some form of action".

vez. Por su parte, los expositivos definitorios hacen mención al tipo de contexto donde se reformula o reitera una definición. Los ejercitivos definitorios pueden ser divididos en *explícitos* o *implícitos*, mientras que los expositivos definitorios se dividen a su vez en *formales* y *semi-formales*. En la siguiente figura se observa una representación de las distintas divisiones de los actos performativos definitorios.

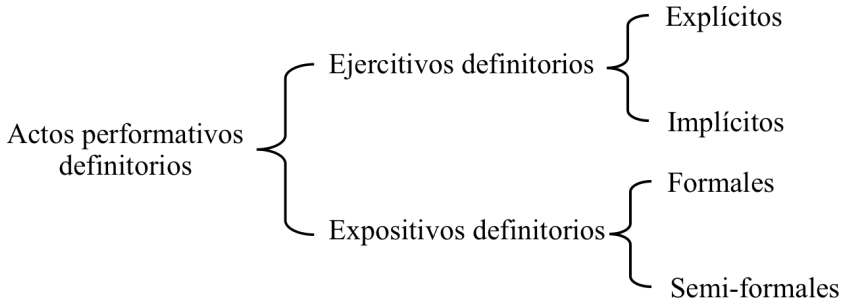

Figura 4.1. Tipología de actos preformativos definitorios según Pearson (tomado de Pearson (1998: 105-120))

Los ejercitivos definitorios pueden ser vistos como los actos de definir por primera vez un término. En estos casos, un individuo o grupo de individuos elaboran un nuevo concepto y lo vinculan con un significado particular, o bien parten de un término ya existente de un dominio de conocimiento distinto y le asignan un significado completamente nuevo. Estos actos, por lo general, los llevan a cabo especialistas relacionados con la práctica terminográfica prescriptiva, como pueden ser comités nacionales o internacionales de normalización, u organizaciones profesionales que delimitan cómo debe ser entendido y usado un término en un contexto particular. Normalmente, los ejercitivos definitorios tienden a ocurrir en situaciones comunicativas entre expertos de un dominio de conocimiento.

Los ejercitivos definitorios siguen un patrón prototípico que Pearson describe como: *I hereby define X as Y*, el cual suele encontrarse en situaciones comunicativas donde no existe un consenso precedente sobre el significado que describe el término X. Si este tipo de patrones aparecen en la situación comunicativa, los actos performativos se clasifican como *ejercitivos definitorios explícitos*; de lo contrario, si no aparece una mención de esta clase, se habla entonces de *ejercitivos definitorios implícitos*.

En los ejercitivos definitorios explícitos, el autor del contexto hace evidente la autoría de la definición que propone, o bien hace referencia a su participación en la elaboración de dicha definición. Asimismo, también se pueden hacer explícitas cierto tipo de indicaciones del alcance de la definición. Pearson (1998: 114-115) provee los siguientes contextos como ejemplos de ejercitivos definitorios explícitos:

- It seems clear that to utter the sentence... is not to describe my doing of what I should be said in so uttering to be doing or to state that I am doing it: it is to do it… What are we to call a sentence or an utterance of this type? I propose to call it a performative sentence, or a performative utterance, or, for short, 'a performative'.
- 'A phrase can be defined for the moment as a co-occurrence of words which creates a sense that is not the simple combination of each of the words'.

En el primer contexto se observa la evidencia explícita de la autoría en la definición del término *performative sentence* mediante el patrón *I propose to call it*. En el segundo ejemplo, la frase *for the moment* hace explícita una indicación relativa al hecho de que la definición provista no es una versión final.

En los ejercitivos definitorios implícitos no se hace evidente ninguna señal sobre la autoría de la definición. Este tipo de contextos suelen estar relacionados, por ejemplo, con los casos en que autores altamente reconocidos en su campo de estudio no necesitan proveer referencias explícitas sobre la autoría de una definición propia. El hecho de que en estos casos el autor haya propuesto una nueva definición se hace evidente con el proceso de reiterarla posteriormente, o bien cuando otros autores recurren a ésta y la hacen explícita en sus propias situaciones comunicativas.

Por otro lado, los expositivos definitorios incluyen contextos con definiciones que ya han sido formuladas con anterioridad y tienden a ser más numerosos en las distintas situaciones comunicativas entre expertos y semi-expertos, o bien entre expertos y no-expertos. En estos casos las definiciones se reformulan o repiten con el propósito de reiterar el significado de un término y vehicular información pertinente. Los expositivos definitorios están

relacionados asimismo con el matiz polisémico de los términos en el sentido de que éstos pueden tener diferentes significados en distintos dominios de conocimiento, o pueden a su vez tener otro significado dentro del lenguaje general, por lo cual los autores se ven en la necesidad de explicar cuál de estos significados es el pertinente en su caso específico. Pearson agrega que este tipo de definiciones son comunes en tesis donde se definen términos básicos para demostrar que se han comprendido los conceptos pertinentes en la investigación que se desarrolla.

Los expositivos definitorios suelen estar señalados explícitamente en textos donde el nivel de conocimiento especializado es relativamente similar entre el autor y el lector. No obstante, son más comunes los casos en que los expositivos definitorios no se hacen explícitos, siendo la única vía para su identificación frases como *se define como*, *denominado*, *consiste de*, *contiene*, etc., las cuales se usan para introducir un concepto que los autores creen pertinente.

Por último, los expositivos definitorios se pueden clasificar a su vez en *expositivos definitorios formales* y *expositivos definitorios semi-formales*. Los primeros incluyen información relativa sobre la clase a la cual pertenece un término, así como las características que lo distinguen de las demás unidades terminológicas de su mismo dominio conceptual. Estos expositivos se representan mediante la siguiente fórmula:

$$X = Y + \textit{características distintivas}$$

El lugar de *X* lo ocupa el término, mientras que el lugar para "=" representa una predicación definitoria del tipo *se define como*. Por su parte, *Y* representa la clase a la que pertenece dicho término y puede estar constituida por otro concepto.

Los expositivos definitorios semi-formales suelen incluir únicamente ciertas características pertinentes del término que lo permitan distinguir de otros términos del dominio. En este caso la fórmula omite el elemento Y, y se representa de la siguiente forma:

$$X = \textit{características distintivas}$$

Tomamos los ejemplos provistos por Pearson para ejemplificar los expositivos definitorios formales y semi-formales, respectivamente (Pearson 1998: 145-159):

- Telewriting is a communication technique that enables the exchange of handwritten information through telecommunication means.
- Digital transfer links are used to interconnect interface adaptors to form signalling data links.

En el primer caso, observamos que se provee la clase general a la cual pertenece el término *telewriting*, es decir, *a communication technique*, que contiene la característica de: *enables the exchange of handwritten information through telecommunication means*. El segundo ejemplo únicamente hace mención a una característica conceptual del término *digital transfer links*, los cuales son usados *to interconnect interface adaptors to form signalling data links*.

En resumen, la tipología de Pearson parte del estudio de cómo son empleadas las definiciones en las diversas situaciones comunicativas, y describe la forma en que los actos performativos definitorios transmiten en mayor o menor grado cierto tipo de información metalingüística explícita o implícita, la cual provee datos sobre el contexto real de uso de los términos. Ciertamente, lo que clasifica Pearson no son todos los tipos de contextos de aparición de los términos, sino aquellos que incluyen un tipo específico de información definitoria. Dentro de este grupo de contextos clasifica dos clases generales, dependiendo de que la definición se presenta por primera vez, o, por el contrario, sea una reformulación de una definición previa.

Al igual que el trabajo de Auger, la finalidad de la clasificación de Pearson es proveer las bases para sistematizar la extracción de información definitoria y conceptual sobre textos especializados o corpus lingüísticos. Específicamente, el tipo de información de interés para Pearson es el relacionado con aquellos contextos donde se reformulan las definiciones, los cuales tienden a ser más numerosos y presentarse en las diversas situaciones comunicativas del lenguaje especializado.

4.1.4 Contextos ricos en conocimiento

Con el mismo fin de proveer un estudio empírico para sentar las bases de la extracción automática de información definitoria, Meyer (2001) propone una categorización simple y más genérica de los tipos de contextos que contienen información conceptual. Meyer (2001: 281-282) define como *contextos ricos en conocimiento* (CRCs) a aquellos contextos que indican por lo menos una característica conceptual del término, ya sea un atributo o una relación, y que en el marco de la práctica terminográfica sean útiles para:

1. Proveer definiciones.
2. Proveer puntos de partida para formular definiciones.
3. Incrementar el conocimiento del terminógrafo sobre el área en la que trabaja.

Sobre el primer punto, la autora menciona que los CRCs pueden conformar entradas terminográficas de un término, ya que dichos contextos contienen información considerada de *alta calidad*. Cuando la información se considera de *baja calidad*, como es el caso del segundo punto, los contextos sirven de inicio para la formulación de definiciones tomando en cuenta que en ciertos medios de trabajo terminográfico, es el conjunto de varios CRCs lo que permitiría la elaboración de definiciones a partir de la selección de los elementos que se consideren indispensables en dichos contextos. Con respecto al tercer punto, el conjunto de todos los contextos constituyen un bagaje de conocimiento que el especialista debería tener con el fin de poder llevar a cabo tareas como la identificación de sinónimos, el establecimiento de sinónimos entre términos de diferentes lenguas, o bien la propia estandarización de los términos.

Meyer clasifica los CRCs dependiendo del tipo de definiciones que éstos incluyen. Partiendo de un modelo aristotélico de la definición, el cual sigue la fórmula: *X = género próximo + diferencia específica*, donde *X* representa el término, *género próximo* es la categoría general a la cual pertenece dicho término, y *diferencia específica* es lo que distingue la categoría general del término que se define, el estudio de Meyer propone que las definiciones de los

CRCs pueden ser de dos tipos: *CRCs definitorios* y *CRCs explicativos*.

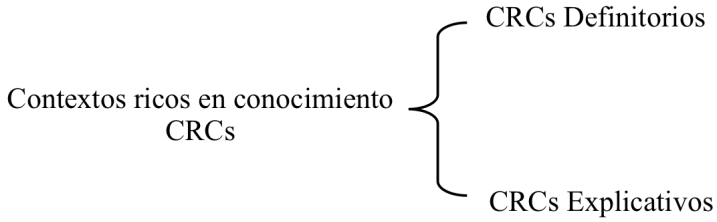

Figura 4.2. Tipología de contextos ricos en conocimiento según Meyer (tomado de Meyer (2001: 283-287))

Los CRCs definitorios son el tipo más común e incluyen una definición que presenta la fórmula aristotélica antes descrita: d*efinición = género próximo + diferencia específica,* que en la fórmula de Meyer está dada como:

$$X = Y + \textit{características distintivas}$$

Esta fórmula contiene cuatro componentes: X representa el término que se define; Y es el hiperónimo o la clase general a la cual pertenece X; las características distintivas representan aquella información que distingue el concepto designado por X de otros miembros de su misma clase; y el elemento "=" indica que tanto X como las características distintivas (*definiendum* y *definiens* respectivamente) deberían poder ser intercambiados por sí mismos en cualquier contexto sin alterar el sentido de la oración.

Los CRCs explicativos son aquellos donde sólo se proporciona información sobre las características del término sin incluir la clase general a la cual pertenece. En la fórmula de Meyer, estos contextos se representan con la fórmula:

$$X \supset \textit{características}$$

En este caso X corresponde de igual forma al término, mientras que el símbolo "⊃" señala que X debe tener una o varias características conceptuales. A diferencia del trabajo de Pearson, el símbolo "⊃" sustituye a "=", en el sentido de que X *contiene* dichas características en lugar de ser equiparable a ellas. También elimina el adjetivo de *distintivas* con la finalidad de expresar que no

siempre dichas características proveerán información que permita distinguir a X de otros elementos de su misma clase.

Los CRCs definitorios se consideran los contextos más completos ya que en ellos se detalla la clase a la que pertenece el término, es decir, su hiperónimo, y se proporciona además información relevante sobre sus atributos. En el caso de los CRCs explicativos, la información que proveen sobre el término sólo permite crear una clasificación de éste a partir de la relación conceptual que establece con otros términos de su misma clase. En este último tipo encontramos relaciones conceptuales como meronimia o sinonimia, por citar algunas. Meyer (2001: 282) propone los siguientes ejemplos para ilustrar la tipología de CRCs:

- Compost is a practical method of recycling organic matters, resulting in improved soil structure and strong, disease-resistant plants.
- Compost, a dark, nutrient-rich soil conditioner, consists of a small amount of soil along with decomposed or partially decomposed plant residues.
- Compost enriches topsoil with organic matter and plant nutrients, improves water infiltration, and increases water availability and nutrient retention in sandy soils.
- Compost contains nutrients, nitrogen, potassium and phosphorus.

En los primeros dos casos encontramos ejemplos de CRCs definitorios de alto nivel que describen el término *compost* a partir de la enunciación de las clases inmediatas a las cuales pertenece, *method of recycling* y *soil conditioner*, junto con una serie de atributos y características específicas.

En los últimos dos ejemplos se omite cualquier hiperónimo pero se aportan descripciones generales sobre las características del término, lo que los lleva a ser considerados contextos de bajo nivel. No obstante, estos casos pueden funcionar como puntos de inicio para la formulación de definiciones, ya que proveen conocimiento conceptual sobre el término.

La tipología de Meyer, como hemos señalado, es un acercamiento más genérico a la descripción de los contextos de uso de los

términos en textos especializados en los que se aporta información definitoria, descriptiva o conceptual. Coincide con los estudios de Auger y Pearson en la finalidad de describir este tipo de contextos para la elaboración de una metodología que permita su extracción automática sobre textos especializados. Por otro lado, a diferencia de aquellos dos autores, la categorización de Meyer se centra básicamente en el estudio de dos grandes bloques a partir de la presencia o ausencia del hiperónimo y de la relación del término con sus características (distintivas o no).

4.1.5 Hacia una definición aplicada de contexto definitorio

Hasta ahora hemos visto algunas consideraciones básicas en el estudio del *contexto* como un elemento clave en la práctica del trabajo terminográfico. Hemos reseñado algunas propuestas que parten del tipo de información que se presenta sobre el término para clasificar los contextos en diversas categorías de acuerdo con sus características, principalmente lingüísticas, metalingüísticas o definitorias.

Respecto a las tipologías de las normas ISO y de De Bessé, ambas coinciden en distinguir diversos tipos de contextos de acuerdo con el contenido de la información que se plasma en ellos. Sin embargo, en las normas ISO no se hace evidente una diferenciación global de los contextos que contienen información definitoria y/o conceptual frente a aquellos que representan simples manifestaciones de los términos.

En la tabla 4.5 ejemplificamos una representación general de ambas tipologías donde podemos ver similitudes y diferencias.

Tipo[2]	ISO	De Bessé
Contextos ligados a un término	Lingüísticos	Del lenguaje Fraseológicos Lingüísticos De uso
Contextos que refieren a un concepto	Metalingüísticos	Materiales
	Explicativos	
	Asociativos	Conceptuales
	Definitorios	Definitorios Enciclopédicos

Tabla 4.5. Comparación entre la tipología de contextos en las normas ISO y De Bessé

Observamos que las normas ISO sugieren cinco clases dentro de las cuales se clasifican a los contextos lingüísticos como contextos que únicamente manifiestan la existencia de un término. Para De Bessé, estos contextos constituyen una clase propia denominada *contextos ligados a un término* y los distingue de otra clase general denominada *contextos que refieren a un concepto*.

En este último grupo, De Bessé clasifica a los CDs como aquellos donde se aporta información sobre los atributos de los términos. De igual forma, las normas ISO sugieren un grupo particular para el caso de los CDs, que a diferencia de los demás contextos contienen información más sustancial que permita entender el concepto que representa el término.

Los contextos conceptuales y materiales, para De Bessé, pueden ser equiparables respectivamente a los contextos asociativos y metalingüísticos de las normas ISO. En los primeros se refieren características sobre las relaciones conceptuales de los términos, mientras que los segundos proveen instrucciones sobre el alcance de los términos y la forma en que éstos operan en un contexto determinado.

Ahora bien, en la práctica, estas clasificaciones se tornan un tanto ambiguas, ya que es difícil que los casos reales de los contextos en

[2] Según De Bessé

los que aparecen los términos se ajusten a las categorías que sugieren. En los textos especializados, tanto los contextos que refieren a un concepto como las denotaciones de explicativo, asociativo, metalingüístico y definitorio suelen constituir un conjunto de información estrechamente entrelazada que es difícil de categorizar por separado, tomando en cuenta que en la mayoría de las ocasiones dicha información no aparece de forma aislada.

Asimismo, toda información que se aporte sobre una unidad del lenguaje, en este caso los términos, constituye por sí misma una enunciación metalingüística, la cual, señala Rodríguez (1999: 65):

> "predica sobre un lenguaje-objeto desde un nivel semiótico superior a partir del cual aporta precisiones sobre su forma de operar y sus posibles contenidos. El sujeto sobre el que recae la predicación es también un lenguaje, o más exactamente un elemento de ese lenguaje considerado desde la perspectiva de una entidad del discurso como cualquier otra".

La información que contienen los contextos explicativos o asociativos, por ejemplo, es a su vez un tipo de información definitoria que puede ser igualmente útil en la formulación de definiciones de un término particular. Considérense los siguientes contextos de uso de diferentes términos[3] (el resaltado en negrita es nuestro):

- Ya se ha hecho mención de que el propio concepto de **gen** ha ido cambiando a medida que ha progresado el conocimiento, pero en la mayoría de los casos se entiende como **gen** una unidad transcripcional, incluyendo sus regiones reguladoras asociadas.
- Como en los seres humanos y en otros organismos superiores los **genes** están formados por ADN, un **gen** puede definir se como la cadena o secuencia lineal de los nucleótidos del ADN que especifica la secuencia lineal de los aminoácidos que forman una proteína (o una cadena polipeptídica).
- "**Desnaturalización**" también se usa para describir la pérdida de la estructura proteica correcta; es un término general que

[3] Tomados del Corpus Técnico del IULA (Bach *et al*. 1997).

significa que la conformación natural de una molécula ha sido alterada.

- Hechas estas consideraciones, debemos señalar que, en el lenguaje habitual, los términos de **inmunógeno** y **antígeno** se utilizan como sinónimos (concediendo al término antígeno el sentido amplio de la definición clásica).

En el primer caso, se habla del concepto de *gen* y de cierta información metalingüística temporal respecto a su significado a lo largo del tiempo; se aporta además información sustancial que permitiría conocer un atributo específico del término: *unidad transcripcional*, al igual que elementos de su dominio conceptual: *regiones reguladoras asociadas*.

De la misma forma, en el segundo ejemplo se menciona un elemento del dominio de conceptual del término: *los genes están formados por ADN*, y se conjunta además con información definitoria: *la cadena o secuencia linear de los nucleótidos*... etc.

El tercer contexto hace mención a una forma en la que también se puede usar el término *desnaturalización*, y se añaden además ciertas características explicativas sobre él: *significa que la conformación natural de una molécula ha sido alterada*.

Por último, el cuarto ejemplo contiene una serie de información que sitúa a los términos *inmunógeno* y *antígeno* en el mismo plano conceptual, estableciendo para dichos términos una relación sinonímica, y señalando además una referencia metalingüística explícita a su carácter de términos.

De esta forma, el conjunto de información que se presenta en cada uno de los contextos constituye en todo caso un enramado de información representativa y relevante que permite comprender los conceptos que representan los términos. Con estos ejemplos no pretendemos asegurar que en los textos especializados resulte inviable categorizar en clases independientes la información que transmiten los contextos de los términos, ya que siempre habrá casos que se ajusten a estas tipologías específicas. Por el contrario, nuestro interés radica en señalar la complejidad de la información que contienen los contextos y que deviene en una difícil categorización en alguna de estas clases preestablecidas.

Para nosotros, los distintos contextos son una serie de predicaciones lingüísticas y metalingüísticas que vehiculan, en mayor o menor grado, conocimiento sobre los términos, y que permiten la adquisición de las pautas necesarias para conocer su significado. Así, coincidimos entonces en mayor medida con las perspectivas de Auger, Pearson y Meyer, tomando en cuenta que en esta investigación los contextos de nuestro interés son los CDs entendidos desde la perspectiva de unidades que transmiten conocimiento sobre los términos de un área de especialidad, y que son útiles para entender la forma en que dichos términos operan y para representar su significado.

La perspectiva de Auger, Pearson y Meyer demuestra la necesidad de explicar los contextos a partir del contenido semántico que vehiculan, así como categorizarlos mediante el estudio y análisis de la forma en que éstos suelen representarse en textos especializados. Auger afirma que los contextos de los términos son *enunciados de interés definitorio* cuando arrojan información relativa a su forma lingüística (sentido, contextos de uso, grafías), sus atributos o sus propiedades.

Hemos mencionado que este tipo de contextos Pearson los denomina *actos performativos definitorios*, partiendo de la idea de que la formulación de definiciones supone un acto performativo usado para describir enunciados del discurso que refieren a las características conceptuales y atributos de un término.

En este mismo sentido, Meyer considera a los contextos donde se presenta información conceptual de un término como *contextos ricos en conocimiento*.

En la figura 4.3 podemos ver un esquema de las distintas clasificaciones con una representación aproximada del lugar que podría ocupar cada clase de contextos frente a las categorías de las demás clasificaciones.

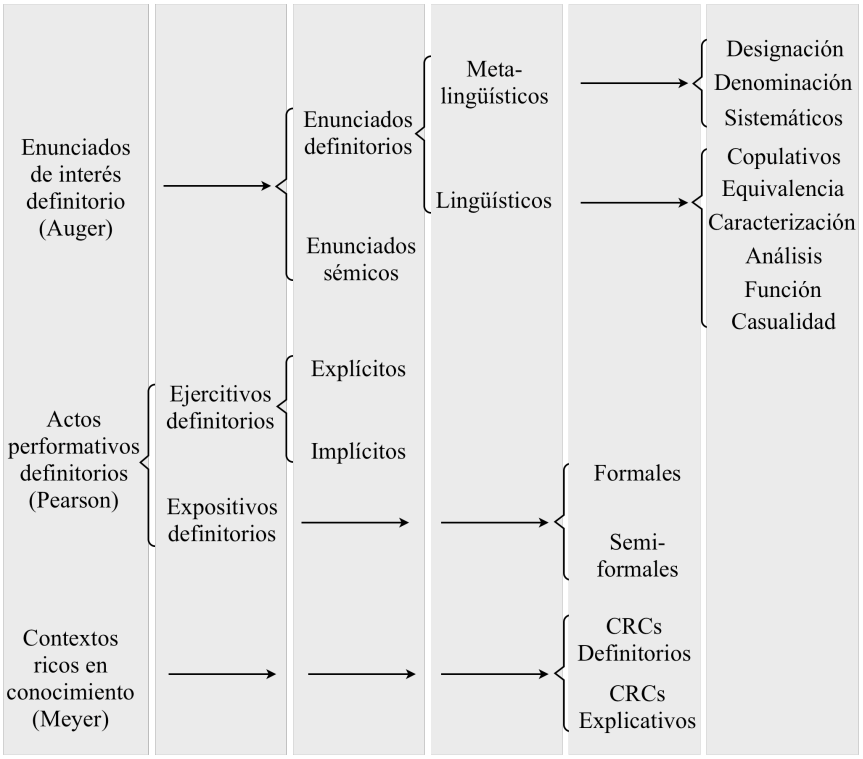

Figura 4.3. Comparación entre las tipologías propuestas por Auger, Pearson y Meyer

La primera distinción que se establece en la tipología de Auger es entre enunciados definitorios y enunciados sémicos, donde la diferencia radica en que la información que se provee sobre el término se manifiesta explícita o implícitamente a través de una serie de verbos o formas lingüístico-sintácticas. En este sentido, coincide con la categorización que hace Pearson al nivel de los ejercitivos definitorios, pero difiere en la profundidad en la representación de las clases precedentes. Pearson, antes de clasificar los ejercitivos definitorios en explícitos o implícitos, sugiere que hay dos clases generales relacionadas con el hecho de que la definición del contexto se exponga por primera vez o sea una reformulación de una definición ya establecida. Estas clases las denomina respectivamente *ejercitivos definitorios* y *expositivos definitorios*.

Retomando los enunciados definitorios, Auger los divide dependiendo de los tipos de verbos o formas lingüístico-sintácticas

que se utiliza en ellos para vincular un término con su respectiva definición. Cuando se utilizan verbos o formas del tipo *llamarse*, *significar*, *el sustantivo*, *el sintagma*, etc., es decir elementos que refieren al mismo lenguaje, se consideran como *enunciados definitorios metalingüísticos*. Cuando los verbos o formas lingüístico sintácticas utilizan elementos del tipo *equivaler a*, *compuesto por*, *características*, *atributos*, los cuales no se utilizan exclusivamente para referirse al propio lenguaje, se consideran entonces como *enunciados definitorios lingüísticos*.

A este mismo nivel no existe otra clasificación por parte de las tipologías de Pearson o Meyer, sin embargo, desde la categorización de enunciados definitorios, expositivos definitorios y contextos ricos en conocimiento cada autor propone ciertas fórmulas que tienen varias características en común.

	Tipo	Fórmula
AUGER	Enunciados definitorios	[N0] → [verbo o forma sintáctico-lingüística] → [N1–X]
PEARSON	Expositivos definitorios formales	X = Y + características distintivas
	Expositivos definitorios semiformales	X = características distintivas
MEYER	CRCs Definitorios	X = Y + características distintivas
	CRCs Explicativos	X ⊃ características

Tabla 4.6. Fórmulas de representación de contextos con información definitoria

Pearson y Meyer proponen un tipo de contextos denominados *expositivos definitorios formales* y *CRCs definitorios* que se expresan por una fórmula donde se refleja la relación que establece X e Y en un plano conceptual a través de predicaciones definitorias representadas por el símbolo "=". En dicha fórmula, X corresponde a un término, Y al hiperónimo o la clase general a la cual pertenece X, y las características distintivas hacen mención a atributos o

relaciones conceptuales del término. Para Auger, esta misma fórmula está representada sustituyendo X por [N_0], el signo "=" por [verbo o forma sintáctico lingüística], Y por [N_1], y las características distintivas por [X].

En lo que difieren Pearson y Meyer de Auger es en proponer una fórmula distinta para los casos en que se elide Y quedando de esta forma un término relacionado con sus atributos o características conceptuales. Meyer propone además que en los casos donde se elide el elemento Y de la definición, el símbolo "=" se sustituya por "⊃", con lo cual se indicaría que el término tiene o debe tener alguna característica conceptual para poder se considerado como un CRC.

Por último, Auger especifica en un último nivel distintos tipos de contextos en que se pueden clasificar los enunciados definitorios metalingüísticos y los enunciados definitorios lingüísticos. Los primeros denotan una diferencia sutil entre distintos tipos de enunciados metalingüísticos, mientras que los contextos del segundo grupo representan diversos tipos de relaciones conceptuales entre X e Y.

En resumen, las tipologías de Auger, Pearson y Meyer no incluyen una clasificación genérica sobre las clases de contextos que representan las ocurrencias simples de los términos, sino se ciernen a los contextos de textos especializados que informan sobre las características definitorias, conceptuales o metalingüísticas de un término.

A partir de lo expuesto, y tomando en cuenta nuestro objetivo de desarrollar una herramienta para la extracción de información definitoria sobre términos especializados, en esta investigación utilizamos el concepto de *contexto definitorio* en un sentido más general. Consideraremos como *contextos definitorios* a aquellos contextos de los textos especializados donde se aporta información relevante sobre los atributos, características y relaciones conceptuales de los términos. Dicha información permite entender el significado y la forma en que operan los términos, al igual que conocer las relaciones que establecen con otros términos para poder situarlos en el contexto global del dominio de conocimiento al cual pertenecen.

La definición de que partimos se acerca más a los conceptos de *enunciados de interés definitorio*, *actos performativos definitorios* y *contextos ricos en conocimiento*, que a la definición propia de CD provista en la norma ISO 12620 (1999: 25) y en De Bessé (1991: 112). En estas dos tipologías, un CD es aquél que incluye información que permite comprender el significado de un término pero que sería insuficiente para la redacción de una definición, ya que no posee el rigor formal de éstas.

Para los fines de nuestra investigación, creemos además poco operativa dicha consideración de que en los CDs la información que se presenta carece del rigor formal de una definición, ya que en ocasiones dichos contextos pueden estar constituidos por reformulaciones o reiteraciones de definiciones ya preestablecidas, o por partes de éstas, que no necesariamente pierden el rigor de una definición terminográfica tal y como se podrían encontrar en un registro ya elaborado. No obstante, coincidimos con las tipologías de las normas ISO y de De Bessé en la división primaria que existe entre los CDs y los contextos lingüísticos o contextos ligados a un término, los cuales representan el uso de las unidades léxicas especializadas en su contexto real pero no aportan información que permita conocer su significado. Si bien los contextos lingüísticos en general permiten adquirir conocimiento sobre la forma en que se usan los términos, en los CDs dicho conocimiento se especifica a nivel definitorio, conceptual o metalingüístico, con lo que su alcance se extiende no sólo a las actividades terminográficas relacionadas con el registro y explicación de los significados de los términos, sino también a aquellas actividades de diversos campos de conocimiento que necesitan estudiar los conceptos y sus significados en las situaciones reales en que se utilizan.

En realidad, los CDs conforman por sí mismos una unidad discursiva que sigue de manera sistemática ciertas pautas léxico-sintácticas. En los siguientes apartados señalamos cuáles son estas pautas recurrentes al igual que ciertas consideraciones básicas respecto a sus elementos constitutivos.

4.2 Análisis de contextos definitorios en español

Este apartado lo enfocamos a describir las características sintácticas de los CDs en español, con base en datos obtenidos en una investigación propia a nivel de tesis de licenciatura que pueden consultarse principalmente en Alarcón (2003) y Alarcón y Sierra (2003). Este análisis que detallamos a continuación ha servido para establecer las bases necesarias en el desarrollo de reglas para la extracción automática de CDs.

Para llevar a cabo un estudio sintáctico de CDs en español, conformamos en primer lugar un corpus de análisis a partir del Corpus Lingüístico de Ingeniería (CLI) (Medina *et al*. 2004), desarrollado en el Grupo de Ingeniería Lingüística de la Universidad Nacional Autónoma de México. De este corpus seleccionamos 20 textos pertenecientes a las siguientes áreas:

- Estructuras bioclimáticas
- Logística
- Sistemas expertos
- Transporte

El corpus de análisis estaba compuesto por tesis, informes a patrocinadores y artículos en congresos, y constaba aproximadamente de 300,000 palabras. Recurrimos a este tipo de documentos debido a que usualmente incluyen un apartado (introducción, presentación o bien un capítulo específico) que funciona como marco teórico donde se definen los términos esenciales para la comprensión del contenido.

4.2.1 Metodología para la delimitación de contextos definitorios

El primer paso de la metodología consistió en el análisis manual de cada documento del corpus con la finalidad de encontrar contextos donde se definiera un término. En esta delimitación manual se tuvo en cuenta que un CD sería aquél fragmento textual donde se aportara información que permitiera comprender el significado de un término, tomando en cuenta que la información contenida en el contexto podía proporcionar datos sobre sus características y

atributos, así como funciones, partes o bien relaciones de éste con otros términos.

A modo de ilustración presentamos algunos de los tipos de CDs encontrados en el corpus de análisis:

- La logística empresarial puede definirse como el conjunto de actividades que tienen por objetivo la colocación, al menor costo, de una cantidad de producto en el lugar y en el tiempo donde una demanda existe.
- *La imaginación* se entiende como la capacidad de meditar o elucubrar sobre ciertas cosas o acontecimientos que no están presentes en el entorno o que pueden manifestarse en el futuro.
- Se utilizan Sistemas Expertos para el monitoreo de procesos, para evitar ciertas desviaciones de las normas establecidas y así se tomen medidas preventivas y/o correctivas en tiempo real.
- Los sistemas de información integrados permiten el acceso en tiempo real a la información asociada a la mercancía facilitando una adecuada gestión de flujos.

Los primeros dos contextos son ejemplos donde se introduce información definitoria de un término y se provee datos sobre sus características y atributos. Por su parte, en los últimos dos ejemplos se informa asimismo de características de los términos, pero en estos casos se habla específicamente sobre su funcionalidad. Más adelante profundizaremos en las características semánticas de los distintos tipos de CDs. Por ahora describiremos sus características estructurales.

Una de estas características estructurales está en relación con la extensión de los CDs. En este primer análisis delimitamos su extensión a un párrafo. No obstante, se encontró que en ocasiones en este mismo párrafo se presentaban oraciones adicionales que ya no contenían información definitoria sobre el término, o bien casos donde un CD abarcaba más de un solo párrafo. Algunos ejemplos de la extensión de los CDs son los siguientes:

- La operación del autotransporte comprende fundamentalmente la actividad de trasladar mercancías y personas de un lugar a otro a través de las carreteras. Asociadas al traslado, se pueden identificar otras tres actividades relevantes: la carga y descarga de mercancías o el abordaje de pasajeros, la administración de los traslados y el mantenimiento de los equipos utilizados.
- Se entiende por paradigma una forma epistemológica que, como instrumento cognoscitivo, permite diferenciar la realidad e identificar y escoger ciertos fragmentos de ella, con el fin de definir el objeto de estudio, así como el modelo que los sustituye en las siguientes fases de la investigación. ¶
  Es así que el paradigma determina el desarrollo del proceso cognoscitivo orientado a descubrir las regularidades y leyes que caracterizan a los fenómenos como objeto de estudio, así como a explicar y aprovecharlas para el control de los mismos [Kuhn, 1982; Gelman, 1974]. ¶
  Considerado como uno de los conceptos básicos de la metodología moderna, el paradigma, entre otras funciones y aplicaciones, sirve también como el instrumento principal para plantear los problemas a través de la interpretación de la problemática, contemplada como la manifestación de los problemas reales.

De los ejemplos anteriores, en el primero de ellos se especifican algunas características que comprende el término *operación del autotransporte*, y en la oración siguiente se comienza a hablar de otra información distinta a la información definitoria provista sobre el término.

En el segundo ejemplo, los saltos de párrafos los indicamos mediante el símbolo ¶. En el primer párrafo se introduce la definición del término *paradigma*, mientras que en el segundo párrafo se continúa definiendo el mismo término señalando algunos de sus atributos. Finalmente, en el tercer párrafo se aporta información sobre algunas funciones del término.

A pesar de la longitud de los CD, por lo general cada párrafo sigue a su vez una estructura prototípica que incluye dos elementos mínimos constitutivos: un término (t) y una definición (d), los cuales se encuentran conectados entre sí mediante un *patrón*

definitorio (pd). Además, encontramos que algunos CDs suelen presentar otro tipo de información metalingüística y pragmática referente a la forma, las condiciones de uso o el alcance operativo de los términos. Dicha información corresponde a lo que denominamos un *patrón pragmático* (ppr).

- <ppr>Tradicionalmente</ppr>, <t>la logística</t> <pd>se define como</pd> <d>el arte militar que estudia el movimiento, transporte y estacionamiento de las tropas fuera del campo de batalla</d>.

En este contexto se define el término *logística* a través de la enunciación de ciertas características distintivas: *el arte militar que estudia el movimiento…*, para lo cual se recurre a un patrón definitorio que corresponde a la estructura *se define como*. Asimismo, este contexto incluye un patrón pragmático, *tradicionalmente*, el cual se utiliza para indicar un matiz especial sobre el significado del término.

De esta manera, un esquema básico y general sobre la estructura de los CDs lo podemos observar en la siguiente representación:

$$\{ T \leftrightarrow PD \leftrightarrow D \} PPR$$

En este caso, los símbolos "**{ }**" enmarcan a los elementos mínimos constitutivos T y D, los cuales se encuentran ligados mediante un patrón definitorio en un sentido bidireccional que se representa con el símbolo "↔", y que en conjunto pueden ser modificados por un patrón pragmático.

Así, una vez identificados los CDs del corpus, el siguiente paso consistió en un análisis con el fin de establecer una primera clasificación de carácter estructural. Como hemos observado a lo largo de esta investigación, en los CDs se emplean recurrentemente una serie de claves tipográficas y sintácticas para conectar al término con la información definitoria que se introduce sobre ellos.

Asimismo, las claves tipográficas suelen emplearse además para resaltar la presencia del término o la definición. Tomando en cuenta lo anterior, se observó que los CDs del corpus de análisis

pertenecían generalmente a alguna de las clases que describimos en seguida.

Los contextos más simples eran aquellos que contenían sólo una marca tipográfica para unir al término con la definición, a la vez que la misma tipografía textual podía usarse para resaltar cualquiera de estos dos elementos.

- **Diseño**: *Desarrollo de configuraciones para la resolución de algún problema en base y sujetándose a sus restricciones*.
- *IMPACTOS AGREGADOS SOCIALES* ¶ Los que impactan a la sociedad, produciendo, por ejemplo, la perturbación de las relaciones familiares.

En el primer ejemplo, el término *diseño* se une a su definición mediante *dos puntos*, a la vez que el término se presenta en negritas y la definición en cursivas. En el segundo caso el término *impactos agregados sociales* se resalta en mayúsculas y cursivas, mientras que la liga a su definición se establece situando al término a modo de título, seguido de un salto de párrafo que representamos con el símbolo ¶.

Por otro lado, otro tipo de patrones igualmente simples eran aquellos donde el término se ligaba a su definición mediante una estructura sintáctica, que generalmente era una frase verbal, aunque también se presentaban marcadores reformulativos:

- De manera general, un Operador Logístico (OL) es una firma que realiza prestaciones logísticas en servicio público que adapta a necesidades específicas de cada cliente.
- Definimos un ramal como aquella sección del acueducto constituida por uno o más tubos interconectados y a lo largo de los cuales no existe derivación alguna, de manera que todos los tubos conducen un mismo caudal.
- Delphi es una Two-Way-Tool, es decir, una herramienta de dos direcciones, porque permite crear el desarrollo de programas de dos formas: una de forma visual en la pantalla, por medio de las funciones de Drag & Drop (Arrastrar y colocar) y la otra a través de la programación convencional, escribiendo el código.

El primer caso es un ejemplo prototípico del uso del verbo *ser* más un determinante para expresar una definición del término *operador logístico*. En el segundo ejemplo el término *ramal* se define empleando una estructura sintáctica donde el verbo *definir* se utiliza en conjunto con el adverbio *como* para introducir la información definitoria. El tercer ejemplo recurre igualmente al uso del verbo *ser* y se incluye además el marcador reformulativo *es decir* para expandir el contenido de información sobre el término *Two-Way-Tool*. En estos tres casos notamos que no se recurre a la tipografía textual para resaltar la presencia del término o la definición.

Los dos casos anteriores constituyen una parte significativa del total de CDs encontrados en el corpus. No obstante, los contextos más recurrentes eran aquellos donde se utilizaba tanto un patrón verbal sintáctico para conectar al término con su definición, a la par que se recurría a alguna marca tipográfica para resaltar dichos elementos.

- **La energía primaria**, por definición, es aquel recurso energético que no ha sufrido transformación alguna, con excepción de su extracción.
- Se denomina "equipo de salud" a todo el personal del hospital que tiene una función directa o indirecta para el paciente.

En estos dos contextos observamos que los términos se resaltan con negritas y comillas, a la vez que se emplean frases sintácticas con verbos como *ser* y *denominar*, respectivamente. Podríamos decir que este tipo de contextos presentan una estructura más sólida en tanto que utilizan elementos que permiten resaltar visual y gramaticalmente la presencia de un contexto con información definitoria.

Finalmente, y siguiendo esta clasificación estructural, otro grupo encontrado fue aquél donde en un mismo CD se definen dos o más términos, y donde a veces la definición de un término puede servir a su vez como CD para otro término y su correspondiente definición.

- Los aleros son elementos horizontales que sobresalen de la parte superior de la ventana y obstruyen la componente vertical de la radiación y los quiebrasoles son elementos verticales colocados junto a las ventanas y que obstruyen la componente horizontal de la radiación.

- PROCESOS ADMINISTRATIVOS DE LOS VEHÍCULOS COMERCIALES ¶ Son sistemas administrativos que graban la bitácora de viaje de los vehículos comerciales. Dicha bitácora registra datos de las jurisdicciones políticas atravesadas, de los consumos de combustible y del kilometraje recorrido en cada jurisdicción.

En el primer caso se proveen dos definiciones en el mismo CD para los términos *aleros* y *quiebrasoles* a partir del verbo *ser*. Mientras que en el segundo ejemplo se define principalmente el término *procesos administrativos de los vehículos comerciales*, resaltándolo en mayúsculas y situándolo a modo de título, pero a la vez se presenta en el mismo CD cierta información funcional sobre el término *bitácora de viaje*, cuya introducción se realiza mediante una referencia anafórica. Este último tipo de casos, si bien no ocurrieron en un gran porcentaje con respecto a los demás, nos permitieron darnos cuenta de la complejidad de formas en que pueden introducirse CDs en los textos especializados.

A partir de esta primera clasificación estructural pudimos entonces delimitar cuatro tipos primarios de CDs:

1. **CDs tipográficos**. Incluyen únicamente una marca tipográfica para unir al término con su definición.
2. **CDs sintácticos**. Se conforman principalmente por frases verbales o reformulativas para conectar los términos y sus definiciones. En estos casos no se incluye ningún tipo de marca tipográfica para resaltar los elementos constitutivos de los CDs.
3. **CDs mixtos**. Este tipo de patrones son una combinación de los dos anteriores, donde se emplea una frase verbal o un marcador reformulativo como conector entre el término y la definición, y donde además se resalta tipográficamente la presencia de cualquiera de estos dos elementos.
4. **CDs complejos**. Los cuales representan los casos donde en un CD se definen dos o más términos.

En total encontramos 254 CDs divididos en los siguientes casos:

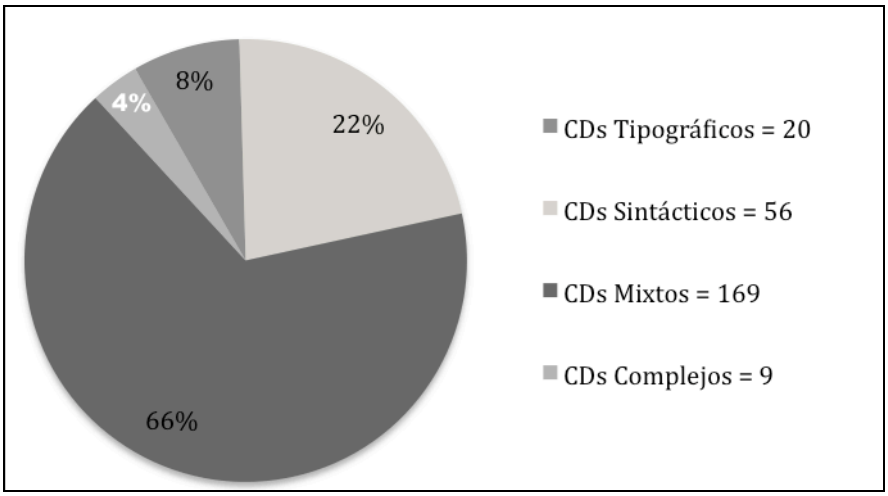

Figura 4.4. Ocurrencias de CDs en el CLI

Como podemos observar del gráfico anterior, la mayor ocurrencia se daba en CDs donde se emplea una marca o patrón sintáctico para conectar el término y la definición, y además alguno de estos dos elementos está resaltado visualmente por un patrón tipográfico, conformando el 66% del total con 169 ocurrencias. El segundo lugar era para los casos donde se emplea solamente un patrón sintáctico, con 56 ocurrencias que representaban el 22% del total. En una cantidad menor ocurrieron CDs donde se utiliza únicamente un signo tipográfico para ligar al término y la definición, los cuales aparecían 20 veces y formaban el 8% de ocurrencias. Por último, los patrones complejos donde en un mismo contexto se define más de un término aparecieron 9 veces, conformando el 4% del total.

La ayuda de marcas tipográficas es un recurso visual que denota un elemento importante para la comprensión del texto. De estos elementos no sólo se utilizan marcas tipográficas, sino también se emplean viñetas, títulos, listas y otros elementos estructurales que se demarcan del formato global del texto para señalar una definición.

Otro elemento de suma importancia lo constituyen las construcciones sintácticas verbales que se usan para ligar al término con su definición. En este sentido, encontramos dos tipos de construcciones sintácticas verbales: aquellas donde sólo se emplea un verbo de manera aislada, como *entendemos* o *definimos*, y

aquellas donde se recurre al uso de alguna partícula gramatical, por ejemplo el pronombre *se* y el adverbio *como* en la construcción *se entiende como*. En la siguiente tabla presentamos algunas de estas construcciones verbales encontradas en los CDs.

Verbos en contextos definitorios	
Se comprende como	Se entiende como
Concebimos a	Incluye
Conocemos como	Se interpreta como
Consiste en	Se llama
Consta de	Permite
Se describe como	Se refiere a
Se define como	Se utiliza como
Denominamos a	Se usa en

Tabla 4.7. Ejemplos de verbos en CDs

En este primer análisis encontramos una variedad de verbos que podían emplearse para introducir la definición de un término. Algunos de estos verbos pertenecen a un grupo claramente más metalingüístico, es decir que se emplean para referirse al propio lenguaje, mientras que otros podría decirse que son de lengua general, ya que se utilizan en una gran variedad de enunciados que no necesariamente son enunciados definitorios. En el siguiente capítulo detallaremos la metodología seguida para determinar el paradigma verbal utilizado en la extracción automática de CDs.

Finalmente, como resultado de este análisis pudimos establecer una tipología de patrones definitorios recurrentes en español, la cual detallaremos en los siguientes apartados.

4.2.2 Tipología de patrones definitorios

Hasta ahora hemos señalado una primera tipología estructural de los CDs encontrados en el corpus de análisis. Señalamos que en esta tipología tomamos en cuenta los tipos de patrones que se emplean para conectar al término con su definición o para resaltar visualmente su presencia dentro del texto. Estos patrones, que denominamos *patrones definitorios*, indican explícitamente cierto

tipo de información conceptual y constituyen una clave esencial en el proceso de reconocer CDs de forma automática.

Para explicar el concepto de *patrones definitorios* retomamos las tipologías de Pearson y Meyer. La primera autora señala que los patrones definitorios pueden ser *expositivos definitorios formales* y *expositivos definitorios semi-formales*, donde los primeros pueden a su vez clasificarse en *simples* y *complejos*.

Los *expositivos definitorios formales simples* corresponden directamente a la tipología de definiciones formales de Trimble[4], en la que se proporciona tres elementos: el término, su hiperónimo y sus características distintivas. En estos casos los patrones definitorios son verbos o frases verbales, como los verbos conectivos *comprender*, *consistir*, *definir*, *denotar*, *designar*, *es/son*, o las frases verbales *es/son llamado*(*s*), *es/son definido*(*s*) *como*, *es/son conocido*(*s*) *como*.

Los *expositivos definitorios formales complejos* incluyen de igual forma los tres elementos de las definiciones formales de Trimble pero requieren un procesamiento adicional para poder incorporarse en la fórmula *X = Y + características distintivas*. Estos casos suelen tener dos o más de dos párrafos y Pearson señala como ejemplos estructuras donde se menciona el término en un primer párrafo, y en los posteriores se aclara su significado: '*Defining statement*. *This is called…*'; '*Heading*. *This is…*'; '*X*. *This is a...*'.

Los expositivos definitorios semi-formales sólo incluyen dos elementos, el término y las características distintivas, y esta clase incluye igualmente verbos conectivos como *contener*, *producir*, *proveer*, o frases verbales como *es/son usado*(*s*) *para*, *es/son caracterizado*(*s*) *para*, etc.

Por último, Pearson añade una clase de patrones definitorios denominados de *tipo diccionario*, los cuales incluyen contextos que normalmente se pueden encontrar en recursos terminológicos. Estos contextos están formados por patrones tipográficos que resaltan, por ejemplo, la presencia del término o la definición mediante

[4] Véase el apartado 2.6.

caracteres como las comillas o paréntesis, o bien resaltan el término, la definición o ambos en itálicas, negritas o subrayado.

Meyer, por su parte, denomina a los patrones definitorios como *patrones de conocimiento* y los clasifica en tres clases: *patrones léxicos*, *patrones gramaticales* y *patrones paralingüísticos* (Meyer 2001: 290).

Los patrones léxicos son el tipo más común e incluyen una o más unidades léxicas como *clasificado como*, *definido como*, *sirve para*, *designado como*, etc. Los patrones gramaticales se encuentran en un número reducido de atributos y relaciones. Por ejemplo el patrón NOMBRE + VERBO, que en inglés resulta productivo para indicar funciones de relación, o el patrón ADJETIVO + NOMBRE que igualmente resulta de provecho para encontrar atributos. Los patrones paralingüísticos son un tipo de patrón que no son gramaticales ni léxicos estrictamente. En este grupo, que coincide con las definiciones de tipo diccionario de Pearson, Meyer considera igualmente a los signos de puntuación, o la tipología del texto.

A partir de estas consideraciones y del estudio que realizamos de CDs en español, proponemos una tipología que denote principalmente la división entre elementos tipográficos y sintácticos. Esta tipología la ejemplificamos en la siguiente figura.

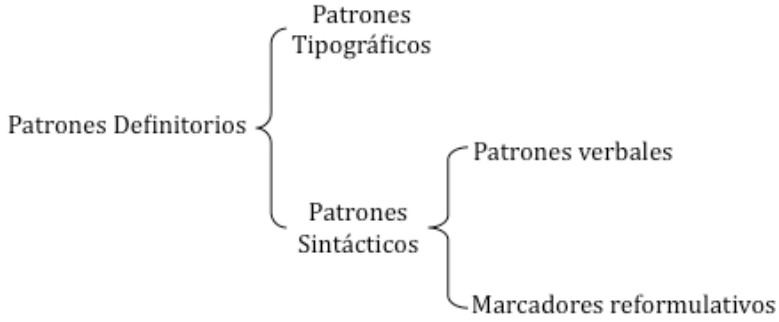

Figura 4.5. Tipología de patrones definitorios

Consideramos dos clases generales de patrones definitorios: los patrones tipográficos y los patrones sintácticos. Estos patrones no son excluyentes, es decir, pueden darse por separado o en conjunto. En el caso de los patrones sintácticos, y de acuerdo con el tipo de

elemento que se presente en el patrón, se pueden dividir en patrones verbales y marcadores reformulativos. En los siguientes apartados explicaremos cada uno de ellos.

4.2.2.1 Patrones tipográficos

La tipografía de un texto sirve por lo general como una ayuda visual para identificar fácilmente algún elemento importante y diferenciarlo del resto del texto común. Los términos son uno de estos elementos que tienden a ser frecuentemente resaltados. Si dichos términos incluyen una definición, es común además que ésta se encuentre también señalizada con algún elemento tipográfico o alguna tipografía específica.

Asimismo, en los textos especializados existen casos como las definiciones de tipo diccionario o los contextos con patrones paralingüísticos referidos por Pearson y Meyer, en que se define un término sin la necesidad de emplear algún verbo definitorio que funcione como conector.

- **Desastre**. *Perturbación de la actividad normal que ocasiona pérdidas o daños extensos o graves*.
- MITIGACION: Disminuir los efectos de los impactos de las calamidades.
- *Calamidad* ¶ Acontecimiento que puede impactar al sistema afectable y transformar su estado normal o deficiente en un estado de desastre.

En estos ejemplos los términos están resaltados en negritas, mayúsculas y cursivas. Observamos en los dos primeros casos que el término se une a la definición a partir de un signo de puntuación. En el tercer ejemplo la definición aparece después de un salto de párrafo, con lo cual se puede considerar que el término además de estar en cursivas, puede estar resaltado con el hecho de aparecer en un párrafo anterior a modo de título donde su presencia se hace más notoria.

En el estudio del corpus preliminar a la tesis se encontró que las tipografías textuales más recurrentes para resaltar los elementos constitutivos mínimos de los CD eran: cursivas, negritas,

subrayados, mayúsculas, encabezados, viñetas y paréntesis. En cuanto al uso de signos de puntuación en los casos en los que se elide el verbo definitorio, se encontró que los más usados eran dos puntos, punto y guión, o punto y seguido.

En resumen, la estructura de este tipo de patrones es la más simple y refiere a los tipos de definiciones que se pueden encontrar en diccionarios. Se utilizan ya sea para resaltar a los elementos constitutivos mínimos de los CDs o bien para conectar dichos elementos.

4.2.2.2 Patrones sintácticos

Por otro lado, la sintaxis a nivel de los CDs permite establecer normas comunes para identificar las estructuras recurrentes tanto de los elementos mínimos constitutivos como de los conectores que unen a estos dos elementos. Cuando dichos conectores sintácticos tienen como núcleo un verbo, consideramos al patrón como un *patrón verbal*. Por otro lado, se suelen emplear otro tipo de formas sintácticas cuya finalidad es establecer una reformulación de una idea o concepto, y que en el caso de los CDs se utilizan para esclarecer el significado de un término. Este tipo de estructuras corresponden a *marcadores reformulativos*. A continuación detallamos cada uno de ellos.

En los CDs se suelen utilizar verbos y formas verbales para unir a un término con su definición y referir atributos y características conceptuales de dicho término. Algunos de los verbos más comunes suelen considerarse como verbos *metalingüísticos*, por ejemplo *definir*, *entender* o *denominar*, los cuales se utilizan para referirse al propio lenguaje. No obstante, en los patrones verbales también pueden encontrarse verbos como *ser* o *considerar*, los cuales suelen emplearse con mayor frecuencia en otro tipo de situaciones comunicativas, no solamente definitorias.

Además, en la construcción de los patrones verbales pueden estar incluidas una serie de partículas gramaticales, por ejemplo el pronombre impersonal *se* en posición proclítica o enclítica en relación con el verbo definitorio, las preposiciones *a* o *por*, o bien el adverbio *como*. Algunas de las construcciones con estas partículas

podrían ser: *se entiende por*, *se denomina a*, *definirse como*, etc. A continuación presentamos algunos ejemplos:

- En este sentido, el estado de un sistema se define como una característica global que está determinada por un conjunto de valores en que se encuentran los parámetros relevantes para su funcionamiento en un momento dado.
- Se denomina "equipo de salud" a todo el personal del hospital que tiene una función directa o indirecta para el paciente.
- El tanque de almacenamiento es un recipiente en el cual se almacena el agua caliente para tenerla disponible a la hora que sea requerida su utilización.

En los ejemplos anteriores se introduce información definitoria a partir de los verbos *definir*, *denominar* y *ser*. Podemos observar, además, la ocurrencia del pronombre *se* en los casos en los que precede a *definir* y *denominar*, al igual que el adverbio *como* y la preposición *a* que forman los patrones *se define como* y *se denomina a*. En el tercer caso, el verbo *ser* va acompañado del determinante *un*, lo cual suele considerarse una estructura prototípica cuando este verbo se utiliza para definir un término.

En conclusión, los patrones verbales constituyen un nexo entre el término y la definición, y estructuralmente pueden estar formados por sintagma verbales que incluyen cierto tipo de partículas gramaticales. En el apartado 4.3 ahondaremos en la relación que establecen los patrones verbales con el contenido semántico de los CDs.

Al mismo nivel sintáctico existen otro tipo de conectores que no constan de un núcleo verbal, pero que igualmente sirven para conectar al término con su respectiva definición. Este tipo de conectores se denominan *marcadores reformulativos* y son estructuras sintácticas que se encuentran relacionadas con un proceso igualmente metalingüístico, que en el caso de los CDs sirve para referirse a los términos como elementos del propio lenguaje.

La reformulación, señala Bach (2005: 2), "es una operación de autorreflexión sobre la lengua, y una muestra clara de la función metacomunicativa del lenguaje." Además, añade, "la reformulación es un proceso de reinterpretación textual mediante el cual un locutor

determinado retoma algún elemento del discurso anterior para presentarlo de otra forma y con una función discursiva determinada […], garantiza la cohesión textual y, a su vez, facilita la progresión discursiva […] porque permite puntualizar el significado de algunos enunciados presentados anteriormente".

En el grupo de marcadores reformulativos encontramos estructuras como *por ejemplo*, *es decir*, *esto es*, *en otras palabras*, *dicho de otra manera*, etc. Algunos ejemplos los encontramos en el siguiente contexto:

- *El pronóstico de daños*, esto es, la cuantificación de la magnitud de las consecuencias o daños del fenómeno destructivo sobre el sistema afectable, conteniendo una relación de la cantidad de daños humanos, económicos, sociales y ecológicos que puede producir la calamidad.
- El índice secundario es a menudo un índice denso, es decir, contiene todos los valores posibles de la clave primaria.

En el primer ejemplo se define el término *pronóstico de daños* utilizando el marcador *esto es* como un conector con la definición. Por su parte, en el segundo ejemplo se presenta un proceso de reformulación para explicar que el término *índice secundario* implica que *contiene todos los valores posibles de clave primaria*.

En resumen, este tipo de patrones sintácticos conforman un proceso de reformulación en el que se explica el significado de un término a partir de estructuras sintácticas no verbales y que pueden ser igualmente útiles para desarrollar una metodología de extracción automática de CDs.

4.2.2.3 Patrones pragmáticos

Por último, en un CD se puede encontrar, además de la definición, otro tipo de información relevante para entender al término dentro del contexto en el cual aparece. Esta información está en relación con la forma en que se introduce el término en el texto especializado y que manifiesta explícitamente las condiciones de uso o de alcance de dicho término. Como señala Cabré:

> "Los términos están asociados a características gramaticales y pragmáticas. Las características pragmáticas describen los usos de los términos y los efectos derivados de estos usos: ámbitos temáticos, zonas geográficas u organismos en que se usan, nivel de especialidad de cada denominación, connotaciones asociadas al término situación en relación a su grado de normalización, frecuencia de uso, etc." (Cabré, 1999, p.139)

Este tipo de patrones los denominamos *patrones pragmáticos* (PPR) y los dividimos en tres clases generales[5]: patrones que corresponden al autor que propone la definición del término, patrones pragmáticos temporales y patrones pragmáticos instruccionales.

En los patrones pragmáticos de autor encontraremos patrones que hacen referencia directa al autor que propone el término. Estos patrones pueden ser sencillos, del tipo *Rosch* (nombre propio), o bien estructuras más complejas como: *los genetistas clásicos desde Mendel a Morgan*.

Los patrones pragmáticos temporales están en relación con la fecha de introducción o modificación del término, y ayudan por lo general a situar históricamente al término y su definición. Encontramos frases como *en 1889*, o bien estructuras más complejas como *a principios del siglo XX*.

Por último, los patrones pragmáticos instruccionales incluyen estructuras que aportan matices diferentes para entender el término: *de manera general*, *desde un punto de vista práctico*, etc. Se denominan instruccionales ya que presuponen una condición de uso del término, es decir, el autor que introduce el CD aclara, mediante estas estructuras, cómo se debe entender el término o cual es su alcance en un contexto determinado.

Algunos ejemplos de patrones pragmáticos los podemos ver en la siguiente tabla:

[5] Es importante aclarar que el estudio de los patrones pragmáticos no lo abordamos de manera directa en esta tesis; sin embargo, a partir de observaciones preliminares hemos distinguido la tipología que aquí detallamos.

Tipo	Ejemplo
Autor	Inicialmente, **Rosch** definió el prototipo como el ejemplar que mejor se reconoce, el más representativo y distintivo de una categoría […]
Temporal	Por ejemplo, la unidad de longitud – el metro - se definió **en 1889** como la longitud de una determinada barra de platino iridiado […]
Instruccional	**Desde el punto de vista genético**, el desarrollo puede definir se como «un proceso regulado de crecimiento y diferenciación resultante de […]

Tabla 4.8. Ejemplos de patrones pragmáticos (tomados del CTIULA (Bach *et al*. 1997))

Es de reconocer que los patrones pragmáticos pertenecen a un paradigma estructural amplio ya que su composición puede variar de acuerdo con formas estructurales o estilísticas utilizadas por cada autor. No obstante, encontramos patrones recurrentes compuestos por: adverbios y frases adverbiales (*usualmente*, *de manera general*), frases prepositivas (*desde el punto de vista genético*), palabras simples (*definición*, *concepto*, *término*), y estructuras formadas por nombres propios (*Rosca*, *El norteamericano Instituto Nacional de la Salud*).

En resumen, este tipo de patrones desempeñan un papel sintáctico importante en la composición de los patrones recurrentes. Cuando no existen patrones tipográficos, los PPR, junto con las PVD nos permiten identificar que existe un posible CD dentro del texto. Además, nos permiten diferenciar fragmentos textuales donde el significado del verbo, por sí solo, no nos ofrece la seguridad de estar funcionando como un nexo entre un término y una definición.

4.3 Tipología semántica de contextos definitorios

En el análisis de CDs hemos observado la importancia de los patrones en la introducción de contextos donde se define un término. Unos de los patrones sintácticos recurrentes son los patrones verbales donde se utiliza una serie de verbos específicos para ligar al término con información sobre sus características o atributos. Igualmente, los patrones verbales pueden introducir

información conceptual sobre el término a partir del contenido semántico implícito en los verbos.

A lo largo de nuestra investigación hemos ido delimitando una tipología semántica de CDs que está en relación precisamente con el tipo de información conceptual que en ellos se transmite (Sierra *et al*. 2003; Aguilar *et al*. 2006; Sierra *et al*. 2008). Con el fin de delimitar una metodología de extracción automática de CDs, propusimos una clasificación basada principalmente en un esquema de definición analítica.

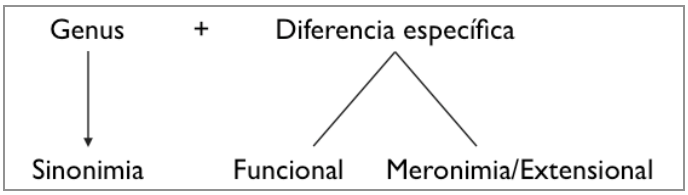

Figura 4.6. Tipología de información definitoria basada en un modelo de definición analítica

En el nivel superior de la figura se observa el tipo de información que se provee en una definición analítica: el *genus*, o género próximo, más la diferencia específica. A partir del genus se deriva un tipo de información definitoria de sinonimia, mientras que a partir de la diferencia específica se puede derivar información funcional o extensional.

Consideramos que el tipo de información sinonímica corresponde al mismo nivel del género próximo en una definición analítica, ya que a este nivel no se provee algún rasgo que distinga al término que se define de otros términos de su misma especie. En el caso de la información funcional y extensional, éstas se consideraron a nivel de la diferencia específica porque en ellas se proveen datos sobre atributos o características de los términos, más no se expresa el género al cual pertenecen.

De esta forma, establecimos cuatro grupos de CDs de acuerdo con su contenido semántico:

- CDs analíticos. En los cuales se presenta tanto el género próximo al cual pertenece el término, como información

específica que lo diferencia de otros elementos de su campo de conocimiento.
- CDs funcionales. En estos contextos no se incluye el género próximo del término pero se indica la función del término.
- CDs extensionales. Como en el caso anterior, en estos contextos tampoco se presenta el género próximo, sin embargo se enumeran características que conforman las partes de la entidad.
- CDs sinonímicos. En este tipo de contextos se presenta otro término que puede equivaler semánticamente al término definido.

Somos concientes de que esta clasificación sólo representa una parte básica del tipo de información conceptual que se puede presentar en los CDs. Cabe aclarar que esta clasificación tiene como principal cometido el delimitar grupos específicos para el desarrollo de reglas que nos permitan extraer automáticamente CDs asociados a un tipo de información específica. A continuación detallamos cada uno de los tipos antes mencionados.

4.3.1 Contextos definitorios analíticos

Los CDs analíticos constituyen los enunciados más completos, ya que en ellos se provee información característica sobre un término y además se presenta el género próximo al cual pertenece. En este grupo se encuentran patrones definitorios con verbos como *definir*, *entender*, *concebir* o *ser*. Algunos ejemplos de CDs analíticos son los siguientes:

- De manera general, un Operador Logístico (OL) es una firma que realiza prestaciones logísticas en servicio público que adapta a necesidades específicas de cada cliente.
- De acuerdo con el enfoque integral expuesto, el sistema de gestión se concibe como una organización, cuyo funcionamiento busca lograr ciertos objetivos a través de la operación de los diversos subsistemas interrelacionados que la componen.

En el primer CD se define el término *operador logístico* a partir del verbo *ser*, y se expresa que dicho término tiene como género

próximo *firma*, cuya diferencia específica es que *realiza prestaciones logísticas en servicio público*... etc. En el segundo ejemplo, el *sistema de gestión* tiene como género próximo *organización*, y se se caracteriza porque su *funcionamiento busca lograr ciertos objetivos a través de la operación de*… etc.

4.3.2 Contextos definitorios funcionales

Los CDs de tipo funcional aportan información de un término a partir de la descripción de su uso o aplicación en una situación determinada. Esto es, dependiendo de la combinación entre el verbo del patrón definitorio y ciertas partículas gramaticales, los CDs funcionales pueden asignar un valor funcional al objeto que en ellos se define, por ejemplo mediante un patrón como *se usa para*. Además, pueden expresar una localización donde se describe la función del objeto definido, por ejemplo mediante *se usa en*.

En este grupo se presentan patrones definitorios con verbos como *usar*, *utilizar, servir*, y combinaciones con partículas gramaticales como: *usar en*, *usar para*, *usar como*, etc.

- Los Sistemas Expertos se pueden utilizar para definir los requerimientos (entradas) de alguna aplicación o para la interpretación de resultados entregados por alguna aplicación
- Las escalas de Likert se usan habitualmente para cuantificar actitudes y conductas.
- Los motores neumáticos se usan en una variedad amplia de herramientas de mano.

De los ejemplos anteriores vemos las construcciones verbales *se pueden utilizar para*, *se usan para* y *se usan en*. Los dos primeros contextos corresponden a descripciones sobre la función específica de los términos, mientras que el tercer ejemplo es una descripción sobre una localización determinada de la función.

4.3.2 Contextos definitorios extensionales

Por su parte, los CDs extensionales aportan información de un término a partir de la enumeración de sus componentes. Estos

contextos suelen introducirse mediante verbos como *comprender, constar*, *constituir*, los cuales se encuentran normalmente en patrones definitorios como *comprende*, *constar de*, *estar constituido por*, etc. Algunos ejemplos de CDs extensionales son los siguientes:

- Un nucleótido está formado por una base nitrogenada (bien una purina o una pirimidina), un azúcar que es una pentosa (la ribosa o la desoxirribosa), y un grupo fosfato
- El terminal utilizado es el videoteléfono, que consta básicamente de una pantalla, cámara, teclado, micrófono, altavoz.

En estos dos ejemplos se enumeran una serie de componentes a partir de los patrones verbales *está formado por* y *consta de*. Por lo general, el contenido de la información definitoria suele ser estructuralmente una lista de frases nominales que a su vez corresponden a otros términos del domino del término que se define.

Cabe señalar que, en algunas ocasiones, las partículas gramaticales pueden llegar a modificar el contenido conceptual del verbo, por ejemplo en las siguientes construcciones hipotéticas con el verbo *consistir*: "el término X consiste **en** Y"; "el término X consiste **de** A, B, C…", donde el primer caso es una declaración sobre los atributos de X, frente al segundo caso en donde se enuncian las partes que conforman a X.

4.3.4 Contextos definitorios sinonímicos

Por último, en los CDs sinonímicos se establece una relación de similitud entre dos términos que se refieren a un mismo concepto[6].

- En la mujer, el conducto vaginal se llama también conducto de Nuck.

[6] Para un estudio detallado sobre variación denominativa y su relación con la sinonimia véase Bach, et. al. (2003).

En este tipo de contextos, se presentan dos términos al mismo nivel conceptual. En el ejemplo anterior, tanto el término *conducto vaginal* como *conducto de Nuck* se encuentran a un mismo nivel semántico y podrían utilizarse indistintamente.

Para cerrar este capítulo cabe señalar que los CDs sinonímicos utilizan patrones verbales como *llamado también*, *se denomina también*, *se conoce también como*, donde la finalidad del adverbio *también* es la de reforzar la función de similitud entre los dos términos propuestos.

5. ECODE: Extractor de Contextos Definitorios

En el capítulo anterior hemos expuesto un análisis de CDs en español, partiendo de la descripción del mismo concepto de contexto definitorio con un enfoque terminográfico. Asimismo, hemos visto cuáles son las ideas comunes en torno a la descripción del concepto de CD con fines de su extracción automática. Hicimos mención a las características distintivas de los patrones definitorios, describiendo los elementos constitutivos por los que están formados, así como ciertas características semánticas que introducen los distintos verbos que se emplean como conectores entre el término y la definición.

Tomando en cuenta lo expuesto hasta ahora, en este capítulo presentamos la metodología de extracción de CDs que hemos desarrollado a lo largo de nuestra investigación. El sistema que aquí describimos se denomina *ECODE* (Extractor de Contextos Definitorios). Es un sistema basado en reglas lingüísticas que incorpora diferentes procesos para la extracción de contextos donde se aporta información definitoria sobre un término. Comenzaremos por describir el corpus de pruebas utilizado en el desarrollo de nuestra metodología (5.1). Enseguida presentaremos una descripción general del algoritmo (5.2). Detallaremos el primer proceso correspondiente a la identificación de patrones verbales (5.3). Posteriormente explicaremos la metodología del siguiente proceso referente al filtro de contextos no relevantes (5.4). Describiremos el proceso de identificación de elementos constitutivos, es decir, términos y definiciones (5.5). Y finalmente, explicaremos un proceso de ranking de los resultados (5.6)

5.1 Corpus de pruebas

Para conformar un corpus de pruebas y elaborar la metodología del sistema aquí propuesto, utilizamos el Corpus Técnico del IULA en español, desarrollado por el Instituto Universitario de Lingüística Aplicada. Este corpus incluye documentos en las áreas de Derecho, Genoma, Economía, Medio Ambiente, Medicina, Informática y Lenguaje General (Bach *et al*. 1997).

El corpus está etiquetado morfosintácticamente con POS (Part of Speech) mediante el estándar EAGLES[1] para representar los distintos tipos de palabra y sus características específicas. Existen códigos para representar categorías mayores, por ejemplo verbo (V), nombre (N) y adverbio (D); códigos para representar rasgos recurrentes, como persona, número o género; y códigos particulares de las categorías mayores, como el tiempo verbal en el caso de los verbos.

Por ejemplo, *definir* sería representado mediante **VI----**, *definida* mediante **VC--SF**, y *definen* mediante **VDR1P-**. Observamos que cada caso incluye un código para la categoría mayor de verbo seguido de 5 posiciones que pueden estar ocupadas por distintos códigos. En algunos casos, las etiquetas pueden estar formadas por un número desigual de posiciones, para lo cual se utiliza el símbolo "-" con el fin de completar los lugares que puedan quedar vacíos[2].

Por otro lado, el corpus permite hacer consultas a través del sistema CQP (Corpus Query Processor)[3], por medio del cual se puede buscar una palabra o conjuntos de palabras específicas, formas gramaticales de una clase de palabra y/o conjunto de palabras, o bien las ocurrencias de uno o varios lemas específicos.

El CQP utiliza los siguientes operadores:

- **WORD**. Si buscamos la palabra "genoma" se hará mediante la ecuación: [word = "genoma"].
- **POS**. Si buscamos un nombre (N), propio (4), femenino (F), singular (S), se hará mediante la ecuación: [pos = "N4-FS"].
- **LEMMA**. Si buscamos las ocurrencias del término *gen* (*gen*, *genes*) se hará mediante la ecuación: [lema = "gen"].

Ahora bien, el corpus de pruebas lo conformamos específicamente a través de la extracción de los contextos donde apareciera uno de los términos de una lista obtenida del *Vocabulario Básico de Genoma*

[1] Página Web de EAGLES: http://www.ilc.cnr.it/EAGLES96/home.html
[2] Para la propuesta de marcaje gramatical del Corpus Técnico del IULA véase: http://www.iula.upf.es/corpus/etqfrmes.htm
[3] Para mayor referencia consultar: http://bwananet.iula.upf.edu/ajuda/cqpajuda3es.htm

Humano[4]. Este vocabulario cuenta con 163 términos y contiene tanto términos simples como *genoma* o *plásmido*, al igual que términos compuestos como *ingeniería genética* o *complejo de Golgi*.

Cada uno de los términos se buscó mediante el operador LEMMA del CQP. Algunos ejemplos de las ecuaciones de búsqueda se pueden observar en la siguiente tabla.

Ecuaciones de búsqueda
[lemma="genoma"]
[lemma="plásmido"]
[lemma="ingeniería"] [lemma="genética"]
[lemma="complejo"] [lemma="de"] [lemma="Golgi"]

Tabla 5.1. Ejemplos de ecuaciones de búsqueda para obtener el corpus de pruebas

Asimismo, cada contexto se recuperó con etiquetas POS. Un ejemplo de un contexto anotado es el siguiente:

<doc_codi m00862>: Las/AFP enfermedades/N5-FP monogénicas/JQ--FP son/VDR3P- producidas/VC--PF por/P la/AFS ##mutación/N5-FS## o/C deleción/N5-FS de/P un/J6--MS gen/N5-MS ./Z

Tabla 5.2. Ejemplo de contexto anotado con etiquetas POS

Observamos que la línea comienza con un código específico referente al documento y área donde se encontró el término (<doc_codi m00862>[5]), cada palabra se expresa seguida de su etiqueta POS, y el término se enmarca entre los símbolos "##".

Para obtener el corpus de pruebas realizamos además un preprocesamiento de los contextos obtenidos. Este preproceso constaba de dos pasos:

1. Eliminamos las ocurrencias pertenecientes al área del Lenguaje General, ya que nuestro interés radica en textos de dominio especializado.

[4] http://www.iula.upf.edu/rec/vbgenoma/esp/index.html.
[5] Donde "m" indica que pertenece al área de medicina.

2. Eliminamos los símbolos "##" que enmarcan los términos de búsqueda, con el fin de eliminar ocurrencias repetidas. Esto debido a que el sistema CQP considera una ocurrencia por cada vez que aparece el término dentro de una misma línea[6].

Así, nuestro corpus de pruebas estaba conformado por un total de 1,091,946 *tokens* (secuencias de caracteres entre espacios en blanco), en un total de 38,247 oraciones.

5.2 Descripción general del algoritmo

Como hemos visto a lo largo de esta investigación, la extracción automática de CDs es posible mediante la búsqueda de ocurrencias de patrones definitorios. Un extractor que obtenga contextos donde ocurren dichos patrones constituiría de entrada una herramienta útil para la adquisición de conocimiento definitorio. Sin embargo, el análisis manual de las ocurrencias extraídas supondría todavía un esfuerzo que podría simplificarse mediante un extractor que incluyera además un procesamiento automático de los resultados obtenidos, principalmente para excluir contextos no relevantes y clasificar los elementos constitutivos presentes en cada contexto.

De esta forma, el sistema que aquí proponemos es un sistema basado en reglas lingüísticas para la extracción de CDs a partir de patrones verbales[7]. Este sistema incluye un filtro de contextos no relevantes, es decir, contextos donde se encuentra un patrón verbal definitorio pero no se define un término; incluye también la identificación de los elementos constitutivos del CD, es decir el término y la definición; y realiza un ranking de resultados para determinar cuáles son los mejores contextos propuestos por el

[6] Por ejemplo el siguiente contexto era recuperado dos veces por la ocurrencia del término *genes* en diferentes posiciones:
<doc_codi m00784>: Junto/D a/P **##genes/N5-MP##** codificantes/JQ--6P de/P proteínas/N5-FP ,/Z existen/VDR3P- otros/EN--66 muchos/EF--MP genes/N5-MP que/RR---66…
<doc_codi m00784>: Junto/D a/P genes/N5-MP codificantes/JQ--6P de/P proteínas /N5-FP ,/Z existen/VDR3P- otros/EN--66 muchos/EF--MP **##genes/N5-MP##** que/RR---66…

[7] Tenemos contemplado como trabajo futuro la inclusión de otro tipo de patrones definitorios, ya mencionados en el capítulo anterior.

sistema. En la figura 5-1 podemos observar un panorama general de la arquitectura del ECODE.

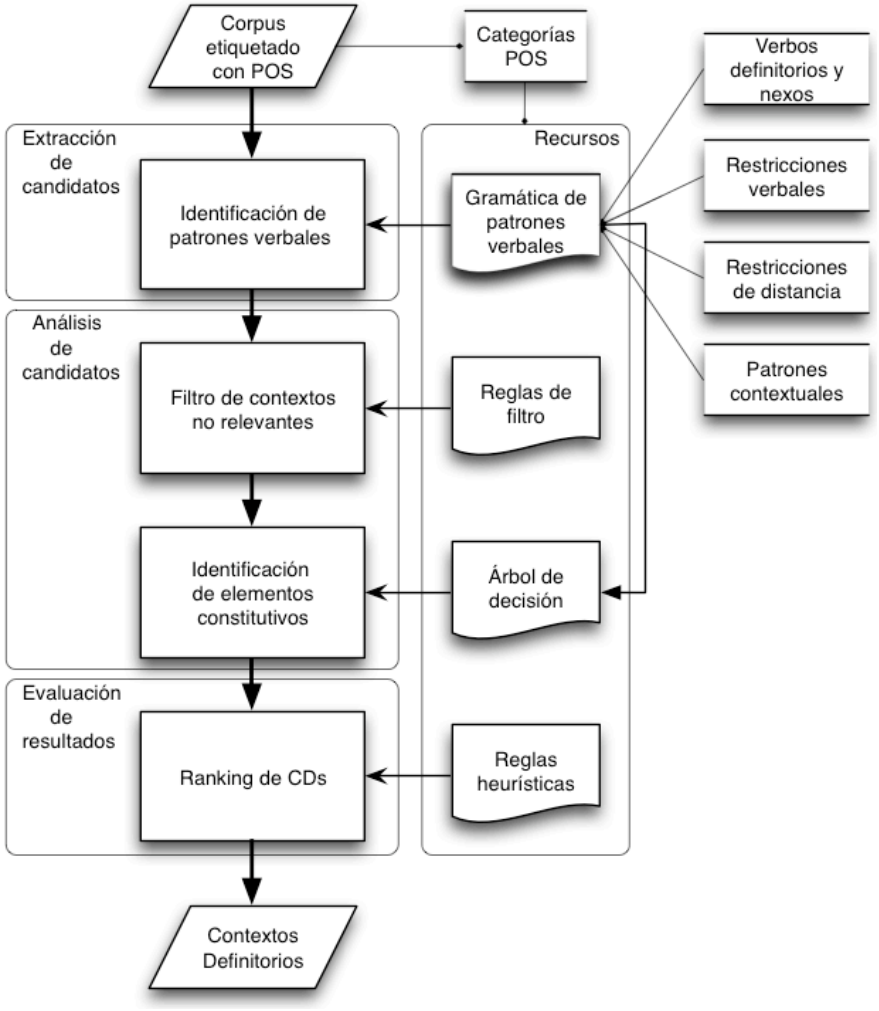

Figura 5.1. Panorama general de la arquitectura del ECODE

En la figura anterior, la entrada consiste en un corpus anotado con etiquetas POS del cual se utilizan ciertas categorías que serán necesarias en los recursos que conforman cada proceso principal. Podemos también observar tres procesos generales que consisten en a) la extracción de candidatos, b) el análisis de candidatos, y c) la evaluación de los resultados.

La extracción de candidatos es el proceso principal que se realiza a partir de una gramática de patrones verbales, donde se especifican una serie de parámetros: los verbos definitorios a buscar, los nexos que pueden acompañar a cada verbo, restricciones verbales referentes al tiempo y a la persona gramatical, así como la anotación del tipo de información definitoria que vehicula cada patrón verbal. También incluye restricciones de distancia entre el verbo y su nexo, así como las posiciones en las que puede aparecer el término en relación con el verbo de cada patrón.

Una vez obtenidos los candidatos, el análisis de éstos incluye dos procesos principales: el filtro de contextos no relevantes y la identificación de los elementos constitutivos. El filtro de contextos no relevantes consiste en una serie de reglas lingüísticas y contextuales para determinar aquellos casos en los que es probable que un patrón verbal no esté introduciendo información definitoria. Por su parte, en la identificación de los elementos constitutivos se utiliza un árbol de decisión que recurre igualmente a la gramática de patrones verbales para tratar de identificar cuál es el término y la definición en los candidatos restantes que no fueron filtrados.

Por último, la evaluación de los resultados está formada por un proceso principal que utiliza una serie de reglas heurísticas asociadas a la estructura del término y la definición para determinar cuáles son los mejores resultados propuestos por el sistema.

Finalmente, la salida consiste en una lista de CDs asignados a alguno de los grupos que explicamos en el capítulo anterior: *CDs analíticos*, *CDs funcionales*, *CDs extensionales* o *CDs sinonímicos*, y reorganizados según la probabilidad de que sean en mayor o menor medida mejores CDs.

El algoritmo se implementó en PERL a partir de una serie de módulos propios quedescribimos en la siguiente tabla[8]:

[8] El sistema, junto con instrucciones detalladas para su ejecución, se puede consultar en la carpeta *ecode* del disco anexo.

Módulo		Descripción
1	vds.pm	**Etiquetamiento de verbos definitorios.** Este módulo etiqueta todas las ocurrencias de los verbos definitorios partiendo de la raíz expresada en la gramática de patrones verbales.
2	pvds.pm	**Etiquetamiento de patrones verbales**. Este módulo se basa igualmente en la gramática antes descrita para etiquetar los patrones verbales de los verbos definitorios anotados previamente.
3	td.pm	**Etiquetamiento de tipo de definición**. Asigna el tipo de definición correspondiente a cada contexto de acuerdo con la información provista en la gramática de patrones verbales.
4	filtro.pm	**Filtro de contextos no relevantes**. Este módulo se basa en las reglas de filtro para excluir aquellos contextos que no contienen información definitoria.
5	arbol.pm	**Árbol de decisión**. En este módulo se implementan unas series de reglas para decidir qué parte del contexto pertenece a cada elemento constitutivo de los CDs. Como resultado etiqueta el término y la definición.
6	retagging.pm	**Reetiquetamiento**. Aplica ciertas reglas para reetiquetar lo anotado en el módulo anterior, con el fin de corregir posibles problemas.
7	rankingTD.pm	**Ranking de término y definición**. Este módulo realiza un ranking a partir de reglas heurísticas específicas para la estructura del término y la definición. Como resultado anota cada elemento constitutivo con un valor.
8	rankingG.pm	**Ranking global**. En este módulo, los valores anotados previamente se combinan para generar un ranking global por cada CD.
9	final.pm	**Escritura de resultados**. Este último módulo genera una lista final de CDs organizados de acuerdo con los valores del ranking global y explicitando los elementos constitutivos, además de una lista con los contextos excluidos como contextos no relevantes.

Tabla 5.3. Módulos específicos del ECODE

De acuerdo con la figura 5.1, los tres primero módulos corresponden a la identificación de patrones verbales; el módulo 4 corresponde al filtro de contextos no relevantes; los módulos 5 y 6 son parte del proceso de identificación de elementos constitutivos; los módulos 7 y 8 pertenecen al ranking de CDs; mientras que el módulo 9 escribe la salida final de los resultados.

En los siguientes apartados nos enfocaremos a describir cada uno de los procesos principales del sistema.

5.3 Identificación de patrones verbales

Como podemos observar en la figura 5.2, el primer proceso para la extracción de CDs consiste en la identificación de patrones verbales a partir de una gramática desarrollada para este fin. En este proceso, la entrada es un corpus anotado con POS, del cual se utilizan ciertas categorías específicas para formar variables que serán indispensables en el funcionamiento del sistema. La gramática de patrones verbales consiste en una serie de datos incorporados manualmente donde se establecen diferentes parámetros relativos a cada verbo definitorio. La salida de este primer proceso está formada por una lista de contextos con patrones verbales anotados.

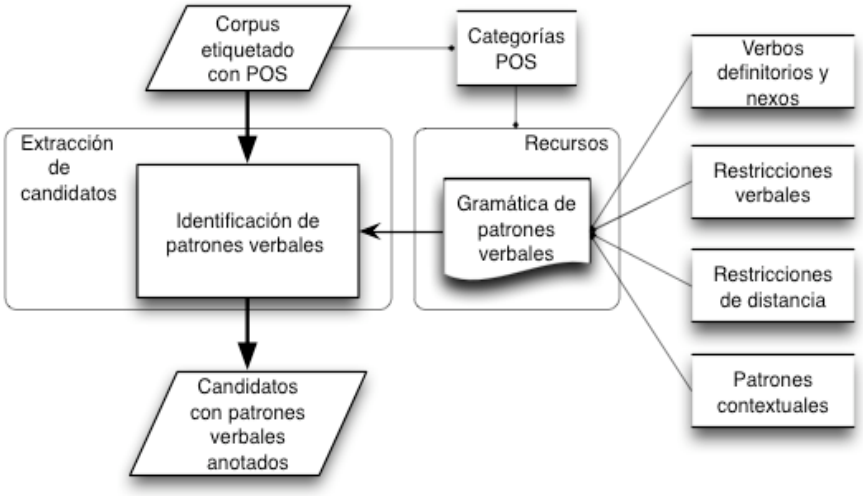

Figura 5.2. Identificación de patrones verbales del ECODE

En el caso de las categorías POS, las expresiones regulares que conforman las variables que necesita el sistema las encontramos en la tabla 5.4.

Variable	Expresión regular	Ejemplos
Generales		
$token	[^ /]*	gen
$tag	/[A-Z][^]*	/N5-MS
$palabra	$token.$tag;	gen/ N5-MS
$inicioLinea	<doc_codi [^]*?>:	<doc_codi m00449>:
Clases de palabras		
$adj	$token./J[^]*	rojo/JQ--MS
$adv	$token./D[^]*	completamente/D
$conj	$token./C	y/C
$det	$token./(A|E|J(N|6))[^]*	los/AMP
$prep	$token./P	de/P
$pron	$token./R[^]*	se/REE6366
$signo	$token./Z	./Z
$sus	$token./N[^]*	cromosoma/N5-MS
$vbo	$token./V[^]*	entender/VI----
Tipos de verbos		
$vTag	/V[^G][^]*	/VDR3P-
$verboCon	$token./V(D|J|R)[^]*	define/VDR3S-
$verboInf	$token./VI[^]*	definir/VI----
$verboPar	$token./VC[^]*	definidos/VC--PM
$verboGer	$token./VG[^]*	entendiendo/VG----

Tabla 5.4. Expresiones regulares de las variables con etiquetas POS del Corpus Técnico del IULA

En el grupo de las variables generales se especifican expresiones regulares para:

- $token: cualquier carácter excepto un espacio en blanco o el símbolo "/" que expresa el comienzo de las etiquetas POS.
- $tag: en el caso específico del Corpus Técnico del IULA, una etiqueta equivale a una diagonal más cualquier letra en mayúsculas seguida de cualquier carácter excepto un espacio en blanco. Por ejemplo el *gen* se expresa con: gen/N5-MS.

- $palabra: cualquier combinación de *token* y *tag*. En este y otros casos la concatenación se expresa mediante un punto.
- $inicioLinea: se usa para determinar dónde comienza cada línea del corpus, en este caso comienza con <*doc_codi* seguido de cualquier carácter excepto un espacio en blanco hasta el siguiente >:. Cabe señalar que cada palabra se considera como lo que hay entre espacios en blancos, por lo que al inicio y al final de cada línea se deberá tener un espacio vacío, independientemente de cuál sea el inicio específico en el corpus.

En el grupo de las clases de palabras se encuentran las siguientes expresiones, todas ellas formadas por un token más una diagonal y una letra mayúscula específica. En algunos casos se añade también la secuencia de cualquier carácter excepto un espacio en blanco:

- $adj: Adjetivo, código J.
- $adv: Adverbio, código D.
- $conj: Conjunción, código C.
- $det: Determinante, código (A|E|J|(N|6)).
- $prep: Preposición, código P.
- $pron: Pronombre, código R.
- $signo: Signo, código Z.
- $sus: Sustantivo, código N.
- $vbo: Verbo, código V.

Por último, el grupo de tipos de verbos contiene variables para cada forma verbal: verbos en forma conjugada, en infinitivo, en participio o en gerundio:

- $verboCon: Verbo en forma conjugada, código V(D|J|R).
- $verboInf: Verbo en infinitivo, código VI.
- $verboPar: Verbo en participio, código VC.
- $verboGer: Verbo en gerundio, código VG.

5.3.1 Verbos definitorios, nexos y restricciones verbales

En cuanto a la gramática de patrones verbales, ya hemos señalado que dicho elemento constituye un paradigma donde se establecen datos específicos relativos a cada verbo definitorio. Esto permite la

flexibilidad de añadir nuevos verbos para conformar así patrones que no se hayan contemplado previamente, o que puedan considerarse importantes de acuerdo con un escenario de aplicación específico, por ejemplo para incluir otro tipo de información definitoria no contemplada previamente.

El primer dato que se debe incorporar manualmente a la gramática de patrones verbales es una lista de verbos definitorios junto con los nexos que pueden acompañarlos y con las restricciones verbales referentes al tiempo y la persona gramatical para cada uno de ellos. En esta investigación delimitamos un grupo de 29 verbos asociados a distintos tipos de información. El paradigma de verbos definitorios lo hemos conformado a partir de una serie de experimentos propios realizados a nivel de tesis de licenciatura y de proyecto de tesis de doctorado (Alarcón 2003; Alarcón 2006); de experimentos y trabajos desarrollados en conjunto con colegas en el mismo ámbito de la extracción de CDs (Sierra *et al*. 2003; Aguilar *et al*. 2006; Sierra *et al.* 2008); y de otra investigación a nivel doctoral enfocada en una descripción lingüística de definiciones en CDs (Aguilar, 2009).

Como podemos ver en la tabla 5.5, en algunos casos un mismo verbo puede estar asociado a distintos tipos de información definitoria, dependiendo principalmente del nexo que lo acompañe. En esta misma tabla presentamos la raíz con la que se buscó cada verbo, lo cual está en relación con las primeras restricciones verbales establecidas para esta etapa. En el caso de la columna *nexo*, el símbolo ∅ expresa la ausencia de éste.

En este modelo observamos que los 29 verbos conforman un total de 33 combinaciones distintas dependiendo del tipo de nexo al que se une cada verbo. Los verbos que se repiten con nexos distintos y que pueden expresar diferentes tipos de información definitoria son:

1. Conocer (como, también)
2. Denominar (∅, como, también)
3. Llamar (∅, como, también)
4. Nombrar (∅, como, también)

Tipo	Lema	Raíz	Nexo
Analítico	ser	(es|son)	determinante
	caracterizar	caracteriz	como, por
	concebir	conc(e|i)b	como
	considerar	consider	como
	describir	describ	como
	definir	defin	como
	entender	ent(ie|e)nd	como
	conocer	conoc	como
	denominar	denomin	∅, como
	llamar	llam	∅, como
	nombrar	nombr	∅, como
Extensional	comprender	comprend	∅
	contener	cont(ien|en|uv)	∅
	incluir	inclu(í|i|y)	∅
	integrar	integr	∅
	constar	const(a|e|ó)	de
	contar	c(ue|o)nt(a|e|á|é|ó)	con
	formar	form	de, por
	componer	comp(on|us|uest)	de, por
	constituir	constit	de, por
Funcional	permitir	permit	∅
	encargar	encarg	de
	consistir	consist	en
	funcionar	funcion	como, para
	ocupar	ocup	como, para
	servir	s(i|e)rv	como, en, para
	usar	us	como, en, para
	emplear	emple	como, en, para
	utilizar	util	como, en, para
Sinonímico	conocer	conoc	también
	denominar	denomin	también
	llamar	llam	también
	nombrar	nombr	también

Tabla 5.5. Modelo inicial de verbos definitorios, raíces y nexos en la gramática de patrones verbales del ECODE

Los cuatro verbos que se repiten suelen emplearse en definiciones analíticas, con o sin nexo, o en definiciones sinonímicas si el nexo es *también*. A continuación presentamos algunos ejemplos:

Patrón	Contexto definitorio
llamar	El tipo de teorías que reemplazaron a los modelos instructivos **se llama** teoría selectiva porque se cree que un antígeno selecciona al anticuerpo apropiado de reserva preexistente de moléculas de anticuerpo.
llamar también	En el hombre, las moléculas MHC **se llaman también** antígenos leucocitarios humanos, o moléculas HLA.

Tabla 5.6. Diferentes tipos de información definitoria introducida por los mismos verbos

En estos ejemplos los dos primeros casos introducen información definitoria a través del verbo *llamar*. No obstante, el primero de ellos constituye una definición analítica mientras que en el segundo, al añadirse el adverbio *también*, se expresa entonces un sinónimo.

Por otro lado, cada verbo del modelo es buscado con una raíz específica. En otras investigaciones (Alarcón y Sierra 2003; Alarcón *et al*. 2008) hemos notado que el tipo de restricción verbal que se impone a esta raíz puede contribuir para filtrar contextos no relevantes. Ciertos verbos como *definir* o *denominar* suelen recuperar información definitoria en una mayor variedad de tiempos, formas verbales o personas gramaticales; mientras que otros verbos como *permitir* suelen ocurrir en determinadas formas verbales (por ejemplo participio) en una mayor cantidad de contextos donde no se expresa información definitoria.

Patrón	Contexto
contar con	Cómo se va a regular la aplicación de estudios de escrutinio conforme **contemos con** el conocimiento de dichos genes?
definir como	La radiación provoca mutaciones , que **definimos** antes **como** cambios en la secuencia de las bases de l ADN.

Tabla 5.7. Ejemplos de CDs y contextos no relevantes

En estos ejemplos se encuentran dos verbos en primera persona de plural. El primero de ellos, *contar*, claramente introduce otro tipo de información que no es definitoria; en cambio, el verbo *definir* en esta misma persona gramatical es una forma recurrente que se utiliza para introducir la definición de un término.

El modelo presentado en la tabla 5.5 incluye un conjunto de raíces verbales en sus restricciones más amplias. Es decir, para la mayoría de los verbos se considera su raíz única, sin delimitar el tiempo, la forma verbal o la persona gramatical. La única excepción se presenta en la raíz del verbo *ser*, que se usa únicamente en tercera persona singular o plural del tiempo presente en forma conjugada. Esto se consideró tomando en cuenta la gran cantidad de oraciones en las que ocurre una secuencia como *ser* + *determinante*. En el próximo capítulo presentaremos y evaluaremos modelos distintos relacionados con una mayor especificación para cada raíz verbal.

Finalmente, otro dato importante a incorporar en la gramática es que cada raíz puede a su vez recuperar otras palabras no relevantes. Si consideramos las raíces de los verbos tal y como las presentamos en el modelo de la tabla 5.5, entonces deberían considerarse las siguientes excepciones:

Lema	**Raíz**	**Excepción**	**Ejemplo**
Constar	*const(a|e|ó)*	*constat*	constatar
Formar	form	form(u|al)	formalizar
Usar	us	usu	usurpar

Tabla 5.8. Modelo de excepciones para raíces en la gramática de patrones verbales del ECODE

Por ejemplo, la raíz *consta* puede recuperar el verbo *constatar*; la raíz *form* podría reconocer también las conjugaciones del verbo *formalizar*; y el lema *usu* se encuentra asimismo en el verbo *usurpar*. En estos casos las excepciones se enmarcan dentro de las etiquetas *<raizEX></raizEX>*. Debe considerarse además que cada inclusión de una nueva raíz para un verbo definitorio no contemplado en el modelo anterior supone la comprobación manual de que dicha raíz no recupera casos no deseados.

5.3.2 Patrones contextuales

El siguiente dato que debe incorporarse a la gramática de patrones verbales corresponde a lo que aquí denominamos *patrones contextuales*. Con el fin de identificar automáticamente los elementos constitutivos de los CDs, debe tomarse en cuenta que términos y definiciones suelen seguir ciertos patrones de posición en relación con el verbo que los introduce.

Para explicar lo anterior, tomemos el caso de los siguientes patrones del verbo *definir*: *T se define como D*, *D es definido como T*, *Se define T como D*. Observamos que el término puede aparecer tanto en posición izquierda como derecha, al igual que entre el verbo definitorio y el nexo. Otros verbos que siguen estos patrones son *entender* y *conocer*. Por su parte, en contextos con verbos como *permitir*, el término aparecerá en posición izquierda: *T permite D*. En otras ocasiones, y dependiendo del nexo con que ocurra, un mismo verbo permite diferentes combinaciones de posiciones; por ejemplo, el verbo *denominar* sin nexo permitiría únicamente que el término apareciera en posición derecha, *D se denomina T* (si excluimos la secuencia *T se denomina a D*, pensando en el uso de la preposición *a* como nexo), pero si incluyera el nexo *como* entonces el término y la definición podrían estar en la posición derecha: *se denomina como T a D*. En la siguiente tabla podemos ver algunos ejemplos de patrones contextuales.

Patrón	**Contexto definitorio**
T + PVD + D	**<t>**La COMT**</t>** **<pvd><vd>**es**</vd>** **<nx>**una **</nx></pvd>** **<d>**enzima de distribución amplia, presente tanto en tejidos neuronales como en los no neuronales.**</d>**
VD + T + NX	**<pvd>**Se ha **<vd>**definido**</vd>** **<t>**el genotipo **</t>** **<nx>**como**<nx></pvd>** **<d>**la constitución genética del individuo en un locus. **</d>**
PVD + T + D	**<pvd>**Se denomina**</pvd>** **<t>**digestión**</t>** **<d>** al proceso por el cual las moléculas ingeridas son fraccionadas en otras más pequeñas mediante reacciones catalizadas por enzimas , bien en la luz o bien en la superficie orientada hacia la luz de l tracto GI.**</d>**

Tabla 5.9. Ejemplos de patrones contextuales

En la tabla anterior, el término se expresa entre las etiquetas <t></t>, la definición entre <d></d>, el patrón verbal entre <pvd></pvd>, el verbo definitorio entre <vd></vd>, y el nexo entre <nx></nx>. En el primer caso el término aparece a la izquierda del verbo *ser*. El segundo caso incluye el término entre el verbo y el nexo *como*. Por último, en el tercer ejemplo el término se encuentra a la derecha del verbo *denominar*.

Tomando en cuenta lo anterior, definimos 3 reglas contextuales simples:

1. Si el término puede aparecer a la izquierda del patrón verbal, entonces se dirá que su patrón contextual incluye la posición IZQUIERDA.
2. De la misma manera, si el término puede aparecer a la derecha del patrón verbal, entonces su patrón contextual incluye la posición DERECHA.
3. Por último, si el patrón verbal permite que el término aparezca entre el verbo definitorio y el nexo, entonces su patrón contextual incluye la posición NEXO

En el modelo de verbos definitorios presentado en la tabla 5.5, la mayoría de los patrones generados a partir de dichos verbos pueden incluir al término en posición izquierda, excepto los patrones para las definiciones analíticas con los verbos *denominar*, *llamar* y *nombrar*, así como en las definiciones sinonímicas, donde decidimos delimitar al término a la posición derecha (pensando en la relación tema-rema, donde el tema és el término 2, la información conocida, y el rema és el término 1, la información nueva), a pesar de que en este tipo de definiciones lo que ocurre realmente es una equivalencia entre dos términos. Por su parte, algunos patrones permiten la ocurrencia del término en la posición de nexo, pero no en la posición derecha. Tal es el caso, por ejemplo, de los verbos *usar*, *emplear* y *utilizar* que raramente introducirían secuencias como *se emplea como T D*. Finalmente, entre los verbos que permiten la ocurrencia del término en posición derecha tenemos a *definir, entender* y *conocer*.

Una representación de lo aclarado en las líneas anteriores la podemos encontrar en la siguiente tabla:

Tipo	Lema	Nexo	Patrón Contextual[9]
Analítico	ser	determinante	I
	caracterizar	como, por	I
	concebir	como	I N
	considerar	como	I N
	describir	como	I N
	definir	como	I N D
	entender	como	I N D
	conocer	como	I N D
	denominar	∅, como	D
	llamar	∅, como	D
	nombrar	∅, como	D
Extensional	comprender	∅	I
	contener	∅	I
	incluir	∅	I
	integrar	∅	I
	constar	de	I
	contar	con	I
	formar	de, por	I
	componer	de, por	I
	constituir	de, por	I
Funcional	permitir	∅	I
	encargar	de	I
	consistir	en	I
	funcionar	como, para	I
	ocupar	como, para	I
	servir	como, en, para	I
	usar	como, en, para	I N
	emplear	como, en, para	I N
	utilizar	como, en, para	I N
Sinonímico	conocer	también	D
	denominar	también	D
	llamar	también	D
	nombrar	también	D

Tabla 5.10. Modelo inicial de patrones contextuales en la gramática de patrones verbales del ECODE

[9] I = izquierda, D = derecha y N = nexo.

En resumen, el hecho de incorporar patrones contextuales en la gramática de patrones verbales responde a la necesidad de delimitar las posiciones en las que podrían aparecer el término y la definición respecto al verbo definitorio del contexto donde aparece. Este tipo de patrones son una base importante para el algoritmo, ya que permiten clasificar los elementos constitutivos de los CDs.

5.3.3 Restricciones de distancia

Una vez especificado el paradigma de verbos definitorios con sus respectivas raíces y nexos, así como el tipo de patrones contextuales que pueden seguir, los siguientes datos que debemos incorporar a la gramática son las restricciones de distancia que puede haber entre el verbo definitorio y su nexo.

Ya hemos señalado que en cada patrón verbal se debe considerar la posible inclusión de diferentes nexos que acompañen al verbo. En algunas ocasiones, dicho verbo puede estar acompañado o no de diferentes nexos para expresar definiciones de uno o varios tipos. Ahora bien, tomando en cuenta los patrones contextuales antes descritos, podemos darnos cuenta de que, en ocasiones, entre el verbo y el nexo puede aparecer otro tipo de información no prevista.

Lo anterior nos ha llevado a considerar la inclusión de una ventana de palabras posibles entre el verbo definitorio y el nexo, pensando que en dicha ventana puede aparecer desde un simple adverbio, hasta términos compuestos acompañados asimismo de adverbios o frases adverbiales. En la siguiente tabla podemos ver algunos ejemplos de lo anterior.

Tamaño	Contexto definitorio
1 palabra	Los vectores lambda **se utilizan** ampliamente **para** la construcción de bibliotecas genómicas.
3 palabras	**Se define** una librería genómica **como** un conjunto de clones en el que está representado todo el genoma de un organismo.
8 palabras	En 1977, Oshimura et al **describieron** las deleciones del brazo largo del cromosoma 6 **como** una anomalía recurrente en leucemias.

Tabla 5.11. Ejemplos de distancias entre verbos definitorios y sus nexos

El primer caso incluye un adverbio en la ventana entre el verbo *utilizar* y el nexo *para*. En el segundo ejemplo se introduce el término *una librería genómica*. Finalmente, el tercer caso es un ejemplo de un término compuesto estructuralmente más largo: *las deleciones del brazo largo del cromosoma 6*.

Debe considerarse que en estos dos ejemplos podría además incluirse un adverbio, frase adverbial o frase prepositiva que especifique el alcance del término (recuérdese los patrones pragmáticos de los que hablamos en el capítulo anterior), con lo cual la ventana entre el verbo y el nexo se ampliaría considerablemente.

Por otro lado, debe tenerse en cuenta que el tamaño de la ventana que se establezca puede a su vez ser una causa para recuperar ruido. Para ejemplificar esto tomemos los siguientes casos.

[...] cuyo número se sitúa entre 3 mil y 3 mil 500 millones, y que contienen la información genética que **define** al individuo, todas sus características internas y externas, así **como** su tendencia congénita a desarrollar ciertas enfermedades.
La clasificación de las distrofias musculares ha ido evolucionando con el tiempo: desde finales de l siglo pasado y hasta los años cuarenta, las descripciones anatomoclínicas **definían** los criterios de clasificación; en una segunda etapa, los distintos patrones de herencia se contemplaron **como** parámetros atener [...]

Tabla 5.12. Ejemplos de distancias sin restricciones entre verbos Definitorios y sus nexos

Los ejemplos anteriores se encontraron a partir de delimitar una ventana de cualquier número de palabras entre el verbo *definir* y el adverbio *como*. Observamos que la información que contienen estas ventanas claramente no corresponde a un término o a un patrón pragmático que nos pueda aportar información relevante sobre un término, sino que recuperan ruido que puede afectar en el rendimiento final del sistema. No obstante, cabe aclarar que en estas ventanas también podemos encontrar ciertas claves léxico-sintácticas que nos ayuden a delimitar los casos en los que los nexos no están funcionando como conectores entre el verbo definitorio y la definición. Esto lo veremos con detalle en el apartado 5.4

En síntesis, el tamaño de la ventana entre el verbo definitorio y el nexo es un dato importante que debe incorporarse a la gramática de patrones verbales, con el fin de considerar aquellas ocurrencias de patrones en las que, tanto el término como otro tipo de secuencias, pueden aparecer entre el verbo definitorio y el nexo. En esta decisión es importante notar si el verbo definitorio permite la inclusión del término en la posición de nexo, a partir de los datos introducidos en los patrones contextuales, con lo cual se debería permitir una distancia mayor para tratar de encontrar términos cuyo tamaño estructural sea mayor a la media.

De esta forma, en esta primera etapa decidimos establecer unas distancias no restrictivas para observar el tipo de datos que podrían recuperarse. El modelo inicial de la gramática de patroes verbales consideramos cualquier número de palabras entre el verbo y su nexo, excepto para los lemas *ser*, de los patrones analíticos, y *conocer*, *denominar*, *llamar* y *nombrar*, en el caso de los patrones sinonímicos. Respecto al verbo *ser*, decidimos no permitir ninguna palabra entre éste y el determinante, con el fin de tratar de disminuir la cantidad de ruido que puede recuperar dicho patrón. En los patrones sinonímicos la distancia se limitó a cero, pensando que es improbable que pueda aparecer alguna otra palabra entre el verbo definitorio y el nexo.

5.3.4 Gramática de patrones verbales

Hasta ahora hemos abordado una descripción de los datos que deben incorporarse en la gramática de patrones verbales con el fin de identificar una lista de primeros candidatos a CDs. Hemos observado que se deben incluir datos específicos sobe los verbos definitorios, como sus raíces, los nexos que los pueden acompañar, la distancia entre dichos verbos y sus nexos, así como información sobre los lugares en los que puede aparecer el término con respecto al verbo definitorio.

Ahora bien, en este apartado describiremos algunos ejemplos de la implementación de dicha gramática para la identificación automática de patrones verbales. En primer lugar comenzaremos

por resumir el modelo inicial de patrones seguidos para la extracción de candidatos en el corpus de pruebas.

	Lema	Raíz	Dist	Nexo	PC
A	ser	(es|son)	0	determinante	I
	caracterizar	caracteriz	-	como, por	I
	concebir	conc(e|i)b	-	como	I N
	considerar	consider	-	como	I N
	describir	describ	-	como	I N
	definir	defin	-	como	I N D
	entender	ent(ie|e)nd	-	como	I N D
	conocer	conoc	-	como	I N D
	denominar	denomin	-	∅, como	D
	llamar	llam	-	∅, como	D
	nombrar	nombr	-	∅, como	D
E	comprender	comprend	0	∅	I
	contener	cont(ien|en|uv)	0	∅	I
	incluir	inclu(í|i|y)	0	∅	I
	integrar	integr	0	∅	I
	constar	const(a|e|ó)	-	de	I
	contar	c(ue|o)nt(a|e|á|é|ó)	-	con	I
	formar	form	-	de, por	I
	componer	comp(on|us|uest)	-	de, por	I
	constituir	constit	-	de, por	I
F	permitir	permit	0	∅	I
	encargar	encarg	-	de	I
	consistir	consist	-	en	I
	funcionar	funcion	-	como, para	I
	ocupar	ocup	-	como, para	I
	servir	s(i|e)rv	-	como, en, para	I
	usar	us	-	como, en, para	I N
	emplear	emple	-	como, en, para	I N
	utilizar	util	-	como, en, para	I N
S	conocer	conoc	0	también	D
	denominar	denomin	0	también	D
	llamar	llam	0	también	D
	nombrar	nombr	0	también	D

Tabla 5.13. Modelo de la gramática de patrones verbales del ECODE

A partir de este modelo, escribimos entonces las reglas de la gramática con las siguientes etiquetas:

- <td></td>: aquí se especifica el tipo de definición.
- <lm></lm>: en este caso se declara el lema del verbo definitorio.
- <raiz></raiz>: contiene la raíz que se buscará.
- <dist></dist>: equivale a un valor numérico que indica la distancia entre el verbo definitorio y su nexo. Se escribe la palabra *any* si la distancia es igual a cualquier número de palabras hasta el primer nexo.
- <nx></nx>: contiene el nexo o los nexos que pueden acompañar al verbo definitorio.
- <lt></lt>: en esta etiqueta se expresa el lugar que puede ocupar el término. Los valores permitidos son I, N o D, referentes a izquierda, nexo y derecha.

En la siguiente tabla se muestran un par de ejemplos de cómo queda finalmente la gramática[10]:

Reglas de la gramática de patrones verbales
<td>analítica</td> <lm>conocer</lm> <raiz>conoc</raiz> <dist>any</dist> <nx>como</nx> <lt>IND</lt>
<td>extensional</td> <lm>formar</lm> <raiz>form</raiz> <dist>5</dist> <nx>de|por</nx> <lt>I</lt>

Tabla 5.14. Ejemplos de notación de la gramática de patrones verbales del ECODE

Como mencionamos en la tabla 5.3, el primer proceso para la implementación del algoritmo consiste en tres módulos donde el primero de ellos anota los verbos definitorios a partir de las raíces expuestas en la gramática de patrones verbales. El segundo módulo anota, por cada uno de los verbos identificados previamente, los casos que conforman algún patrón verbal, igualmente a partir de los datos de la gramática. Por último, el tercer módulo analiza cada línea donde aparece un patrón verbal para anotar el tipo de información definitoria al que está ligado.

[10] La gramática de patrones verbales se puede consultar en la carpeta *ecode/gramaticas/02_gramPVDs.pm* del disco anexo.

En el caso de los verbos definitorios, éstos se anotan con las etiquetas *<vd lema= ".*"></vd>*. Algunos ejemplos de verbos etiquetados, de acuerdo con el modelo inicial de la gramática expuesta en la tabla anterior, son:

- <vd lema="consistir">consiste</vd>
- <vd lema="conocer">conocer</vd>
- <vd lema="denominar">denominaremos</vd>
- <vd lema="ocupar">ocupaban</vd>
- <vd lema="servir">Sirva</vd>

Ahora bien, una vez etiquetados los verbos definitorios, el primer módulo selecciona las líneas que contienen al menos uno de ellos, y sobre estas líneas se corre el segundo módulo que consiste en la detección de patrones verbales con base en los datos de la gramática.

El primer paso en este proceso es identificar tres grupos de verbos distintos:

1. SIN NEXO, SIN DISTANCIA: aquellos que no tienen nexo y por lo tanto no hay ninguna distancia entre el verbo y el nexo.
2. CON NEXO, SIN DISTANCIA: aquellos que sí tienen nexo pero la distancia entre éste y el verbo es cero.
3. CON NEXO, CON DISTANCIA: aquellos que sí tienen nexo y también una distancia mayor a cero.

El proceso de etiquetamiento comienza por los verbos del segundo grupo, es decir, aquellos que sí tienen nexo pero la distancia es igual a cero. Enseguida se etiquetan los verbos del primer grupo, los que no tienen ni nexo ni distancia. Al último se etiquetan los verbos que sí tienen nexo y distancia. Con este orden de ejecución nos aseguramos que los patrones que pueden o no tener nexo se etiqueten correctamente. Por ejemplo el verbo *denominar* puede incluir o no el nexo *como*, por lo cual, con este orden se etiquetaría correctamente el patrón *denominamos como*.

Para anotar los patrones verbales del segundo grupo se utilizan las siguientes expresiones regulares:

- $vd $sele $nexo

- $vd $nexo

En estos casos, la variable $vd equivale al verbo definitorio anotado en el proceso anterior (por ejemplo <vd lema="ser">es</vd>); la variable $sele conforma una combinación de los pronombres *se* más (*les*|*las*|*los*|*le*|*la*|*lo*), la cual nos permite identificar estructuras como *denominar se como* o *denominár se les como* (tomando en cuenta que en el Corpus Técnico del IULA los pronombres posclíticos se separan para su etiquetamiento POS); por último, en la variable $nexo se almacenan los nexos específicos para cada una de las partículas expresadas en la gramática de patrones verbales. Si estas expresiones regulares se encuentran en el texto, entonces se añaden etiquetas específicas para cada elemento: el nexo se enmarca en *<nx></nx>*, el pronombre en *<pr></pr>*, y el conjunto total en *<pvd></pvd>*. Por ejemplo, los patrones *denomina como* y *denominar se como* se etiquetan de la siguente forma:

- <pvd><vd lema="denominar">denomina</vd> <nx>como</nx> </pvd>
- <pvd><vd lema="denominar">denominar</vd> <pr>se</pr> <nx>como</nx></pvd>

Una vez etiquetado el segundo grupo, entonces se etiqueta el grupo de aquellos verbos que no tienen nexo ni distancia. En este caso los verbos se etiquetan directamente como patrones verbales añadiendo las etiquetas *<pvd></pvd>*. Por ejemplo:

- <pvd><vd lema="contener">contiene</vd></pvd>
- <pvd><vd lema="permitir">permitirá</vd></pvd>

Finalmente, se etiqueta el tercer grupo de verbos que incluyen tanto nexo como distancia. Para ello se utilizan las siguientes expresiones regulares:

- $vd $nexo
- $vd $distancia $nexo

En este caso, la variable $distancia equivale a los datos expresados en la gramática entre las etiquetas <dist></dist>. Si es un valor numérico, éste se sustituirá por una expresión regular que denote el número de palabras requeridas; si el valor equivale a la palabra *any*,

entonces se sustituirá por una expresión regular de cualquier palabra hasta el siguiente nexo inmediato, excepto un verbo en forma conjugada o un elemento dentro de las etiquetas de verbo definitorio. Lo anterior lo establecimos con la finalidad de hacer más precisa la búsqueda de los nexos. Un par de ejemplos que se obtendrían son los siguientes:

- <pvd> <vd lema="utilizar">utilizan</vd> normalmente <nx>en </nx> </pvd>
- <pv> <vd lema="definir">definido</vd> inicialmente con un significado citológico puramente descriptivo <nx>como</nx> </pvd>

El último proceso consiste en expandir la etiqueta de inicio de patrón verbal a la izquierda, con el fin de incluir verbos auxiliares o pronombres que sean partes del patrón verbal definitorio. El pronombre se etiqueta igualmente dentro de *<pr></pr>*, mientras que el verbo auxiliar dentro de *<aux></aux>*. De esta forma, nos aseguramos que un patrón, como *se ha definido como*, incluya el pronombre y el verbo auxiliar dentro del patrón verbal.

Con esta metodología podemos identificar automáticamente una variedad de patrones verbales como los que se muestran en la siguiente tabla:

	Patrones verbales
1	se les denomina
2	pueden utilizarse para
3	conocidos habitualmente como
4	se conocen en castellano como
5	se emplean con mayor frecuencia en
6	definimos la aptitud absoluta del genotipo xxx como
7	es obligado conocer mejor la interacción entre dichas moléculas y otras proteínas, así como
8	se considera la manera para determinar las frecuencias y las tasas de mutación de ciertos genes de importancia clínica, tanto autosómicos como

Tabla 5.15. Ejemplos de patrones verbales extraídos automáticamente

En los ejemplos anteriores observamos que se identifican automáticamente patrones sencillos como 1 y 2; patrones que

incluyen información extra entre el verbo definitorio y el nexo, como 3, 4 y 5; patrones donde el término se puede encontrar en la posición de nexo, es decir entre éste y el verbo definitorio, como en el ejemplo 6; y casos donde la ventana que puede estar constituida por cualquier palabra hasta el siguiente nexo recupera una serie de ruido, como en 7 y 8.

Algunos de los contextos que presentaban algún verbo definitorio pero ningún patrón verbal son los siguientes:

Los atributos esenciales del gen fueron <vd lema="definir"> definidos </vd> por Mendel hace más de un siglo.
El glucógeno se puede <vd lema="formar">formar</vd> a partir de los tres principales azúcares de la dieta.
No obstante, ofrecen sitios de unión para componentes <vd lema="conocer">conocidos</vd> de transducción de señales, una propiedad que sugiere que ellas también coadyuvan a transmitir mensajes desde las integrinas a los genes y a otras partes de la célula.

Tabla 5.16. Contextos con verbos definitorios pero sin patrones verbales

En estos ejemplos de los verbos *definir*, *formar* y *conocer*, los contextos fueron excluidos de los candidatos por no contener uno de los nexos que se expresan en la gramática.

Por último, el tercer módulo que incluye el proceso de identificación de patrones verbales, consiste en etiquetar cada línea de los candidatos con el tipo de información definitoria que se encuentra en ellos, de acuerdo con los datos de la gramática relacionados con los patrones verbales y su asociación a un tipo de información específica. En este proceso simple, únicamente se anota dentro de la etiqueta <*tipoD=“”*/> el valor expresado dentro de <td></td> en la gramática de patrones verbales, por ejemplo <tipoD=“ana”>.

5.4 Filtro de contextos no relevantes

Una vez que hemos extraído y anotado candidatos con ocurrencias de patrones verbales, el siguiente proceso que incluye el algoritmo es el análisis de los candidatos. A su vez, en este análisis el primer proceso consiste en la búsqueda automática de reglas de

excepciones que nos ayude a filtrar contextos donde probablemente no se defina un término, mientras que el segundo proceso es el análisis de los candidatos para identificar los elementos constitutivos, es decir los términos y definiciones. Retomando los datos del algoritmo presentado en la figura 5.1, en la siguiente figura podemos observar un esquema general del filtro de contextos no relevantes.

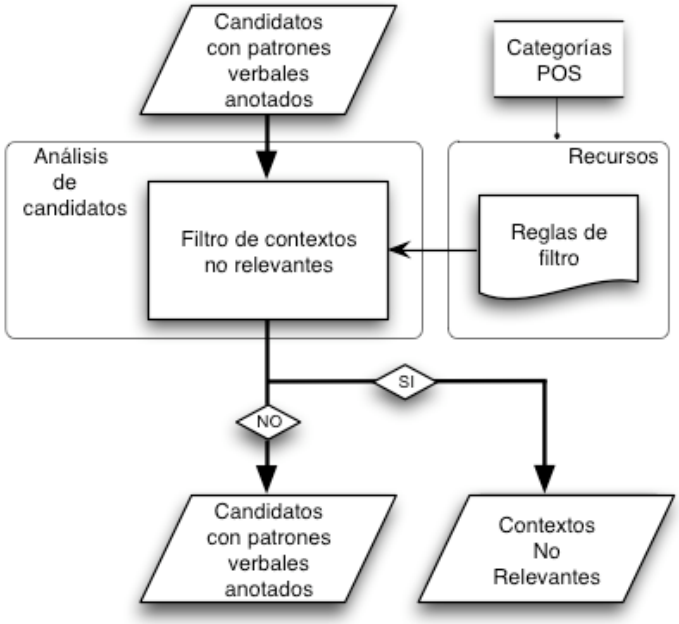

Figura 5.3. Filtro de contextos no relevantes del ECODE

La entrada está formada por candidatos con patrones verbales anotados. El proceso principal consiste en el filtro de contextos no relevantes, para lo cual se toma como recurso principal una serie de reglas de filtro apoyadas en ciertas categorías POS. En este caso, la salida está condicionada: si los candidatos cumplen las reglas del filtro, entonces se consideran como contextos no relevantes, si no las cumplen, los contextos restantes conforman un nuevo grupo de candidatos con patrones verbales anotados.

Como hemos visto a lo largo de este trabajo, existen patrones verbales que pueden servir para conectar al término con su definición y que pueden ser el punto de partida para la búsqueda automática de CDs. No obstante, los patrones verbales no se mplean únicamente en enunciados donde se aporta información relevante

sobre términos. Por su naturaleza, ciertos verbos tienden a utilizarse en enunciados con un carácter mayoritariamente definitorio, como *definir*, *conocer*, *llamar*, etc. Por otro lado, algunos verbos se utilizan recurrentemente en una gran variedad de situaciones donde no necesariamente se define un término, tal es el caso de verbos como *ser*, *permitir* o *concebir*. A su vez, los mismos verbos con un carácter más definitorio no se utilizan siempre en enunciados con el fin de definir un término.

Las células del hibridoma pueden separarse unas de otras y cultivarse indefinidamente, produciendo cada clon un único anticuerpo **conocido como** anticuerpo «monoclonal».
En este estudio se han analizado grupos de sueros de genotipo **conocidos** , así **como** todos los sueros de aquellas personas con un genotipo C282Y/C282Y xxx y C282Y/H63D xxx (Tabla 16).

Tabla 5.17. Ejemplo de contexto definitorio y contexto no relevante

En estos casos vemos ejemplos de ocurrencias del verbo *conocer* en participio acompañado del nexo *como*. En el primero de ellos el verbo se utiliza claramente para introducir información definitoria de un término, pero en el segundo caso el verbo no tiene el mismo uso definitorio ni se utiliza con la misma intención. En este último ejemplo podemos ver la partícula *así*, antes del adverbio *como*, formando una estructura que nos puede dar pistas sobre los casos en que dichos adverbios no constituyen parte del verbo definitorio y por tanto no conforman un patrón verbal.

Así, con el fin de filtrar aquellos candidatos donde no se define un término, hemos generado una lista de restricciones basada en ciertas partículas gramaticales y sus ocurrencias en posiciones específicas dentro de los candidatos. Las restricciones son reglas formadas por partículas gramaticales que incluyen principalmente preposiciones, pronombres, adverbios y verbos en forma conjugada. Estas partículas aparecen en ciertas posiciones específicas: a la izquierda del patrón verbal, en la posición de nexo (es decir, entre éste y el verbo definitorio), y a la derecha del patrón verbal.

En la siguiente tabla presentamos las reglas de filtro.

Posición		Regla
Izquierda	1	para <pvd>
Nexo	2	</vd> .* verbo conjugado .* </nx>
	3	</vd>.*? se .*?<nx>
	4	</vd>.*? tanto .*?<nx>
	5	</vd>.*? sino .*?<nx>
	6	</vd> , <nx>
	7	así <nx>
	8	cerca <nx>
	9	parte <nx>
	10	partir <nx>
	11	más <nx>
	12	menos <nx>
	13	mientras (que ,|que) <nx>
	14	no <nx>
	15	poco <nx>
	16	poco más <nx>
	17	(que ,|que) <nx>
	18	sino <nx>
	19	tales <nx>
	20	tal <nx>
	21	y <nx>
	22	ya <nx>
Derecha	23	</pvd> antes
	24	</pvd> cuan
	25	</pvd> para
	26	</pvd> si
	27	</pvd> se
	28	</pvd> verbo Conjugado

Tabla 5.18. Reglas de filtro de contextos no relevantes implementadas en el ECODE

Observamos que en las posiciones izquierda y derecha, las reglas aparecen inmediatamente antes o después de la etiqueta del patrón verbal. En la posición de nexo, las reglas pueden aparecer en la posición inmediata posterior a la etiqueta de cierre del verbo definitorio, o en la posición anterior a la etiqueta de abertura del nexo, respectivamente. Las reglas 2, 3, 4 y 5 contemplan que puede aparecer cualquier palabra antes y después de la partícula de filtro,

mientras que en el resto de las reglas las partículas ocupan una posición específica. En la siguiente tabla podemos ver algunos ejemplos de contextos filtrados con esta metodología.

Regla		Contexto Filtrado
Izquierda	1	**Para** <pvd> <vd>entender</vd> <nx>como </nx> </pvd> funciona el ADN, es necesario conocer algo sobre su estructura y organización.
	1	La codificación de este ejemplo se podría ampliar **para** <pvd> <vd>incluir</vd> </pvd> tipos registro (record).
Nexo	4	Asimismo, <pvd> <pr>se</pr> <vd>considera </vd> la manera para determinar las frecuencias y las tasas de mutación de ciertos genes de importancia clínica, **tanto** autosómicos <nx>como</nx></pvd> ligados al X.
	7	<pvd><pr>Se</pr> <vd>conocen</vd> ya las secuencias de bases de muchos genes salvajes y mutantes, **así** <nx>como</nx></pvd> las secuencias de aminoácidos de las proteínas que codifican.
Derecha	25	[…] tan pronto hayan transcurrido treinta días a contar desde la delación a su favor , que señale un plazo a l <pvd> <vd>llamado</vd> </pvd> **para** que manifieste si acepta o repudia la herencia .
	*28	En resumen, <pvd> <aux>podría</aux> <vd>definir</vd> <aux>se</aux> a cualquier organismo o individuo <nx>como</nx> </pvd> **aquello** que estructuralmente determina su ADN que sea.

Tabla 5.19. Ejemplos de contextos filtrados con las reglas implementadas en el ECODE

En la tabla anterior, los ejemplos para las reglas de filtro en la posición izquierda incluyen la preposición *para* en la posición inmediata anterior a la etiqueta de abertura del patrón verbal. Los ejemplos para la posición de nexo incluyen las partículas *tanto* y *así* dentro de la información recuperada entre el verbo y el nexo. Por su parte, los ejemplos para la posición derecha incluyen la partícula *para* y verbos en forma conjugada, el cual cumple la regla de filtro debido a un mal etiquetado POS en el corpus original, ya que el pronombre aquello es erróneamente considerado como un verbo

(aquello/VDR1S-). En este caso, la regla se cumple pero filtra un candidato bueno.

En resumen, las reglas de filtro representan las posiciones en las que pueden aparecer ciertas partículas adyacentes al patrón verbal o dentro de éste. El filtrado automático de excepciones se realiza a partir de la búsqueda de dichas partículas en posiciones específicas, apoyándose en las etiquetas anotadas en el proceso anterior, así como ciertas categorías de las etiquetas POS del corpus.

5.5 Identificación de elementos constitutivos

En el segundo proceso que incluye el análisis de los candidatos es la identificación de los términos y las definiciones en los contextos que no fueron filtrados como excepciones. Como hemos explicado anteriormente en la descripción de los patrones contextuales, los términos y las definiciones pueden ocupar un lugar específico en los CDs. Por ejemplo, un patrón como *consiste en* generalmente presenta al término en una sola posición, es decir, a la izquierda del patrón verbal, por lo cual resulta fácil determinar el término y la definición. En el caso de otros patrones verbales, y dependiendo específicamente del verbo que se utilice para conectar al término con su definición, el número de posiciones podría aumentar considerablemente, como en el caso de algunas combinaciones que podría seguir un patrón con el verbo *definir*:

Patrón contextual	**Ejemplo**
T + VD + NX + D	T se define como D
D + VD + NX + T	D se define como T
VD + T + NX + D	se define T como D
VD + NX + T + D	se define como T D
PPR + T + VD + NX + D	generalmente T se define como D
T + PPR + VD + NX + D	T generalmente se define como D
VD + PPR + T + NX + D	se define generalmente T como D
VD + T + PPR + NX + D	se define T generalmente como D

Tabla 5.20. Ejemplos patrones contextuales para el verbo *definir*

Observamos que T y D pueden aparecer a izquierda, derecha, o bien entre el patrón verbal definitorio y el nexo. A su vez, si

consideramos la inclusión de patrones pragmáticos, como el adverbio *generalmente*, el número de combinaciones posibles en el orden de aparición de los elementos constitutivos se incrementa considerablemente.

Así, el problema en este proceso de identificar los elementos constitutivos radica en una clasificación del contenido de los candidatos con patrones verbales. Específicamente, esta etapa está relacionada con la toma de decisiones para clasificar qué palabras de los candidatos pertenecen al término y cuáles pertenecen a la definición. El esquema de este proceso lo podemos ver en la siguiente figura.

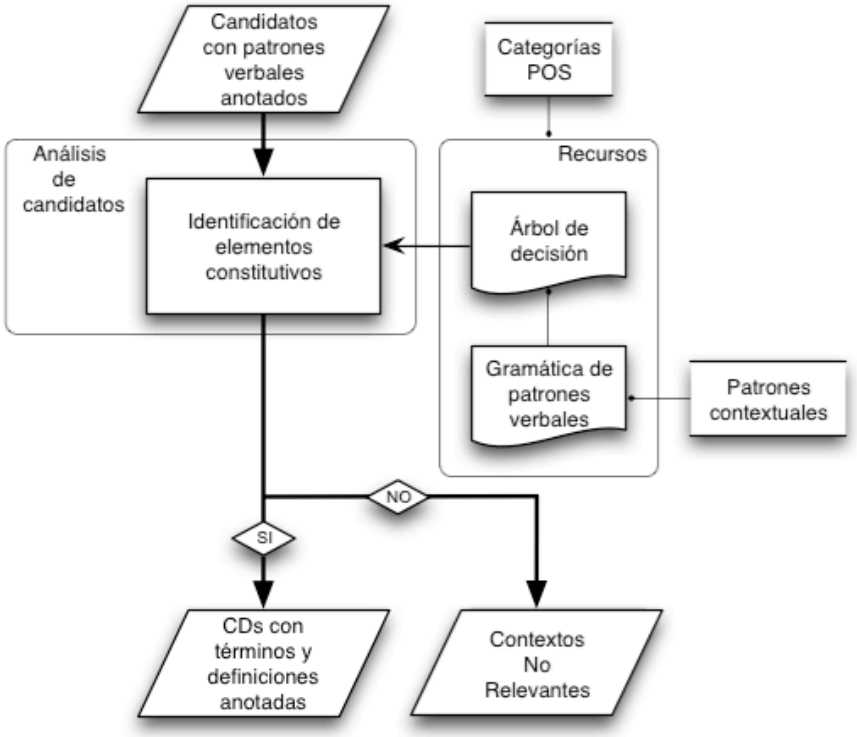

Figura 5.4. Identificación de elementos constitutivos del ECODE

Aquí, la entrada está formada por los candidatos con patrones verbales que no fueron filtrados como contextos no relevantes. El proceso principal, es decir, la identificación del término y la definición, toma como recurso principal un árbol de decisiones basado en la gramática de patrones verbales, específicamente en los

patrones contextuales, y se apoya de igual forma en algunas categorías de las etiquetas POS. Como en el proceso anterior, la salida está condicionada: si los candidatos cumplen las reglas del árbol de decisión, se consideran entonces como candidatos con términos y definiciones anotadas; de lo contrario, se consideran contextos no relevantes por exclusión.

Con el fin de implementar una metodología para el reconocimiento automático de los elementos constitutivos, el módulo encargado de este proceso parte de una simple anotación de los candidatos con unas etiquetas que hemos denominado *etiquetas contextuales*. Recordemos que hasta ahora lo que tenemos anotado son los patrones verbales dentro de las etiquetas de *<pvd></pvd>*. En este paso, se añaden además las etiquetas *<izq></izq>* y *<der></der>* a la izquierda y derecha del patrón verbal. Asimismo, la etiqueta de abertura de nexo se expande a la izquierda hasta el final del verbo definitorio o el pronombre que forme parte de dicho verbo. Por lo tanto, los candidatos quedan anotados de la siguiente forma:

<izq>El turismo, en términos generales,**</izq> <pvd><aux>**ha sido**</aux> <vd>**concebido**<vd> <nx>**como**</nx></pvd> <der>**la reproducción de los hábitos cotidianos en un ambiente diferente.**</der>**
<izq>El metabolismo**</izq> <pvd><aux>**puede**<aux> <vd>**definir **</vd> <pr>**se**</pr> <nx>**en términos generales como**</nx></pvd> <der>**la suma de todos los procesos químicos (y físicos) implicados.**</der>**
<izq>NULL**</izq> <pvd><pr>**Se**</pr> <vd>**define**</vd> <nx>**como **</nx></pvd> <der>**la medida cuadrática de todas las desviaciones de cada valor de la variable con respecto a su media aritmética**.</der>**

Tabla 5.21. Ejemplos de anotación contextual

En el primer caso sólo se anota la parte izquierda y la parte derecha del patrón. En el segundo ejemplo se expande la etiqueta de nexo hasta la palabra inmediatamente después del pronombre *se* ligado al verbo definitorio. Dicha expansión se lleva a cabo hasta este momento debido a que, en el proceso de filtro de contextos no relevantes, la etiqueta de apertura del nexo constituye una frontera de las expresiones regulares usadas por cada regla. Por otro lado, observamos en el tercer ejemplo que si la izquierda del patrón

verbal corresponde a un espacio vacío después del inicio de línea[11], entonces se etiqueta esta posición con la palabra *NULL*.

Una vez que se ha realizado este re-etiquetamiento es posible realizar el proceso para determinar los elementos constitutivos. Para ello, en este módulo se utilizan ciertas expresiones regulares para encontrar el inicio de las estructuras sintácticas de los términos, definiciones y patrones pragmáticos[12]. Las principales expresiones regulares propuestas se pueden observar en la siguiente tabla[13]:

Variable	Expresión Regular
$termino	$fron ($det)? ($sus | $adj) .* $fron
$definicion	$fron $det $sus .* $fron
$pragma	$fron ($coma ($adv | $prep) .* | $adv) $fron

Tabla 5.22. Expresiones regulares de los elementos constitutivos

Donde:

- $termino: término
- $definicion: definición
- $pragma: patrón pragmático
- $det: determinante
- $sus: sustantivo
- $adj: adjetivo
- $adv: adverbio
- $prep: preposición
- $coma. el signo de puntuación ","
- .*: cualquier palabra o número de palabras
- ()?: variable optativa
- $fron: equivale a una frontera que puede ser cualquier etiqueta contextual, cualquier etiqueta de los patrones verbales, o bien

[11] Recuérdese que cada línea del corpus original debe comenzar y terminar con un espacio en blanco.

[12] Si bien no incluimos un proceso para reconocer los patrones pragmáticos, consideramos indispesable la inclusión de expresiones regulares para representar el inicio de sus posibles estructuras, con el fin de tener un mejor funcionamiento del sistema.

[13] Existen variaciones que pueden observarse en el disco adjunto como apéndice en la ruta: *ecode/gramaticas/01_gramaticaPOS.pm*.

la misma expresión regular de cada elemento constitutivo, excluyendo la frontera

Algunos ejemplos de estas expresiones regulares son los siguientes:

<izq>El turismo, en términos generales,</izq> <pvd><aux>ha sido</aux> <vd>concebido</vd></pvd> <nx>como</nx> <der>la reproducción de los hábitos cotidianos en un ambiente diferente.</der>
<izq>Este conjunto de materiales básicamente orgánicos, generados a partir de la fotosíntesis o bien evolucionados en la cadena biológica y que son susceptibles de degradación o fermentación bioquímica</izq> <pvd><aux>son</aux> <vd>definidos</vd> <nx>como</nx></pvd> <der>biomasa.</der>

Tabla 5.23. Ejemplos de contextos y expresiones regulares de los elementos constitutivos

Si tomamos en cuenta que el término y la definición pueden aparecer tanto en posición izquierda como derecha del patrón verbal, en el primer ejemplo se reconocería entonces el término por la expresión regular que representa una frontera (<*izq*>) seguida de un determinante y un sustantivo, más todo lo que esté después hasta otra frontera, que en este caso sería la expresión regular para el inicio de patrón pragmático. Este último se reconocería porque empieza con una coma seguida de una preposición, e incluye todo lo que sigue hasta la próxima frontera que es la etiqueta de cierre de la posición izquierda. En este mismo ejemplo, la definición sería reconocida por su expresión regular correspondiente, es decir, determinante más sustantivo, y las fronteras serían el inicio y cierre de la posición derecha. En el segundo ejemplo, por el contrario, se podría reconocer que la información que está en la posición izquierda seguiría el modelo de la expresión regular de una definición. Esto se podría deducir tomando en cuenta que la posición derecha está ocupada únicamente por un sustantivo, y que la expresión regular para términos, a diferencia de la expresión para definiciones, indica que éstos pueden o no incluir un determinante.

5.5.1 Árbol de decisión

Ahora bien, con el fin de resolver el problema de la identificación de los elementos constitutivos, hemos decidido utilizar un árbol de

decisión para determinar, mediante inferencias lógicas, las distintas posibilidades de aparición de los términos y las definiciones.

Según Moreno *et al* (1994: 49). "un árbol de decisión es una representación posible de los procesos de decisión involucrados en tareas inductivas de clasificación". Los árboles de decisiones son funciones de clasificación que están estructuradas como un árbol: tienen *nodos*, *ramas*, y *hojas*. Los nodos son decisiones tomadas a partir de atributos representados por las ramas y las hojas son los elementos ya clasificados. Como señala Mooney,

> "decision trees are classification functions represented as trees in wich the nodes are feature test, the branches are attribute values, and the leaves are class labels. Rules are implications in either propositional or predicate logic used to draw deductive inferences for data". (Mooney, 2003, p. 377)

En nuestra investigación, el árbol de decisiones representaría las inferencias lógicas para encontrar los posibles términos y definiciones en los candidatos. Las ramas en un primer nivel serían las distintas posiciones en las que pueden aparecer los elementos constitutivos, es decir izquierda, derecha y opcionalmente nexo; y en un segundo nivel serían las expresiones regulares para identificar cada elemento constitutivo. Los nodos corresponderían a las decisiones tomadas a partir de los atributos de cada rama y estarían relacionados entre sí a nivel horizontal por inferencias del tipo *IF, IF NOT*, y a nivel vertical por inferencias del tipo *THEN*. Por último, las hojas serían las distintas posiciones una vez reconvertidas en el término o la definición.

Para ejemplificar el uso de un árbol de decisiones tomemos el caso de las decisiones que se deberían tomar para clasificar la información partiendo del análisis de la posición izquierda. A grandes rasgos, las inferencias que se seguirían en este caso serían las siguientes:

1. SI la posición izquierda *sólo* está ocupada por una Expresión Regular de Término **ERT:** entonces la posición izquierda equivale al término **<izq> = T** y la posición derecha equivale a la definición **<der> = D**; si no:

2. SI la posición izquierda está ocupada por una **ERT** y una Expresión Regular de Patrón Pragmático **ERPPR**: entonces la posición izquierda equivale al término y al patrón pragmático **<izq> = T**, **<izq> = PPR**, y la posición derecha equivale a una definición **<der> = D**; si no:

3. SI la posición izquierda *sólo* está ocupada por una **ERPPR: e**ntonces la posición izquierda equivale a un patrón pragmático **<izq> = PP** y a) la posición nexo puede equivaler al término **<nexo> = T** y la posición derecha puede equivaler a la definición **<der> = D**; o b) la posición derecha puede equivaler al término y a la definición **<der>= T D**; si no:

4. SI la posición izquierda *sólo* está ocupada por una Expresión Regular de Definición **ERD**: entonces la posición izquierda equivale a la definición **<izq> = D** y la posición derecha equivale al término **<der> = T**.

Este análisis de la posición izquierda se puede representar mediante el esquema de la figura 5.5:

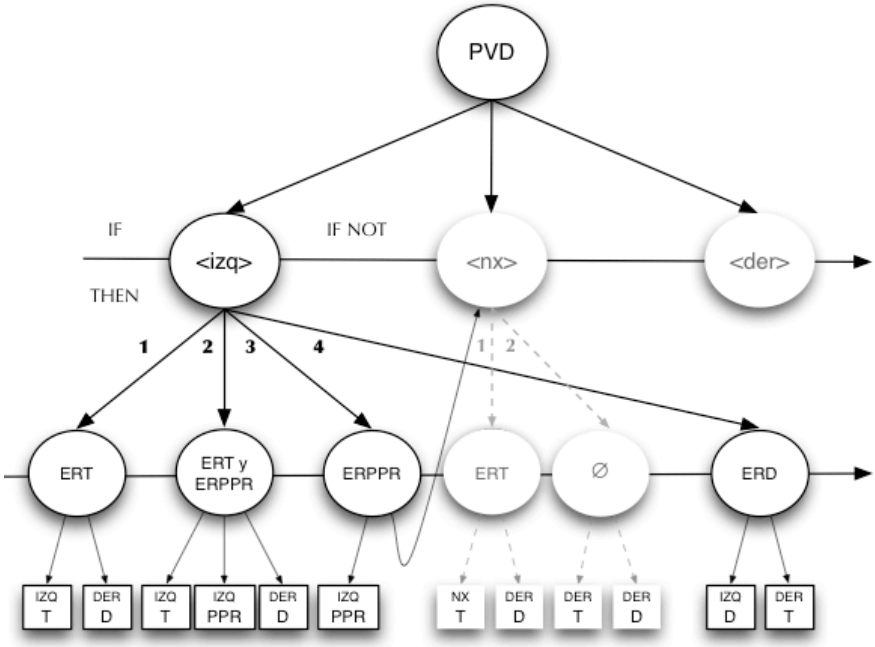

Figura 5.5. Representación del árbol de decisión para el análisis de la posición izquierda

Observamos que para determinar qué elemento se encuentra en la posición izquierda, las decisiones parten del reconocimiento de las expresiones regulares de término, de patrón pragmático o de definición.

Con las inferencias 1 y 2 se puede determinar que la posición izquierda equivale a un término o a un término y un patrón pragmático, los cuales se distinguen por una frontera como un signo de puntuación, mientras que la posición derecha equivale a una definición.

En cambio, con la inferencia 3 se determina que la posición izquierda equivale únicamente a un patrón pragmático, y para saber en qué posición se encuentran el término y la definición se recurre a las inferencias de la posición de nexo.

Entonces, si la primera inferencia de la posición de nexo incluye únicamente una expresión regular de término, la posición nexo corresponde al término y la posición derecha a la definición. Si por el contrario, la posición de nexo no incluye ninguna expresión regular correspondiente a un elemento constitutivo, entonces el término y la definición están en la posición derecha y pueden ser reconocidos por separado a partir de una frontera como un signo de puntuación.

Por su parte, con la inferencia 4 se encuentra a la definición en posición izquierda y al término en posición derecha.

Con este esquema hemos representado el análisis de la posición izquierda. En el caso de las inferencias de la posición nexo, éstas seguirían el modelo de la posición izquierda, sustituyendo *izquierda/<izq>* por *nexo/<nexo>* y omitiendo la inferencia número 3, ya que la definición no podría encontrarse en la posición de nexo.

Para el caso de las posiciones de derecha, se añade una inferencia que permita reconocer los casos en los que el término y la definición se pueden encontrar juntos en esta posición, debido a que las posibilidades de encontrarlos por separado ya habrían sido reconocidas en las posiciones anteriores.

Así, tomando el siguiente ejemplo de candidato a CD:

- <izq>En sus comienzos</izq> <pvd><pr>se</pr> <vd> definió</vd> <nx>la psicología como</nexo></pvd> <der>la descripción y la explicación de los estados de conciencia (Ladd, 1887).</der>

Se encontraría entonces que la posición izquierda:

1. **NO** está ocupada únicamente por una ERT
2. **NO** está ocupada por una ERT y una ERPPR
3. **SI** está ocupada únicamente por una ERPP

Por lo tanto, la posición izquierda corresponde a un patrón pragmático (*En sus comienzos*), y para identificar el término y la definición se recurriría a las inferencias de la posición de nexo, con lo que se encontraría que en este caso:

1. **SI** está ocupada únicamente por una ERT

Por lo tanto, la posición de nexo corresponde a un término, *la psicología*, y la posición derecha corresponde a una definición, *la descripción y la explicación de los estados de conciencia* […]. De esta manera, las etiquetas contextuales y de nexo se reescriben de la siguiente forma:

- En sus comienzos <pvd><pr>se</pr> <vd>definió</vd> <t>la psicología</t> <nx>como</nexo></pvd> <d>la descripción y la explicación de los estados de conciencia (Ladd, 1887).</d>

En síntesis, con la metodología que hemos expuesto pretendemos dar una visión general da la forma en que se puede resolver el problema de la identificación automática de los términos y las definiciones, mediante el uso de un árbol de decisión. Este árbol toma como referencia expresiones regulares que nos permiten identificar los términos, las definiciones y los patrones pragmáticos; apoyado en las etiquetas contextuales y de patrones verbales se puede entonces determinar a qué elementos constitutivos corresponden las palabras que se encuentran en las posiciones izquierda, derecha o nexo.

5.5.2 Implementación y resultados

Para implementar el árbol de decisión se elaboró una serie de reglas en cascada correspondientes a la identificación del término en cada posición. En primer lugar, el módulo encargado de este proceso genera diferentes grupos de verbos partiendo de la gramática de patrones verbales, específicamente de los patrones contextuales, para identificar los casos en los que el término puede aparecer en posición izquierda, nexo o derecha. A lo largo del desarrollo de esta metodología, nos dimos cuenta que los mejores resultados se obtenían implementando en primer lugar las inferencias de la posición de nexo, seguidas de la posición derecha y terminando con la posición izquierda.

De esta forma, el módulo selecciona primeramente los verbos cuyo término puede aparecer en la posición de nexo. Para los verbos de esta posición se elaboraron las siguientes reglas:

Reglas de nexo	
N01	$pragma $termino
N02	$a $termino
N03	$termino2
N04	$termino

Tabla 5.24. Reglas para la identificación de término en la posición nexo

Las dos primeras reglas contemplan la expresión regular de término precedida de un patrón pragmático o de la preposición "a". En el tercer caso, la regla considera una expresión regular distinta para términos, donde no se incluye cualquier palabra sino combinaciones específicas como *$det $sus $adj $prep $sus*, *$det $sus $prep $sus* o *$sus $prep $sus*. Por último, la cuarta regla contempla la expresión regular de término normal, es decir *($det)? ($sus | $adj)* seguida de cualquier palabra hasta la siguiente frontera, en este caso la etiqueta de cierre de la posición nexo. En todas las reglas, la información que se considera como término se etiqueta con *<t></t>*, mientras que las etiquetas de derecha se convierten en *<d></d>*.

N01	[…] es posible <pvd><vd>considerar</vd> entonces <t>cualquier función computable</t> <nx>como</nx></pvd> <d>una función cuyas entradas sean tuplas de enteros no negativos</d> .
N02	<pvd><aux>Podemos</aux> <vd>considerar</vd a <t>l genoma</t> <nx>como</nx></pvd> <d>una lista de todos los productos génicos (RNA y polipéptidos) que pueden ser producidos y el manual de instrucción sobre cuándo , donde , y en qué cantidad deben fabricar se esos productos</d>.
N03	<pvd><pr>Se</pr> <vd>define</vd> <t>una librería genómica</t> <nx>como</nx></pvd> <d>un conjunto de clones en el que está representado todo el genoma de un organismo</d>.
N04	En este método, <pvd><pr>se</pr> <vd>utiliza</vd> <t>la enzima desoxinucleotidil transferasa terminal</t> <nx>para</nx></pvd> <d>crear colas complementarias mediante la adición de fragmentos de poli- dA y de poli- dT</d>.

Tabla 5.25. Ejemplos de los resultados obtenidos con las reglas de la posición nexo

Con estos ejemplos de las reglas de nexo podemos ver los distintos tipos de términos[14] que se pueden recuperar: términos simples como *genoma*, y términos compuestos como *enzima desoxinucleotidil transferasa terminal*. Observamos también las distintas estructuras de los contextos, algunos comienzan directamente con el patrón verbal mientras que otros incluyen otra información en la posición izquierda. Por otro lado, el orden de ejecución de las reglas está pensado para anotar en primer lugar los casos donde se presenta información extra además del término, ya sea un patrón pragmático o la preposición *a*, y excluir así dichos contextos para tener la posibilidad de obtener mejores resultados en las reglas subsecuentes.

Una vez que se procesaron los candidatos con verbos que permiten el término en la posición de nexo, el módulo selecciona aquellos

[14] Estamos conscientes de que estas expresiones regulares no representan la totalidad de las combinaciones sintácticas con las que puede estar formado un término y que pueden implementarse más combinaciones con el fin de mejorar los resultados de la aplicación de las reglas.

que permiten que el término se encuentre en la posición derecha. En la siguiente tabla encontramos las reglas de esta posición:

Reglas de derecha	
D01	($termino2\|$comillas .*? $comillas) $a $definicion
D02	$termino $a $definicion
D03	$comillas [^$verboConjugado]* $comillas
D04	$termino [^$verboConjugado]
D05	$termino $coma
D06	$termino $corte
D07	$termino2 $det

Tabla 5.26. Reglas para la identificación de término en la posición derecha

En las reglas anteriores se incluyen la variable *$comillas*, que corresponde a los símbolos (" | « | »), y la variable *$corte*, que puede ser una coma, cualquier conjunción o un verbo conjugado. Las dos primeras reglas seleccionan al término seguido de la preposición *a* y el comienzo de la definición. Las diferencias de la primera con la segunda son que la primera utiliza la expresión regular de término2, a la cual nos referimos en las reglas de nexo, y añade la posibilidad de que el término sea cualquier elemento enmarcado dentro de comillas. La tercera regla no recurre a la expresión regular de término, sino a cualquier palabra que no sea un verbo conjugado. Lo mismo sucede con la cuarta regla, que selecciona como término todo lo que esté en la posición derecha, exceptuando los casos en los que contenga un verbo conjugado. A decir verdad, estas dos reglas podrían formar una sola, pero decidimos incluir la regla con comillas para tratar de seleccionar en primer lugar a los mejores candidatos. En cuanto a la quinta y sexta regla, éstas consideran al término como todo aquello que comience con la expresión regular correspondiente, hasta la siguiente coma o corte. Por último, la séptima regla incluye igualmente la expresión regular de término hasta el siguiente determinante.

Ahora bien, al momento de analizar los verbos cuyo término aparece en posición derecha, el módulo reconoce aquellos verbos cuyo patrón contextual sólo incluye dicha posición (por ejemplo *llamar* o *denominar*). En estos casos de derecha única, el término se anota en posición derecha, pero la definición se puede anotar tanto en izquierda como en derecha. Por otro lado, si los verbos permiten

más posiciones, entonces puede haber dos casos de anotación: la posición derecha se convierte en etiqueta de término y definición (reglas 1, 2 y 7), o bien la posición derecha se reconvierte en término y la posición izquierda en definición (reglas 3, 4, 5 y 6).

D01	<pvd><pr>Se</pr> <vd>denomina</vd></pvd> <t>conversión génica</t> a <d>la sustitución o reemplazo de la secuencia esperada en una región por otra de un tamaño similar correspondiente a un gen no alélico relacionado</d>.
D02	<pvd><vd>Llamamos</vd></pvd> <t>deletéreas</t> a <d>las variaciones - y por ende a las mutaciones causantes de las mismas - que merman la capacidad de supervivencia o de reproducción de los organismos donde se manifiestan</d>.
D03	Por otra parte, existen <d1>genes,</d1> <pvd><vd> denominados</vd></pvd> <t>"prd-like"</t>, <d2>que codifican proteínas con un homeodominio similar a los genes con secuencia "paired" pero que carecen de la misma</d2>.
D04	<d>La célula que experimenta la meiosis</d> <pvd><pr> se</pr> <vd>llama</vd></pvd> <t>oocito primario</t>.
D05	<d1>Los virus de RNA</d1> <pvd><vd>conocidos</vd> <nx>como</nx></pvd> <t>retrovirus</t>, <d2>poseen la propiedad singular de transcribir RNA en DNA, utilizando la enzima transcriptasa inversa</d2>.
D06	<d1>La gametogénesis masculina</d1> <pvd><vd> denominada</vd></pvd> <t>espermatogénesis</t> <d2>da lugar a cuatro espermatozoides haploides para cada célula que inicie la meiosis</d2>.
D07	Ya se ha hecho mención de que el propio concepto de gen ha ido cambiando a medida que ha progresado el conocimiento, pero en la mayoría de los casos <pvd><pr>se</pr> <vd> entiende</vd> <nx>como</nx></pvd> <t>gen</t> <d>una unidad transcripcional, incluyendo sus regiones reguladoras asociadas</d>.

Tabla 5.27. Ejemplos de los resultados obtenidos con las reglas de la posición derecha

En la tabla anterior presentamos algunos resultados del procesamiento de la posición derecha. Podemos observar la variedad de las formas estructurales de los términos que se recuperan, que va desde términos simples como *gen*, *retrovirus* o *deletéreas*, hasta términos multipalabra como *conversión génica*.

Igualmente, algunos contextos comienzan directamente con el patrón verbal, mientras que algunos incluyen otra información en la posición izquierda. Observamos asimismo que se introducen las etiquetas <*d1*></*d1*> y <*d2*></*d2*>, con las cuales se pretende dar cuenta de que la información definitoria se puede encontrar en diferentes lugares. Por ejemplo, el contexto anotado con la regla D03, habla acerca de *genes,* denominados *prd-like,* que *codifican proteínas con un homeodominio*, es decir, introduce primeramente el género próximo del término que se define y en seguida explica sus características distintivas. Estos casos se reconocen con el análisis de los verbos que permiten la aparición del término en posición izquierda y derecha.

Por otra parte, el orden de ejecución de las reglas también es un factor importante para tratar de encontrar, en primer lugar, a los mejores candidatos. Si se aplica la regla D07 en los primeros lugares, la cual considera como término todo lo que aparece después de la etiqueta <*der*> hasta el siguiente determinante, entonces es bastante probable que se recupere sólo parcialmente algunos términos compuestos.

Finalmente, después de haber analizado la posición nexo y derecha, el módulo analiza los candidatos con patrones verbales que permiten el término en la posición izquierda.

Las reglas de la posición izquierda se pueden observar en la tabla 5.28. En estos casos observamos la inclusión de las variables *$parentesis*, *$que*, *$demos*, *$verboInfinitivo* y *$verboParticipio*. La variable *$demos* incluye todos los pronombres demostrativos, mientras que las demás variables se explican por sí mismas. Se añaden además dos reglas de exclusión, la inicial corresponde a los casos donde la posición izquierda está vacía, y la exclusión final corresponde a todos los contextos restantes no identificados con estas reglas. Esto último se consideró tomando en cuenta que, con estas reglas, es bastante probable no abarcar la cantidad de formas en las que se pueden expresar estructuralmente los CDs. Por tanto, si almacenamos los candidatos que no cumplen alguna de nuestras reglas, tenemos la posibilidad futura de definir reglas que no hayan sido contempladas con anterioridad.

Reglas de izquierda	
I00	Regla de exclusión inicial - <izq>NULL</izq>
Izquierda sin verbo conjugado	
I01	$termino2
I02	$termino $pragma
I03	$termino $parentesis $termino $parentesis
I04	$pragma $termino
I05	$termino [^$coma]
I06	$termino $coma .* $coma
I07	$termino $coma $palabra
I08	.* $coma $termino
I09	$termino
I10	.* $termino
Izquierda con verbo conjugado	
I11	$demos $termino
I12	.* $verboConjugado $termino2
I13	.* $coma $termino2
I14	.* $coma $termino
I15	.* $que $termino
I16	.* $parentesis $termino
I17	.* $verboConjugado $termino
I18	.* $verboParticipio $termino
I19	.* $verboConjugado $verboInfinitivo $termino
I20	.* $que $termino
I21	.* $comas $termino $comas .*
I22	Regla de exclusión final - ELSE

Tabla 5.28. Reglas para la identificación de término en la posición izquierda

Observamos que las reglas se dividen en dos grupos, dependiendo de si en la posición izquierda aparece o no un verbo conjugado. Esto, junto con el orden de ejecución de las reglas, desempeña un papel clave para obtener una mejor anotación, ya que primero se anotan los candidatos más *prototípicos* y, en segundo lugar, aquellos que pueden incluir una mayor cantidad de información, además del término, y por ello son más propensos a presentar errores en la anotación.

Algunas observaciones específicas de estas reglas: la número 5 anota la posición izquierda sin verbo conjugado, siempre y cuando no haya una coma; la regla 7 previene casos con un término seguido

de coma y cualquier otra palabra, por ejemplo un adverbio; la diferencia entre la regla 9 y 10 es que la primera anota la posición izquierda como término sólo si ésta comienza con la expresión regular correspondiente, mientras que la segunda anota todos los casos restantes siempre que se cumpla la expresión regular en otro lugar dentro de la posición izquierda; a partir de la regla 12 se anotan casos donde la posición izquierda comienza con cualquier palabra hasta un lugar específico, que puede estar marcado por un verbo conjugado, una coma, la conjunción *que*, etc.; por último, la regla 21 prevé casos donde el término puede aparecer entre comas en cualquier distancia excepto al principio o al final de la posición.

En la tabla siguiente presentamos algunos ejemplos de los contextos anotados con estas reglas de posición izquierda.

I00	<izq>NULL</izq> <pvd><aux>Está</aux> <vd>compuesto </vd> <nx>por</nx></pvd> <der>millones de clones de linfocitos</der> .
I03	<t>La hibridación in situ con fluorescencia (FISH)</t> <pvd><vd>es</vd></pvd> <d>una técnica en que una sonda marcada se hibrida con cromosomas en metafase, profase o interfase</d>.
I04	En el músculo, <t>el fosfato de creatina</t> <pvd><vd> es</vd></pvd> <d>una molécula de alta energía que tiene una especial importancia</d>.
I06	<t>La técnica de ARMS</t>, descrita recientemente por Newton et al., (1989), <pvd><vd>permite</vd></pvd> <d>detectar todo tipo de mutaciones puntuales y pequeñas deleciones</d> .
I14	La explicación es aplicable a los genes de esas poblaciones; y así, <t>el bantú</t>, en su origen una categoría lingüística, <pvd> <pr>se</pr> <vd>emplea</vd> <nx>ahora</nx> <nx>para</nx> </pvd> <d>designar un conjunto de poblaciones que comparten una base lingüística y genética</d>.
I21	Las algas procarióticas, o sea, <t>las cianobacterias</t>, poseen células que se organizan y <pvd><vd>funcionan</vd> <nx> como</nx></pvd> <d>cloroplastos, en muchas formas.</d>
I22	<izq>Tal</izq> <pvd><vd>es</vd></pvd> <der>el caso de la proopiomelanocortina (POMC) cuyo gen estructural está en el cromosoma 2</der>.

Tabla 5.29. Ejemplos de los resultados obtenidos con las reglas de la posición izquierda

En los ejemplos observamos casos de anotación del término en la posición izquierda. En todas estas reglas, lo que se reescribe como definición es la etiqueta de posición derecha. También observamos la diferencia de términos que se recuperan, por ejemplo *bantú*, *fosfato de creatina*, *entorno*, etc., así como la diversidad en la estructura de cada contexto. Algunos están formados por una estructura más prototípica, donde la única información extra que se introduce además del término y la definición, es un patrón pragmático. Tal es el caso del ejemplo de la regla I04. Pero en otros casos, el CD se introduce en una estructura discursiva mayor, lo cual supone un proceso menos preciso a la hora de anotar los elementos constitutivos, ya que se puede anotar como término o definición algo que no es parte de estos elementos.

Es importante notar además la diversidad de lugares donde el término puede ocurrir dentro de la misma posición izquierda, lo cual nos lleva a considerar el hecho de que se necesitan más reglas que en las posiciones de nexo y derecha. No obstante, debe tenerse en cuenta que entre más reglas se definan a partir de experimentos con un corpus de pruebas, se corre el riesgo de que dichas reglas resulten específicas al ámbito de desarrollo de los experimentos. Cabe señalar que el ruido anotado en estos casos es mayor que en las posiciones anteriores. Esto último lo veremos con detalle en el siguiente capítulo de evaluación.

En cuanto a los ejemplos de las reglas de exclusión, el primero de ellos (regla I00) es un contexto que introduce información extensional sobre un término que probablemente se haya introducido en una línea o párrafo anterior, pero al estar elidido no cumple con las reglas de formación de CDs donde se especifica que el término es uno de los dos elementos constitutivos mínimos. Por su parte, el último contexto de la tabla no cumple con ninguna de las reglas definidas para esta posición, por lo cual se elimina con la regla de exclusión final (I22).

Como último paso en el proceso de identificación de los elementos constitutivos, el ECODE contiene un módulo con reglas de re-etiquetamiento para tratar de mejorar los resultados cuando hay algunas palabras o secuencias específicas dentro de los elementos que previamente han sido etiquetados como término o definición. Con ello se busca complementar al módulo anterior para tener una

mayor precisión en el etiquetamiento de los elementos constitutivos. A grandes rasgos, en este proceso se eliminan comas, adverbios o secuencias como *el cual* en la última posición del término, o bien la etiqueta de inicio o fin de término se mueve para tratar de incorporar resultados más precisos. Tómese el siguiente contexto re-etiquetado para ejemplificar lo anterior:

- <d1>Los enzimas específicos</d1> <pvd><vd>denominados </vd></pvd> <t>**quinasas catalizan la inserción de grupos fosforilo**</t> <d2>, mientras que las fosfatasas catalizan su separación por hidrólisis</d2>.
- <d1>Los enzimas específicos</d1> <pvd><vd>denominados </vd></pvd> <t>**quinasas**</t> <d2>catalizan la inserción de grupos fosforilo , mientras que las fosfatasas catalizan su separación por hidrólisis</d2>.

En este caso se aplica una regla que mueve la etiqueta de cierre de término cuando lo etiquetado como tal contiene un verbo conjugado, un pronombre *se* más un verbo conjugado o una conjunción, e inmediatamente después de la etiqueta de cierre del término hay una etiqueta de abertura de definición <*d2*>.

En resumen, el proceso de identificación de elementos constitutivos parte del análisis de la gramática de patrones verbales para etiquetar los posibles términos en los candidatos, dependiendo de las posiciones que permita cada patrón verbal. Para lograr este fin, desarrollamos un árbol de decisión que contempla ciertas expresiones regulares básicas para tratar de representar el inicio de lo que pueden ser términos, definiciones y patrones pragmáticos. Este árbol lo implementamos en un módulo del sistema a través de una serie de reglas en cascada, las cuales, a lo largo del desarrollo de nuestra investigación, nos dimos cuenta que aportan mejores resultados si comienzan buscando los posibles términos de la posición nexo, seguido de la posición derecha y finalizando en la posición izquierda. Por último, hemos visto además que se incluye un módulo para realizar un re-etiquetamiento, cuya finalidad es tratar de mejorar la precisión de lo anotado anteriormente.

5.6 Ranking de CDs

Hasta ahora hemos hecho mención de los procesos generales de extracción y análisis de candidatos. Estos procesos incluyen la identificación de contextos con patrones verbales, el filtrado de contextos no relevantes y la identificación del término y la definición. Hemos visto asimismo que la salida de estos procesos es un grupo de CDs anotados con etiquetas de término, patrón verbal y definición, asignados a un tipo de información definitoria específica que, de manera general, serviría para entender el significado del término que en ellos se presenta.

Ahora bien, la información presente en cada CD puede considerarse *completa* en mayor o menor medida, dependiendo no sólo del tipo de verbo ligado a una definición[15], sino a la misma estructura del contexto recuperado automáticamente. Tomemos los siguientes resultados para ejemplificar lo anterior.

<t>Un gen</t> <pvd><vd>es</vd></pvd> <d>una secuencia de ADN que codifica una proteína, ARNt o ARNr</d>.
<t>Un cromosoma</t> <pvd><vd>es</vd></pvd> <d>una fibra larga e ininterrumpida de ADN a lo largo de la cual hay muchos genes</d>.
<t>La hélice</t> <pvd><vd>es</vd></pvd> <d>una estructura semejante a un cilindro</d>.
<t>El DNA</t> <pvd><vd>es</vd></pvd> <d>una hélice doble</d>.

Tabla 5.30. Ejemplos de CDs obtenidos hasta el proceso de identificación de elementos constitutivos del ECODE

Los primeros dos CDs representan ejemplos de definiciones analíticas donde se introduce el género próximo y la diferencia específica para los términos *gen* y *cromosoma*. Por el contrario, en el tercer y cuarto ejemplo se introduce información sobre los términos *hélice* y *DNA* que, a diferencia de los ejemplos anteriores,

[15] Recordemos que nuestro modelo de verbos definitorios incluye cuatro tipos de definición que son analítica, funcional, extensional y sinonímica. Mientras que en las definiciones del primer tipo se incluye información sobre el género próximo del término y sus características distintivas, en los otros tres tipos sólo se presentan características relacionadas con sus partes o extensión, su función, o bien términos relacionados que pueden funcionar al mismo nivel semántico.

puede considerarse más corta en el sentido de la cantidad de información relevante que aporta sobre el término.

En este sentido, con la idea de complementar los procesos explicados hasta ahora, se decidió incluir un ranking para la evaluación de los resultados. Este último proceso tiene la finalidad de identificar aquellos contextos que presenten estructuras más prototípicas de término y definición, con lo cual se pueda reorganizar los resultados de acuerdo con su probabilidad de ser buenos CDs. Así, en este último apartado describimos un acercamiento al desarrollo de una metodología para realizar un ranking de CDs.

El esquema general de este proceso lo podemos ver en la siguiente figura:

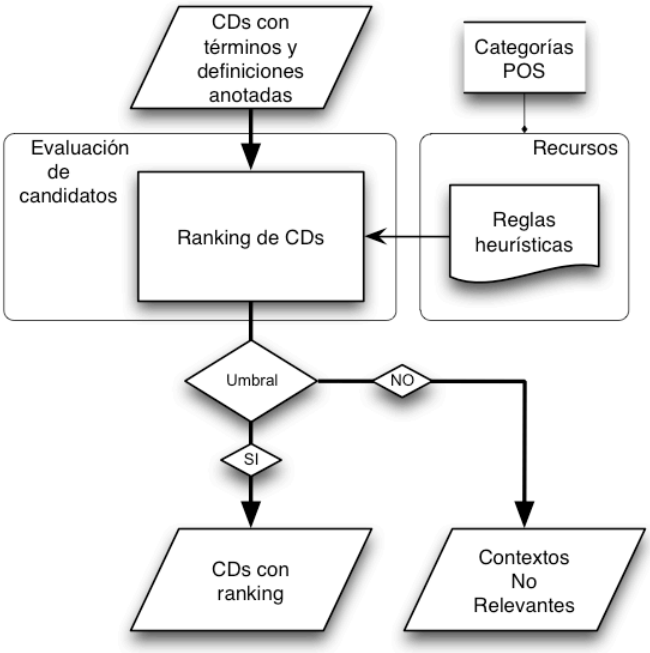

Figura 5.6. Ranking de CDs del ECODE

En este caso, la entrada está constituida por CDs con términos y definiciones anotadas. El recurso utilizado es un conjunto de reglas heurísticas que se utilizan para analizar las estructuras sintácticas de lo etiquetado como término y definición. Este recurso se basa también en algunas categorías POS del corpus original. El proceso

de ranking asigna un valor numérico a cada elemento, combina los resultados y analiza si cada contexto supera o no un umbral que sirve como un último filtro de contextos no relevantes.

Para llevar a cabo el proceso se definieron manualmente reglas para evaluar el término y la definición etiquetados de manera automática. En la siguiente tabla podemos ver algunas de estas reglas[16] de evaluación del término:

Reglas de término	
Valor	Regla
1	<t>$comillas.*$comillas</t>
1	<t>.*$parentesis $termino $parentesis</t>
2	<t>$coma.*$coma .*</t>
2	<t>$parenteis.*$parentesis .*</t>
3	<t>$demos\|$demos.*</t>
3	<t>.*$pron.*</t>

Tabla 5.31. Ejemplos de reglas de ranking para el término

A cada regla le corresponde un valor numérico simple, que va del 1 a 3, para indicar un gradiente de *bueno* a *malo*, donde 1 es el mejor resultado. Podemos observar que el primer ejemplo de las reglas considera como buen candidato a los casos en los que la estructura anotada como término está enmarcada en comillas. En el segundo caso se asigna igualmente el mejor valor cuando el término incluye información entre paréntesis que cumpla con la expresión regular de término, justo al final de la etiqueta. Lo anterior se tomó en cuenta pensando en que puede ser parte de un acrónimo o una referencia a un término en otro idioma. La tercera y cuarta regla asignan un valor 2 para representar aquellos casos donde lo etiquetado como término comienza con una coma o un paréntesis, seguido de cualquier palabra hasta otra coma u otro paréntesis, lo cual puede deberse a un mal etiquetamiento del proceso anterior. Estas reglas se asignan sobre todo para que estos casos no aparezcan dentro de los primeros resultados. En las últimas dos reglas, la primera de ellas busca que lo etiquetado como término corresponda únicamente a un pronombre demostrativo, o comience con uno de ellos. Cabe señalar que en un texto especializado es común que los términos no

[16] Las reglas de ranking para término y definición pueden consultarse en la ruta *ecode/lib/07_rankingTyD.pm* en el disco anexo.

se repitan constantemente y en su lugar se recurra a referencias anafóricas para sustituir su presencia. Así, estos casos se anotaron con un valor 3 y se les asignó además una etiqueta para denotar posibles anáforas: <*anf*></*anf*>. La última regla busca en lo etiquetado como término los casos que contienen dentro un pronombre y les asigna el valor 3.

En la siguiente tabla presentamos algunos ejemplos de los resultados de estas reglas.

Valor	Término
1	<t>«intrones»</t>
1	<t>La dopamina beta-hidroxilasa (DBH)</t>
2	<t>, por tanto , anticolinesterásicos</t>
2	<t>(de una forma vaga) genomas plasmídicos</t>
3	<t><anf>Este cloroplasto</anf></t>
3	<t>Cada una de ellas</t>

Tabla 5.32. Ejemplos términos etiquetados con las reglas de ranking

Observamos que el primer ejemplo presenta un término dentro de comillas, lo cual lo lleva a ser considerado con el valor más alto. De igual manera, el segundo caso se considera bueno en tanto presenta información extra dentro de un paréntesis, que cumple a su vez con la expresión regular de término; en este caso la información corresponde a un acrónimo. En el tercer caso observamos que se introduce parte de una reformulación con la frase *por tanto*, que se debe a un mal etiquetamiento del proceso anterior. Igualmente, el cuarto ejemplo incluye la frase *de una forma vaga* que no corresponde a la estructura sintáctica del término. Como mencionamos anteriormente, estas reglas se aplicaron con el fin de excluir los casos con problemas de etiquetamiento. Por último el quinto ejemplo presenta un caso de una referencia anafórica, mientras que el sexto ejemplo corresponde a la regla que excluye como términos a las formas que contengan un pronombre.

Por otro lado, en la siguiente tabla presentamos algunas de las reglas de ranking de las definiciones:

Reglas de definición	
Valor	Regla
1	<d>.* $que $verboConjugado</d>
1	<d2>$verboConjugado.*</d2>
2	<d>$palabra {,5}</d>
2	<d>.* (;|$noObstante| $sinEmbargo) .*</d>
3	<d>$demos</d>
3	<(d|d1)>NULL</(d|d1)>

Tabla 5.33. Ejemplos de reglas de ranking para la definición

En el primer caso se asignan los mejores valores si el verbo contiene una secuencia del tipo *que* + *verbo conjugado* en la posición derecha. En la segunda regla se toma en cuante que la posición etiquetada como d2 comience inmediatamente con un verbo conjugado. Recuérdese que en algunos casos, dependiendo del verbo definitorio y si el término se encuentra en la derecha, la definición puede presentarse tanto en la posición izquierda como en la derecha.

Por su parte, la tercera regla aplica el valor 2 a los casos en que la definición consta de menos de 5 palabras, lo cual se consideró para tratar de diferenciar contextos donde se aporta una mayor cantidad de palabras y probablemente más información sobre el término. Cabe señalar que en esta regla sólo se aplica a definiciones de tipo analítico, ya que la cantidad de palabras que forman las definiciones en los tipos extensionales, funcionales y sinonímicos puede ser de una o dos palabras y no por ello constituir una definición menos completa.

Por otro lado, en la cuarta regla se incluye un punto y coma junto con las frases *no obstante* y *sin embargo*, para tratar de identificar aquellos casos donde se incluye información extra que probablemente ya no sea parte de la definición. La quinta regla considera con el valor 3 a los casos que incluyen sólo un pronombre demostrativo dentro de las etiquetas, el cual probablemente esté haciendo referencia a otros elementos anteriores que no se encuentran en el mismo contexto; también se etiqueta con *<anf></anf>*.

Finalmente, la última regla asigna el valor 3 a los casos que tienen un espacio vacío representado por la palabra *NULL* en el lugar de la definición. Cabe señalar que estos casos no se filtraron anteriormente en el árbol de decisión, como se hizo en el caso de las posiciones vacías de izquierda, debido principalmente al orden de ejecución de las reglas (en el caso de la posición izquierda, los espacios vacíos se filtran una vez que se han procesado las demás posiciones).

En la siguiente tabla presentamos algunos ejemplos de los valores asignados por le método de ranking a las definiciones propuestas por el ECODE.

Valor	Definición
1	<t>la mutación rutabaga</t> <pvd><vd>es</vd></pvd> <d>una mutación errónea que destruye a la adenilciclasa, interrumpiendo la síntesis de l AMPc</d> .
1	<d1>Los bacteriófagos</d1> <pvd><vd>denominados</vd> </pvd> <t>fagos T par</t> <d2>son virus que contienen ADN que infecta a las E. coli<.d2>.
2	<d>El material de los cromosomas</d> <pvd><pr>se</pr> <vd>nombra</vd></pvd> <t>cromatina</t>.
2	<t>El Mycobacterium tuberculosis</t> <pvd><vd>es</vd> </pvd> <d>un bacilo inmóvil, aerobio de 0'8 a 4 micras de longitud; se tiñe con dificultad, pero una vez teñido resiste a la decoloración ante los ácidos fuertes y el alcohol</d>.
3	<d>Esto</d> <pvd><pr>se</pr> <vd>conoce</vd> <nx> como</nx></pvd> <t>mutación</t>.
3	<d>NULL</d> <pvd><pr>Se</pr> <aux>han</aux> <vd>definido</vd> <nx>como</nx></pvd> <t>" talofitas eucariotas aclorofílicas"</t>.

Tabla 5.34. Ejemplos de definiciones etiquetadas con reglas de ranking

En la tabla anterior observamos que el primer contexto incluye una serie de información que podría considerarse más completa ya que introduce el género próximo y la diferencia específica de un término, y que fue reconocida por la introducción de la conjunción *que* seguida de un verbo en forma conjugada.

En el segundo caso, observamos un ejemplo de un verbo que permite el término en la posición derecha y la información definitoria aparece antes y después del patrón verbal.

El tercer ejemplo constituye un caso donde la definición está formada por menos de 5 palabras y se considera con el segundo valor principalmente para hacer una distinción de aquellos casos que podrían considerarse más completos.

En el cuarto ejemplo vemos que el punto y coma corta el contenido de la definición e introduce otro tipo de información que, si bien podría considerarse pertinente para entender el significado del término, rompe con el esquema estructural de lo que podría ser una definición prototípica. Estos dos casos anotados con el valor 2 son igualmente útiles en la descripción del significado de un término, pero se consideran de esta forma con el fin de reorganizar los resultados y presentar los que contienen más información en primera posición.

La definición del quinto caso está formada únicamente por un pronombre, lo cual nos hace suponer que la verdadera definición está en un párrafo anterior.

En el último ejemplo, la definición se etiqueta como la posición vacía a la izquierda, y en este caso se presenta un término que igualmente puede estar definido en una oración o párrafo anterior.

Ahora bien, esta idea de aplicar las reglas para asignar los valores que hemos detallado tiene la finalidad de obtener una votación general sobre cada candidato, que nos ayude a determinar la reorganización en un gradiente donde primero se presenten los casos más probables a ser buenos CDs.

En la siguiente tabla podemos ver ejemplos de cada combinación de resultados. En la columna de los valores primero se expresa el número asignado al término y después el que fue asignado a la definición.

T D	Ejemplos
1 – 1	<t>La penetrancia de un genotipo</t> <pvd><pr>se</pr> <vd>define</vd> <nx>como</nx></pvd> <d>el porcentaje de individuos con este genotipo que desarrollan evidencias o síntomas (penetrancia fenotípica) debido a la enfermedad (penetrancia de la enfermedad)</d>.
1 – 2	<t>La enzima</t> <pvd><vd>es</vd></pvd> <d>una proteína de transmembrana</d>.
1 – 3	<d>NULL</d> <pvd><pr>Lo</pr> <vd>denominaron</vd> </pvd> <t>factor promotor de la mitosis o FPM</t>.
2 – 1	Se cree que la reacción alérgica que se produce en el asma alérgica ocurre de la siguiente manera: la persona alérgica típica tiene <t>tendencia a</t> <pvd><vd>formar</vd> cantidades anormalmente elevadas <nx>de</nx></pvd> <d>anticuerpos IgE , y estos anticuerpos producen reacciones alérgicas cuando reaccionan con sus antígenos específicos, como se explica en el Capítulo 34</d>.
2 – 2	Cómo podía sintetizar <t>proteínas con formas diferentes, para</t> <pvd><aux>ser</aux> <vd>usadas</vd> <nx>como</nx></pvd> <d>enzimas ?</d>
2 – 3	<d1>NULL</d1> <pvd><pr>Se</pr> <vd>denomina</vd> </pvd> <t>dNTP a las formas de desoxirribosa de los cuatro nucleósidos trifosfato</t> ; <d2>de l mismo modo , dNMP hace referencia a las formas monofosfato</d2>.
3 – 1	<t><anf>Esta biblioteca</anf></t> <pvd><vd>es</vd> </pvd> <d>un conjunto completo de clones que <pvd><vd>contienen</vd></pvd> la totalidad de l genoma</d>.
3 – 2	El grupo concentró su atención en la región Xq28 del cromosoma X; <t><anf>ese segmento</anf></t> <pvd><vd> contiene</vd></pvd> <d>cientos de genes</d> .
3 – 3	<d1>NULL</d1> <pvd><pr>Se</pr> <vd>nombran</vd> </pvd> <t>atendiendo a la pentosa</t>, <d2>la base nitrogenada y el número de grupos fosfato que presenten , por ejemplo: adenosina trifosfato (ATP), deoxicitosina monofosfato (dCMP), etc</d2>.

Tabla 5.35. Ejemplos de CDs con ranking

Vemos que está organizada de acuerdo con los valores numéricos del término, los tres primeros ejemplos corresponden a casos donde al término se le asignó el valor 1. En el primer ejemplo, este valor también es asignado a la definición, y podemos observar un CD con una forma más prototípica: el término no está constituido por

alguno de los elementos de las reglas, como un pronombre, y la definición contiene la conjunción *que* seguida de un verbo conjugado, la cual puede ser una estructura prototípica para introducir la diferencia específica. En el segundo ejemplo, a la definición se le asigna el valor 2 debido a su tamaño. En el tercer ejemplo, la definición corresponde a un espacio vacío y se le asigna el valor 3.

Por otro lado, los siguientes tres casos corresponden a contextos donde el término tiene el valor 2. El primero de ellos es reconocido de esta forma por la inclusión de una preposición justo antes de la etiqueta de cierre de término, aunque en este caso, a la información de la definición se le asigna el valor 1. El término del segundo ejemplo es considerado igualmente con el valor 2 por la preposición detrás de la etiqueta de cierre, y la definición también se considerada con este valor por su extensión. En el tercer ejemplo de este grupo se considera con el valor 2 al término porque no comienza con una expresión regular de término, sino con un elemento que no fue reconocido como sustantivo en las etiquetas POS originales del corpus, sino como un elemento desconocido *dNTP/W*; en este caso la definición tiene el valor 3 porque corresponde a un espacio vacío.

El último grupo lo conforman los ejemplos donde el término es anotado con el valor 3. Aquí, el primer término es una anáfora, aunque la definición cumple con las reglas del valor 1. El segundo caso corresponde igualmente a una anáfora, pero la definición es marcada con el valor 2 por su extensión. Por último, el tercer caso anota la definición con 3 por estar vacía, y el término se considera con este mismo valor por una regla que identifica los casos que contienen un verbo en gerundio.

Una vez obtenido un valor para cada término y definición, incluimos un módulo que se encarga de hacer un ranking global a partir de las combinaciones encontradas. Para ello, decidimos establecer un umbral para filtrar posibles contextos no relevantes que hayan sido clasificados como CDs. Este umbral lo establecimos para los casos donde al término o a la definición se les hubiera asignado el valor 3, tomando en cuenta que en estos casos la información proporcionada en cada elemento era de bajo nivel. Así,

con el umbral se filtraron las combinaciones: 1 – 3, 2 – 3, 3 – 1, 3 – 2 y 3 – 3.

Finalmente, con este ranking global se asignaron nuevos valores tomando en cuenta las siguientes reglas:

- Si el valor de la combinación es 1–1, el valor global es 1.
- Si el valor es 1–2 o 2–1, el valor global es 2.
- Si el valor es 2–2, el valor global es 3.
- Las combinaciones filtradas en el umbral equivalen a 4.

Algunos resultados los podemos ver en la siguiente tabla:

Valor	Ejemplo
1 - CD	<t>La OLA</t> <pvd><vd>es</vd></pvd> <d>una técnica adecuada para las pruebas automatizadas a gran escala, ya que no implica centrifugación o electroforesis</d>.
2 - CD	<d>Este enzima multifuncional</d> <pvd><pr>se</pr> <vd>llama</vd></pvd> <t>CAD</t>.
3 - CD	La idea <pvd><aux>es</aux> <vd>considerar</vd> <t>cada registro</t> <nx>como</nx></pvd> <d>un tipo básico en la codificación ; una secuencia de bits independiente codifica el tipo de cada campo de l registro</d>.
4 - NR	<t>Las unidades más frecuentes</t> <pvd><vd>incluyen </vd></pvd> <d>tres o cuatro nucleótidos de guanina</d>.

Tabla 5.36. Ejemplos de contextos con ranking global

En la tabla vemos que los valores globales que permanecieron como CDs son 1, 2 y 3, que en este orden expresan la probabilidad de ser mejores candidatos. Los contextos con el valor global 4 fueron considerados contextos no relevantes. En esta categorización, vemos cómo el mejor resultado lo obtiene un contexto que contiene información más completa, como es el caso de la definición del término OLA, donde se introduce su género próximo e información sobre sus diferencias específicas. En segundo lugar aparece un contexto donde se esclarece únicamente un probable género próximo del término, es decir, en este caso podría saberse por lo menos que el término *CAD* es una *enzima multifuncional*, pero no se especifican sus características distintivas. En el tercer puesto se encuentra un contexto donde al término y a la definición se les asigna el valor 2; en el caso del término, porque comienza con un

determinante que forma parte de un grupo definido en las reglas, donde se encuentran *cada*, *algunos*, *otros*, por ejemplo, para representar casos que conforman un matiz de menor estabilidad y fuerza en comparación con los determinantes *el, la, un, una*, etc.; en el caso de la definición, ésta se votó con el valor 2 porque incluye un punto y coma que probablemente esté introduciendo un corte a partir del cual se presenta otro tipo de información no relacionada con la definición. Finalmente, el cuarto caso se excluye por el valor 3 del término, ya que tiene dentro de su estructura sintáctica el adverbio *más*, considerado como una regla para notar construcciones ajenas a las estructuras comunes de los términos.

Para finalizar este apartado consideramos necesario hacer algunas aclaraciones pertinentes.

- La inclusión de esta metodología de ranking tiene la finalidad de reorganizar los datos y presentar en primer lugar aquellos contextos que probablemente sean mejores candidatos a CDs, así como excluir aquellos estableciendo un umbral de relevancia.
- No obstante, es preciso aclarar que en muchas ocasiones es difícil identificar, únicamente por medio de reglas lingüísticas, aquellos contextos que no presentan información definitoria o conceptual sobre un término. Las expresiones regulares expuestas para identificar el término y la definición pueden recuperar cualquier otro tipo de elemento discursivo que no sea parte de un CD, y que estructuralmente cumplen con todas las reglas de los buenos candidatos. Por ejemplo, el sistema identifica como términos con valor 1 los siguientes casos: *repetitivo*, *el término*, *una vez*, *la molestia de la ropa interior teñida*.
- También debe tomarse en cuenta que el ranking se basa únicamente en el análisis de la estructura de lo etiquetado automáticamente en el proceso de identificación de términos y definiciones, por lo cual, si el etiquetado previo fue erróneo, entonces se corre el riesgo de que el ranking asigne un valor menor.

5.7 Recapitulación

Ahora bien, a modo de recapitulación queremos sintetizar algunas ideas que surgen a partir de lo expuesto a lo largo de este capítulo. A grandes rasgos, hemos presentado un sistema para la extracción de CDs, el cual constituye un esfuerzo por incorporar una metodología basada en reglas lingüísticas para la extracción automática de información definitoria en textos especializados en español.

El sistema parte de la búsqueda de patrones verbales asociados a diferentes tipos de definición, los cuales se generan a partir de ciertos parámetros que se especifican en una gramática, específicamente los verbos definitorios por cada tipo de definición, las restricciones en tiempos verbales por cada uno de los verbos, así como restricciones de distancia entre éstos y sus respectivos nexos, y los patrones contextuales de cada verbo, es decir las posiciones en las que puede aparecer el término en el CD.

Con esta metodología, el sistema permite modificar la configuración de la gramática, ya que el usuario puede agregar nuevos patrones a la vez que cambiar los parámetros de cada uno de ellos.

También hemos hecho mención de que el algoritmo consta de un proceso para la exclusión de contextos no relevantes, es decir, aquellos candidatos donde se identifica un patrón verbal pero éste no introduce información definitoria sobre un término. Este filtro se lleva a cabo a partir de una serie de reglas que se especifican igualmente en una gramática, con lo que cualquier modificación es posible sin alterar el código del sistema.

Una vez filtrados los contextos no relevantes, se identifica en los candidatos restantes los posibles términos y definiciones a partir de un árbol de decisión. Y por último, el sistema contiene un proceso para analizar los resultados y tratar de identificar aquellos que presenten estructuras sintácticas más prototípicas de términos y definiciones, lo cual sirve a su vez como un último filtro para excluir contextos no definitorios.

6. Evaluación del ECODE

En el capítulo anterior describimos un sistema basado en reglas lingüísticas para la extracción de CDs denominado ECODE (Extractor de Contextos Definitorios). Detallamos que este sistema está compuesto por diferentes módulos: el primero de ellos se encarga de buscar ocurrencias de patrones definitorios, específicamente patrones verbales; el segundo aplica un filtro para tratar de eliminar contextos no relevantes; el tercer módulo se encarga de identificar los elementos constitutivos de los candidatos a CDs; por último, el cuarto módulo realiza un ranking de los resultados obtenidos automáticamente.

Con el fin de tener un panorama general del desempeño del sistema, en este capítulo realizaremos una evaluación global junto con una serie de evaluaciones específicas. En primer lugar describiremos el corpus utilizado para evaluar el sistema (6.1). Posteriormente detallaremos la metodología de evaluación y los índices utilizados (6.2). Enseguida detallaremos los resultados de precisión y cobertura obtenidos: de manera global (6.3), por tipo de patrón verbal (6.4) y según diferentes tipos de restricciones (6.5). Finalmente explicaremos algunas conclusiones de la evaluación (6.6).

6.1 Corpus de evaluación

Para desarrollar la evaluación del ECODE, en esta etapa recurrimos igualmente al Corpus Técnico del IULA en español. Tomando en cuenta que el objetivo de nuestra evaluación es presentar un panorama del funcionamiento de un sistema basado en patrones cuyo núcleo es un verbo definitorio, decidimos conformar un subcorpus a partir de la búsqueda de los lemas de los verbos contenidos en nuestra gramática de patrones.

Para ello, buscamos cada uno de los verbos o patrones verbales mediante la opción de búsqueda por lema que se puede realizar con el CQP[1], específicamente en el área de medicina. Los resultados numéricos los podemos ver en la siguiente tabla:

[1] Ver el apartado 5.1.

Tipo	Lema	Nexo	Ocurrencias
Analítica	(es|son)	determinante	250
	caracterizar	como, por	47
	concebir	como	19
	considerar	como	250
	describir	como	249
	definir	como	250
	entender	como	59
	conocer	como	250
	denominar	como	9
	llamar	como	1
	denominar	∅	250
	llamar	∅	250
	nombrar	∅	42
Extensional	comprender	∅	250
	contener	∅	250
	incluir	∅	250
	integrar	∅	250
	constar	de	250
	contar	con	199
	formar	de, por	250
	componer	de, por	250
	constituir	de, por	250
Funcional	permitir	∅	250
	encargar	de	130
	consistir	en	250
	funcionar	como, para	142
	ocupar	como, para	47
	servir	como, en, para	250
	usar	como, en, para	250
	emplear	como, en, para	250
	utilizar	como, en, para	250
Sinonímica	conocer	también	42
	denominar	también	54
	llamar	también	92
Total			6132
Total líneas no repetidas			5809

Tabla 6.1. Patrones de búsqueda para la evaluación del ECODE

Para recuperar los patrones que permiten un nexo, se incluía una ventana de 15 palabras entre dicho nexo y el verbo definitorio. Por ejemplo, para obtener ocurrencias con el verbo *definir* se utilizaba la siguiente expresión: *[lemma="definir"] [word!= "como"]{0,15}*[2] *[word="como"]*.

De los resultados obtenidos para cada lema se seleccionaron las primeras 250 ocurrencias, independientemente de que la expresión de búsqueda obtuviera más o menos contextos. Dos de los patrones (*nombrar como* y *nombrar también*) no recuperaron ningún caso, mientras que la mayoría de ellos recuperaron las 250 ocurrencias. El lema *ser* se buscó en la tercera forma de singular y plural con la finalidad de igualar a las formas que se expresan en la gramática de patrones definitorios. Observamos también que algunos patrones incluyen verbos que pueden o no llevar un nexo, por lo cual se duplicó la expresión de búsqueda, una para los patrones con nexo y otra para los patrones sin nexo. En este caso, el único preprocesamiento que se realizó en el corpus fue eliminar las líneas repetidas, con lo cual el total de 6132 ocurrencias iniciales se redujo a 5809 contextos.

6.2 Metodología de evaluación

El primer paso en el proceso de la evaluación fue el análisis manual de las ocurrencias para decidir cuáles de ellas eran o no CDs. En este proceso se recurrió a la ayuda de 3 evaluadores, cada uno de los cuales analizó una parte del corpus. Se escogió a personas involucradas en el área de terminología esperando tener un criterio más *amplio* a la hora de decidir si un contexto introducía o no información definitoria relevante para la comprensión de un término. El único requisito que se les pidió fue que consideraran como CDs a aquellos casos donde se introdujeran explícitamente un término y una definición, tomando en cuenta los distintos tipos de CDs con los que trabaja el sistema: analíticos, extensionales, funcionales y sinonímicos.

[2] En este caso, la expresión *[word!="como"]* indica que se busque cualquier palabra excepto el nexo.

Como resultado de este análisis manual obtuvimos el corpus de evaluación, al cual se le añadieron las etiquetas de <CDs/>, para el caso de los contextos definitorios, y de <NRs/>, para el caso de los contextos no relevantes. Este corpus formó el input de entrada del ECODE, sobre el cual aplicamos los índices de precisión y cobertura para evaluar distintas facetas de los resultados.

La precisión y cobertura se emplean comúnmente en la evaluación de sistemas de recuperación y extracción de información. Como explican Jurafsky y Martin (2000), la precisión es una medida que se utiliza para determinar cuánta información extraída automáticamente por el sistema es correcta, mientras que la cobertura es una medida para saber cuánta de la información relevante en el texto fue extraída automáticamente. Estos índices se suelen representar mediante las siguientes fórmulas:

$$\text{Precisión} = \frac{\text{número de respuestas válidas propuetas por el sistema}}{\text{número de respuestas propuestas por el sistema}}$$

$$\text{Cobertura} = \frac{\text{número de respuestas válidas propuetas por el sistema}}{\text{número total de respuestas en el texto}}$$

Ahora bien, pensando en el escenario de la extracción de CDs, estas fórmulas las podemos interpretar de la siguiente manera:

$$\text{Precisión} = \frac{\text{número de CDs válidos propuetos por el sistema}}{\text{número de CDs propuestos por el sistema}}$$

$$\text{Cobertura} = \frac{\text{número de CDs válidos propuetos por el sistema}}{\text{número total de CDs en el corpus}}$$

Cabe señalar que los resultados de precisión y cobertura tienden a ser proporcionalmente inversos, entre más alta sea la precisión, más baja será la cobertura y viceversa. En estos casos un índice más cercano al 1 indica mejores resultados.

A continuación presentamos los resultados de precisión y cobertura. En primer lugar detallaremos los índices obtenidos de manera

global (6.3). En segundo lugar expondremos los índices por cada uno de los patrones verbales asociados a diferentes tipos de definición (6.4). Finalmente, en tercer lugar señalaremos los resultados obtenidos a partir de diferentes tipos de restricciones (6.5).

6.3 Resultados globales de precisión y cobertura

Como señalamos en el capítulo anterior, la gramática de patrones verbales que utilizamos para desarrollar el sistema no incluía ninguna restricción específica. Por ejemplo, se consideró que la distancia entre el verbo y el nexo podía ser cualquier número de palabras, o bien se tenía en cuenta que los verbos definitorios podían emplearse en cualquier tiempo y forma gramatical.

Tomando en cuenta esta gramática sin restricciones, los primeros resultados numéricos que se encontraron fueron los siguientes:

CDs en el corpus	CDs propuestos por el sistema	CDs válidos propuestos por el sistema
2254	3309	1783

Tabla 6.2. Resultados globales obtenidos con la gramática sin restricciones

La precisión se obtuvo dividiendo el número de CDs válidos propuestos por el sistema sobre el número de CDs propuestos por el sistema (1783/3309), mientras que la cobertura se obtuvo dividiendo el número de CDs válidos propuestos por el sistema sobre el número de CDs en el corpus (1783/2254). De esta forma, los primeros resultados globales fueron los siguientes:

P	C
0.53	0.79

Tabla 6.3. Resultados globales de precisión y cobertura del ECODE con la gramática sin restricciones

De manera general, los índices muestran que se obtuvo una mejor cobertura frente a la precisión. Esto quiere decir que aproximadamente el 80% de CDs presentes en el corpus fueron

clasificados válidamente, mientras que poco menos del 50% de lo clasificado como CDs era ruido, es decir, contextos que los evaluadores no consideraron válidos. En otras palabras, los resultados de la precisión indican que varios contextos propuestos por el sistema no eran CDs anotados por los evaluadores, mientras que en la cobertura el resultado indica que algunos CDs anotados manualmente no fueron clasificados como tal por parte del sistema.

En la siguiente tabla presentamos algunos ejemplos de CDs que el sistema clasificó válidamente:

Tipo	Ejemplo
Analítica	Hasta el momento <pvd><aux>hemos</aux> <vd> descrito</vd> <t v="1">la afasia de Broca</t> <nx> como</nx></pvd> <d v="1">un trastorno de la producción del habla.</d>
Extensional	<t v="1">Una molécula de anticuerpo</t> <pvd> <vd>consta</vd> <nx>de</nx></pvd> <d v="1">dos cadenas ligeras idénticas y dos cadenas pesadas idénticas.</d>
Funcional	<t v="1">El método de Maxam y Gilbert</t> <pvd><aux>fue</aux> <vd>utilizado</vd> <nx>para </nx></pvd> <d v="1">determinar la secuencia completa de nucleótidos en el DNA del virus del mono, SV 40.</d>
Sinonímica	<d1 r11="2">El ímpetu,</d1> <pvd><vd>conocido </vd> <nx>también/D</nx></pvd> como <t v="1"> cantidad de movimiento</t>, <d2>resulta especialmente útil en el estudio de las colisiones.</d2>

Tabla 6.4. Ejemplos de CDs válidos propuestos por el ECODE

El primer contexto es un ejemplo de CD analítico donde se define el término *afasia de Broca* y se especifica que es un tipo de *trastorno*, específicamente un *trastorno de la producción del habla*. En este caso vemos que el árbol de decisión identifica correctamente al término en posición derecha del patrón verbal, seguido de la definición también en posición derecha. Por su parte, el segundo ejemplo presenta información sobre las partes que componen al término *molécula de anticuerpo*, en donde el árbol identifica que el término está en posición izquierda y la definición en derecha. El tercer ejemplo corresponde a un CD de tipo funcional, donde se

establece para qué fue utilizado un método denominado *método de Maxam y Gilbert*. Por último, el caso del CD sinonímico introduce el término *ímpetu* y su sinónimo *cantidad de movimiento*. En este ejemplo el sistema clasifica la definición en dos partes, la primera es lo que está a izquierda del patrón verbal y la segunda lo que está a la derecha del término.

Por otra parte, en la siguiente tabla presentamos ejemplos de los contextos NRs que el sistema excluyó válidamente, entre los que se encuentran contextos excluidos en el ranking, excluidos en el árbol de decisión, y filtrados por las reglas de contextos no relevantes.

Tipo	**Ejemplo**
Excluidos en el ranking	Al acercarse a un problema, el médico comienza con conocimientos médicos pertinentes y sintetiza <t v="1">la información en un concepto</t> <pvd><vd> integrado</vd></pvd> <d v="3">NULL.</d>
	<t v="3"><anf>Este apartado</anf></t> <pvd><vd> incluye</vd></pvd> <d v="1">trastornos sexuales no clasificables en ninguna de las categorías antecedentes.</d>
Excluidos por el árbol de decisión	<izq>Como tal</izq> <pvd><aux>puede</aux> <vd> considerar</vd> <nx>se en gran medida</nx> <nx> como</nx></pvd> <der>un conato de laboratorio .</der>
	<null>NULL</null> <pvd><pr>Se</pr> <vd>encarga </vd> <nx>de</nx></pvd> <der>el estudio de la enfermedad considerando los mecanismos causales, los mecanismos lesivos y la relación con las alteraciones fisiopatológicas y su expresión sintomática.</der>
Filtrados	En este caso conocemos las velocidades inicial y final, <filtro>así como</filtro> la aceleración.
	Postuló la existencia de dos clases de receptores, alfa y beta, definidos sobre base de la potencia de los agonistas, <filtro>como se</filtro> muestra a continuación:

Tabla 6.5. Ejemplos de NRs válidos propuestos por el ECODE

En estos ejemplos, el primer caso fue excluido por el ranking ya que se asignó un valor 3 a lo clasificado automáticamente como definición, y que correspondía a un espacio vacío. En el segundo ejemplo también se le asignó al término el valor 3, ya que éste

presenta una secuencia introducida por un pronombre que puede indicar una posible anáfora, y por tanto el término no está presente de manera explícita. Por otro lado, el tercer contexto, que contiene el patrón verbal analítico *considerar como*, fue excluido por el árbol de decisión, ya que su estructura sintáctica no dio un valor positivo al momento de tratar de identificar expresiones regulares de término o definición. Asimismo, el cuarto caso fue excluido por el árbol porque el verbo definitorio *encargar* permite al término únicamente en posición izquierda, y en este caso se encontró una posición vacía. Finalmente, en el quinto ejemplo se presenta un contexto filtrado a partir de la regla del patrón de exclusión *así como*, mientras que en el sexto ejemplo la regla de contexto no relevante que se cumple es la de *nexo* seguido del pronombre *se*, en este caso *como se*.

Con lo expuesto hasta ahora hemos presentado algunos ejemplos de los CDs y NRs clasificados correctamente por el ECODE. Ahora bien, a partir de los resultados de precisión y cobertura también podemos observar que el sistema clasificó erróneamente algunos contextos.

En el caso de la precisión, se encontró un total de 1526 contextos que el sistema propone como CDs pero que no fueron considerados por los evaluadores. Algunos ejemplos son los siguientes:

Tipo	**Ejemplo**
Analítica	<d v=“2”>Este proceso</d> <pvd><pr>se</pr> <vd> denomina</vd></pvd> <t v="1">enlace peptídico</t>.
Extensional	<t v="1">La escala de medición</t> <pvd><vd> comprende</vd></pvd> <d v="1">los siguientes niveles :</d>
Funcional	<t v="1">Los lactantes del grupo II</t> <pvd><vd> sirvieron</vd> <nx>como</nx></pvd> <d v="2"> controles.</d>
Sinonímica	<d v="1">En primer lugar, debemos conocer por supuesto qué gen examinar, y</d> <pvd><aux>debemos/VDR1P- </aux> <vd>conocer </vd> <nx>también/D </nx></pvd> <t v="1">la secuencia "normal" (silvestre)</t>.

Tabla 6.6. Ejemplos de NRs propuestos por el ECODE como CDs

El primer ejemplo fue catalogado como CD a partir del valor 2 asignado a la definición, es decir, no fue excluido en el ranking; sin embargo, en este caso los evaluadores no lo consideraron como un CD. El segundo ejemplo es clasificado como un CD extensional, ya que cumple con las reglas del árbol de decisión; no obstante, lo etiquetado como definición no corresponde a información definitoria sobre el término. En el mismo sentido, la estructura sintáctica del tercer ejemplo cumple con las reglas del sistema, pero en este caso el contenido semántico del contexto no aporta información funcional sobre un término. Por último, en el cuarto ejemplo el sistema propone un CD sinonímico a partir del patrón *conocer también*, pero claramente no se está introduciendo un sinónimo.

En el caso de la cobertura, hubo 471 CDs que el sistema no identificó como tales sino como NRs. Algunos ejemplos de estos casos los podemos ver en la tabla 6.7.

Tipo	**Ejemplo**
Analítica	<d v="1">EL tejido esqueletógeno se desarrolla a partir del borde inferior de la pared interna del somita y esta parte del somita</d> <pvd> <pr>se</pr_> <vd lema="llamar">llama</vd></pvd> <t v="3">por tanto, esclerotoma</t> (fig. 353).
Extensional	Después de la neurulación, <t v="3">la epidermis de los anfibios , y también la de otros muchos vertebrados ,</t> <pvd><aux>está</aux> <vd>formada</vd> <nx>por </nx></pvd> <d v="1">dos capas de células.</d>
Funcional	<izq>Además de</izq> <pvd><vd>servir</vd> <nx> como</nx></pvd> <der>fuentes de energía, los azúcares también son componentes de glucoproteínas, glucopéptidos y glucolípidos que desempeñan papeles estructurales y funcionales.</der>
Sinonímica	<d v="2">La reductasa</d> <pvd><nx>también</nx> <pr>se</pr> <vd>conoce</vd></pvd> como <t v="3">la ferro-proteína (Pe proteína), y la nitrogenasa como la molibdeno-ferroproteína (MoFe proteína)</t>.
Filtrados	Algunos pacientes sufren, 20 o 30 años después, una debilidad muscular progresiva, por afectación de las neuronas motoras supervivientes, <filtrado>denominada <filtro>*atrofia</filtro> muscular pospolineuritis.

Tabla 6.7. Ejemplos de CDs propuestos por el ECODE como NRs

Observamos en el primer ejemplo que el CD fue filtrado erróneamente como NR en el ranking de término, al cual se le asignó el valor 3 por una regla que determina que la estructura sintáctica de un término no puede contener un adverbio. Lo mismo sucedió en el segundo ejemplo, donde al término se le asignó un valor 3 por contener un adverbio, pero los evaluadores consideraron como un CD con información extensional. En el tercer caso, este CD se filtró como NR ya que el árbol de decisión no incluye una regla para identificar casos donde el término aparece en posición derecha después de la definición. En este ejemplo, se expresa que *los azúcares* sirven como *fuente de energía*. En el contexto con información sinonímica, el sistema identificó como término toda la secuencia de palabras a la derecha del patrón verbal y clasificó esta secuencia con un valor 3 debido al adverbio *como*. Finalmente, el CD filtrado como NR se debió a la regla que excluye los patrones verbales que van seguidos de un verbo en forma conjugada, aunque en este caso el error se debió a un mal etiquetamiento de las categorías gramaticales en el corpus del IULA, ya que el sustantivo *atrofia* fue considerado como un verbo.

En resumen, con estos resultados globales de precisión y cobertura hemos visto algunos ejemplos de los aciertos y errores del sistema. Hemos observado que, con la gramática sin restricciones, el ECODE identifica correctamente casi el 80% de CDs anotados manualmente por los evaluadores. No obstante, casi el 50% de los contextos clasificados automáticamente como CDs no fueron reconocidos manualmente como válidos. Ahora bien, en el siguiente apartado presentaremos resultados específicos según cada patrón verbal.

6.4 Resultados de precisión y cobertura por patrón verbal

Con el fin de tener un panorama más específico de los resultados globales obtenidos por el sistema, en este apartado detallamos los índices de precisión y cobertura de cada uno de los patrones verbales que conforman nuestra gramática de patrones definitorios. Para obtener estos índices utilizamos únicamente los contextos con un sólo patrón, tomando en cuenta el hecho de que en ocasiones algunos contextos pueden contener más de un patrón definitorio y

por tanto estar asociados a más de un tipo distinto de definición. En estos casos, el ECODE puede reconocer los patrones y los tipos de información definitoria a los que están asociados; sin embargo, cuando son más de uno y alguno de ellos es excluido en el ranking y el otro no, entonces entra en conflicto y no puede reconocer cuál de ellos es el que corresponde a la regla por la cual se identificó dicho contexto como un CD.

Tomemos el siguiente contexto para ejemplificar lo anterior:

- <tipoD="ext"/><tipoD="ana"/>En 1897, J. J. Thomson (1856-1940) demostró que <t v="3"><anf>esta corriente</anf></t> <pvd><pr>se</pr> <vd>compone</vd> <nx>de</nx></pvd> <d v="1"><d v="2">partículas cargadas negativamente</d> <pvd> <vd>denominadas</vd></pvd> <t v="1">electrones</t>.</d>

Este CD que incluye dos verbos, el primero de tipo extensional y el segundo analítico, con los patrones *se compone de* y *denominadas*, respectivamente. Dentro del patrón extensional se identifica que el término es *esta corriente* y la definición es *partículas cargadas negativamente*. Por su parte, en el caso del patrón analítico el término es *electrones* y su definición es *partículas cargadas negativamente*. Observamos que el término *esta corriente* obtiene una votación de valor 3 y se considera como posible anáfora. No obstante, el término electrones es votado con el valor 1 y por tanto el sistema lo considera como un buen candidato. Lo que no reconoce el ECODE es que el término con votación 1 pertenece al patrón *denominada* y por tanto a un tipo de definición analítica.

Es por esta razón que decidimos considerar para esta parte de la evaluación sólo aquellos contextos que tuvieran un sólo patrón definitorio y que nos permitieran estar seguros de que el sistema los está reconociendo como CDs o NRs a partir de un patrón asociado a un tipo de definición específica.

Así, de las 5809 ocurrencias con verbos definitorios que forman nuestro corpus de evaluación, en esta etapa se utilizó un total de 4799 ocurrencias correspondientes a los casos con un sólo verbo definitorio.

En la siguiente tabla mostramos los índices de precisión y cobertura en el caso de los contextos asociados a definiciones analíticas, extensionales, funcionales y sinonímicas.

Tipo de CDs	P	C
Analíticos	0.58	0.83
Extensionales	0.48	0.77
Funcionales	0.45	0.83
Sinonímicos	0.76	0.85

Tabla 6.8. Resultados de precisión y cobertura del ECODE por tipo de patrón definitorio

Vemos, a grandes rasgos, que los patrones de definiciones sinonímicas fueron los que mejores índices obtuvieron, los cuales consistieron en 76% de precisión y 85% de cobertura. El grupo que siguió en el orden de mejores resultados fue el de los patrones analíticos; en este caso, la cobertura se mantiene alta mientras que la precisión bajó al 58%. Por el contrario, los patrones de tipo extensional y funcional recuperaron valores de precisión por debajo del 50%; en el caso de los funcionales, la cobertura fue igualmente del 83%, mientras que en el caso de los patrones extensionales fue más baja, con 77%.

Lo anterior quiere decir, en el caso de la precisión, que los patrones extensionales y funcionales recuperaron mayor ruido, ya que poco más del 50% de lo recuperado automáticamente no eran CDs válidos anotados por los evaluadores. Por su parte, en el caso de la cobertura, todos los patrones recuperaron índices por arriba del 77%, lo cual indica que el sistema no identifica un 23% de CDs que los evaluadores sí clasificaron manualmente. Más adelante ahondaremos en un análisis de los resultados obtenidos en comparación con los resultados de otras metodologías para la extracción de información definitoria que hemos reseñado en el capítulo 3.

En la siguiente tabla podemos ver los resultados del grupo de patrones verbales asociados a definiciones analíticas.

Patrones Analíticos		
Patrón verbal	**P**	**C**
(es|son) + det.	0.60	0.78
caracterizar como	0.37	0.75
caracterizar por	0.66	1
concebir como	0.50	0.66
Considerar como	0.57	0.80
describir como	0.65	0.85
definir como	0.85	0.87
entender como	0.69	0.69
conocer como	0.48	0.95
denominar como	0.40	0.66
llamar como	1	1
denominar	0.62	0.89
llamar	0.62	0.78
nombrar	0.14	1

Tabla 6.9. Resultados de precisión y cobertura del ECODE para patrones verbales analíticos

En general, vemos que tres patrones obtuvieron el óptimo resultado en el caso de la cobertura: *caracterizar por*, *llamar como* y *nombrar*. En el caso de estos últimos dos, aunque la precisión del primero también fue de 1, cabe destacar que este patrón recuperó únicamente una ocurrencia; mientras que en el caso de *nombrar*, la precisión fue de apenas 0.14. Observamos también que 4 de los 14 patrones obtuvieron una precisión por debajo de 0.50, mientras que los valores más bajos en el caso de la cobertura no fueron menores a 0.65. A grandes rasgos, lo anterior quiere decir que con la gramática sin restricciones, el ECODE logra identificar una cantidad considerable de CDs a partir de patrones verbales, aunque algunos de ellos anotados manualmente por los evaluadores son identificados como contextos no relevantes por el sistema . Por su parte, estos resultados demuestran que la misma gramática sin restricciones está recuperando una cantidad notable de ruido, es decir, contextos que no son CDs pero que el sistema clasifica automáticamente como si lo fueran.

En el caso de los patrones extensionales, se encontraron los siguientes resultados:

Patrones Extensionales		
Patrón verbal	**P**	**C**
comprender	0.25	0.82
contener	0.64	0.73
incluir	0.42	0.54
integrar	0.18	0.69
constar de	0.73	0.91
contar con	0.26	0.61
formar de	0.37	0.83
formar por	0.83	0.76
componer de	0.60	0.75
componer por	0.72	0.85
constituir de	0.33	1
constituir por	0.40	0.81

Tabla 6.10. Resultados de precisión y cobertura del ECODE para patrones verbales extensionales

Aquí se puede ver que los valores más altos se obtuvieron en la cobertura de los patrones *constituir por*, *constar de* y *componer por*, los cuales estuvieron por arriba de 0.85. De igual forma, en el caso de la precisión los valores más altos fueron para *formar por*, *constar de* y *componer de*, en los cuales se obtuvieron índices por arriba de 0.70, aunque no mayores a 0.85. Los valores más bajos los encontramos en la precisión de *integrar* y *comprender*, por abajo de 0.25. De igual forma que en los patrones analíticos, vemos cómo la cobertura obtiene mejores resultados que la precisión y cómo el sistema está considerando varios contextos no relevantes como CDs. Asimismo, la mayoría de los CDs fueron clasificados correctamente, lo cual indica que el sistema puede identificar con efectividad los patrones definitorios de cada contexto a partir de la gramática y de las reglas de identificación de verbos definitorios y patrones verbales.

Para el caso de los patrones de tipo funcional, los resultados obtenidos los presentamos en la siguiente tabla:

Patrones Funcionales		
Patrón verbal	**P**	**C**
permitir	0.68	0.73
encargar de	0.29	0.74
consistir en	0.44	0.79
funcionar como	0.63	0.90
funcionar para	0.50	0.66
ocupar como	0.20	1
ocupar para	0.50	0.66
servir como	0.66	0.80
servir en	0.25	1
servir para	0.47	0.77
usar como	0.84	1
usar en	0.37	0.68
usar para	0.73	0.84
emplear como	0.42	1
emplear en	0.13	0.85
emplear para	0.29	0.82
utilizar como	0.38	0.88
utilizar en	0.33	0.90
utilizar para	0.46	0.76

Tabla 6.11. Resultados de precisión y cobertura del ECODE para patrones verbales funcionales

En este caso obtenemos una lista mayor a partir de las combinaciones que permiten algunos verbos con diferentes partículas gramaticales. Resulta interesante observar la divergencia de resultados en dichos casos de patrones con un verbo y varios nexos. Por ejemplo, el patrón *ocupar* (*como*|*para*) obtiene una cobertura de 1 con el nexo *como*, pero una cobertura de 0.66 con el nexo *para*. En este mismo caso, la precisión es de apenas 0.20 con el nexo *como*, mientras que con el nexo *para* sube a 0.50. Lo mismo ocurre con el patrón *usar* (*como*|*en*|*para*), donde el nexo *como* recupera una cobertura de 1, mientras que en el caso del nexo *en* baja a 0.68. Otros patrones con resultados de cobertura altos fueron *ocupar como*, *servir en*, *emplear como*, *utilizar en* y *funcionar como*. En cuanto a la precisión, los valores más altos llegaron a 0.84, en el caso del patrón *usar como*, y arriba de 0.60 para *usar para*, *permitir*, *servir como* y *funcionar como*.

Finalmente, los resultados de los patrones sinonímicos los podemos observar en la siguiente tabla:

Patrones Sinonímicos		
Patrón verbal	**P**	**C**
conocer también	0.77	0.80
denominar también	0.72	0.94
llamar también	0.79	0.81

Tabla 6.12. Resultados de precisión y cobertura del ECODE para patrones verbales sinonímicos

Estos 3 casos recuperaron índices por arriba de 0.70 en el caso de la precisión y la cobertura. El mayor de ellos fue la cobertura del patrón *denominar también*, con 0.94, mientras que el más bajo fue para la precisión de este mismo patrón, con 0.72.

En resumen, con estos datos pretendemos dar cuenta de un panorama más específico a partir de los resultados de cada uno de los patrones verbales que incluimos en el sistema. Hemos visto que el sistema logra identificar la mayoría de los CDs clasificados manualmente por los evaluadores, tomando en cuenta que los resultados de la cobertura están en promedio por arriba de 0.75.

Con el fin de mejorar los resultados que hemos expuesto hasta ahora, realizamos una serie de modificaciones a la gramática de patrones verbales, los cuales detallaremos en el siguiente apartado.

6.5 Resultados de precisión y cobertura según tipo de restricciones

El primer tipo de restricción que aplicamos fue a las raíces de los verbos definitorios para buscar únicamente ocurrencias que cumplieran con las siguientes condiciones: 3a persona singular o plural, tiempo presente, infinitivo o participio. Por ejemplo, con esta restricción la expresión regular del verbo definir pasó de ser *defin* a *defin(e|en|ir|id.|id.s)*. En algunos casos, se excluyó de la raíz la forma en participio pensando en que no son formas recurrentes en CDs, por ejemplo la raíz del verbo *incluir* quedó de la siguiente manera: *inclu (?:ye|yen|ir)*.

De esta forma obtuvimos los siguientes resultados:

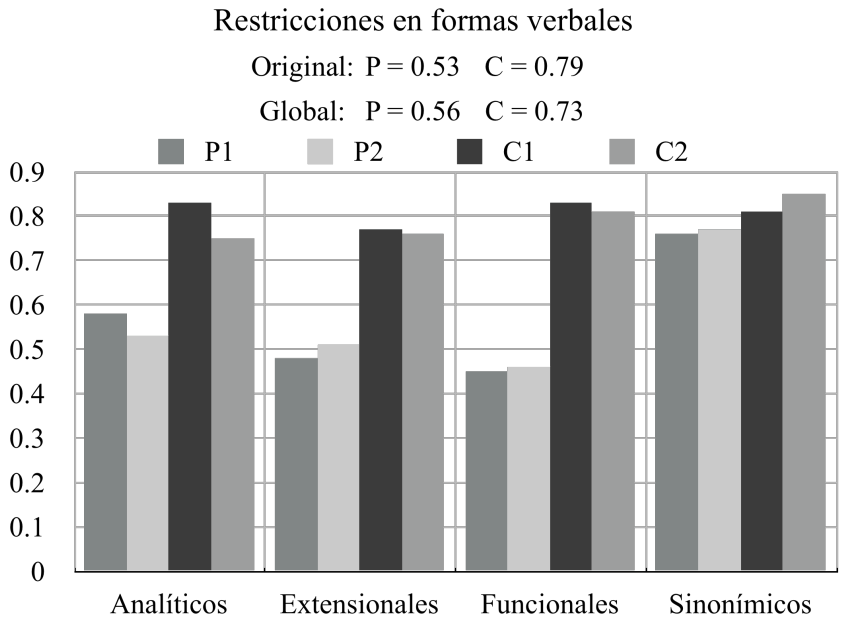

Figura 6.1. Resultados de precisión y cobertura del ECODE con restricciones de raíces verbales

En la figura anterior mostramos por cada tipo de patrón los datos originales de precisión y cobertura comparados contra los datos globales obtenidos en esta nueva evaluación. Se puede observar que los valores obtenidos a partir de aplicar estas restricciones verbales se modificaron mínimamente a favor de la precisión, excepto en el caso de los patrones verbales analíticos. La cobertura, por su parte, también se vio afectada pero en contra, es decir, que disminuyó excepto en el caso de los patrones sinonímicos. A nivel global también vemos un aumento de la precisión pero una disminución de la cobertura, ya que los primeros índices obtenidos eran de P = 0.53, C = 0.79, mientras que con estas restricciones se obtuvieron índices de P = 0.56 y C = 0.73.

Ahora bien, a esta misma gramática con restricciones en las raíces verbales le aplicamos, además, restricciones de distancia entre el verbo y el nexo a una ventana de 15 palabras. Los resultados en este caso fueron los siguientes:

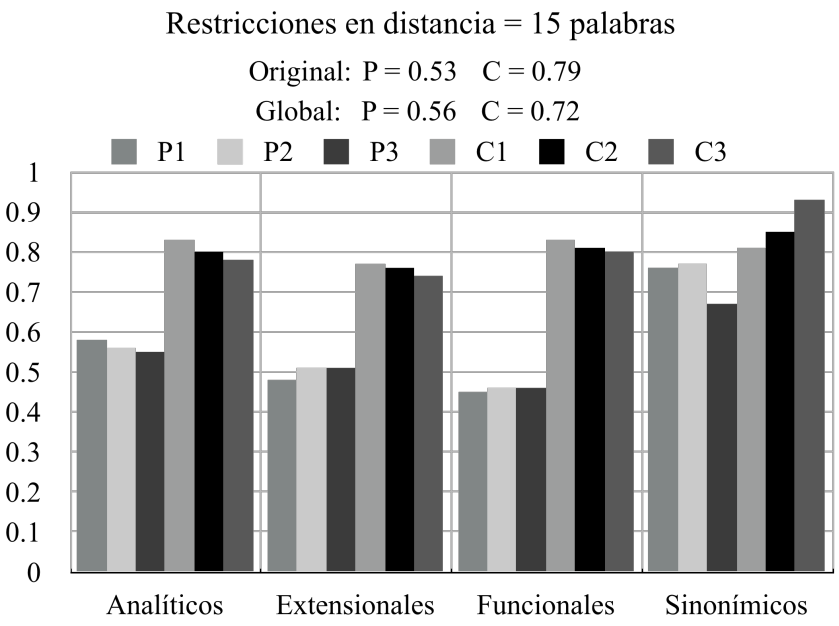

Figura 6.2. Resultados de precisión y cobertura del ECODE con restricciones en distancia

En esta figura añadimos los datos de precisión y cobertura en una tercera barra, e igualmente los datos originales los comparamos con los datos globales obtenidos en esta nueva evaluación. Vemos que la única mejora a favor fue en el caso de la cobertura de los patrones sinonímicos. En los patrones analíticos fue a la baja tanto en precisión como en cobertura; asimismo en los patrones extensionales y funcionales bajó la cobertura, aunque la precisión se mantuvo estable. Los índices globales también se modificaron a favor de la precisión y en contra de la cobertura. En este caso, la precisión aumentó de 0.53 a 0.56, mientras que la cobertura bajó de 0.79 a 0.72.

En cuanto a los patrones contextuales, es decir, la posibilidad que tienen los términos de aparecer a izquierda, derecha o en la posición de nexo respecto al verbo definitorio, eliminamos de la gramática la posibilidad de encontrar términos en las posiciones de nexo y derecha en los patrones que permiten al término en la posición izquierda, y respetamos la posición derecha para los casos en los que el término sólo puede aparecer en dicha posición. Los resultados los presentamos en la siguiente figura:

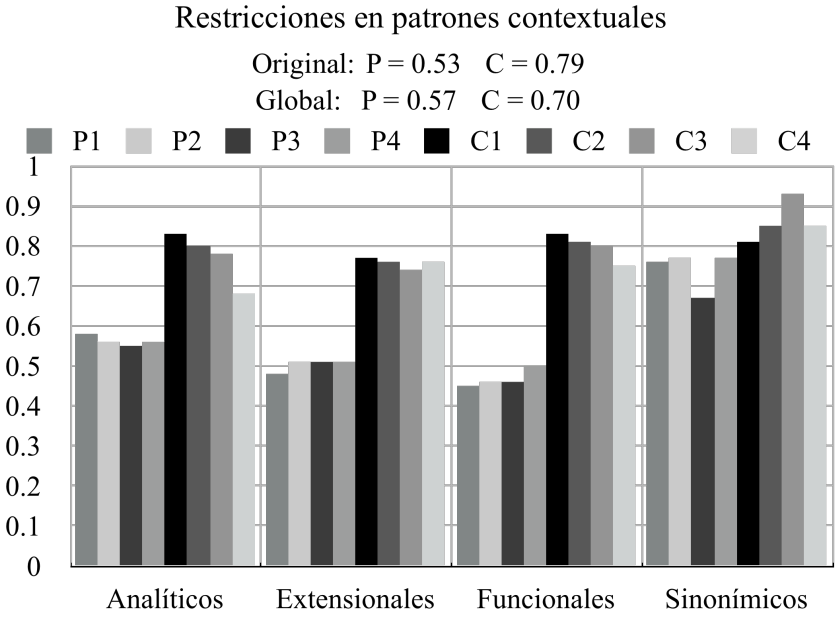

Figura 6.3. Resultados de precisión y cobertura del ECODE con restricciones en distancia

La figura anterior muestra una cuarta barra por cada índice de precisión y cobertura que corresponde a los nuevos datos. Podemos ver que la precisión global sube mínimamente de 0.53 a 0.57; sin embargo, en el caso de la cobertura, ésta baja de 0.79 a 0.70. De manera específica, vemos que en los patrones analíticos la precisión sube por primera vez mientras que en los patrones extensionales se mantiene estable y en los funcionales y sinonímicos incrementa. La cobertura sube en el caso de los patrones extensionales, sin embargo lo hace a un nivel más bajo que el inicial.

Como podemos observar, el beneficio de aplicar estas restricciones recae en una reducción del ruido, es decir, de los contextos no relevantes que el ECODE clasifica como CDs. No obstante, perdemos en cobertura ya que algunos CDs válidos que el sistema había clasificado correctamente, con las restricciones son clasificados como contextos no relevantes.

Ahora bien, pensando en obtener unos resultados más estables decidimos aplicar estas restricciones en los verbos que obtuvieron una precisión por debajo de 50, los cuales se encontraban en todos

los tipos de patrones excepto sinonímicos. De esta forma, obtuvimos los siguientes resultados para los patrones analíticos, extensionales y funcionales:

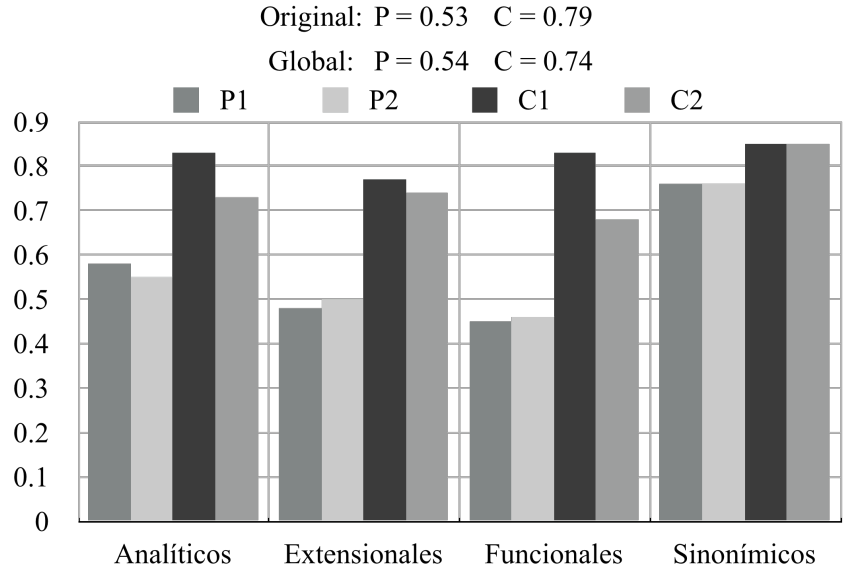

Figura 6.4. Resultados de precisión y cobertura del ECODE con combinación de restricciones 1

Podemos observar, en este caso, que la precisión global sube mínimamente de 0.53 a 0.54. Específicamente, para los patrones analíticos se sigue observando una reducción de los resultados de los dos índices, mientras que la precisión en los patrones extensionales y funcionales sube ligeramente.

Tomando en cuenta lo anterior, decidimos conservar la configuración original para los patrones analíticos y aplicar combinaciones de restricciones en las raíces de los verbos extensionales y funcionales, así como en las distancias que pueden ocurrir entre estos verbos y sus nexos. Los resultados los podemos ver en la siguiente figura:

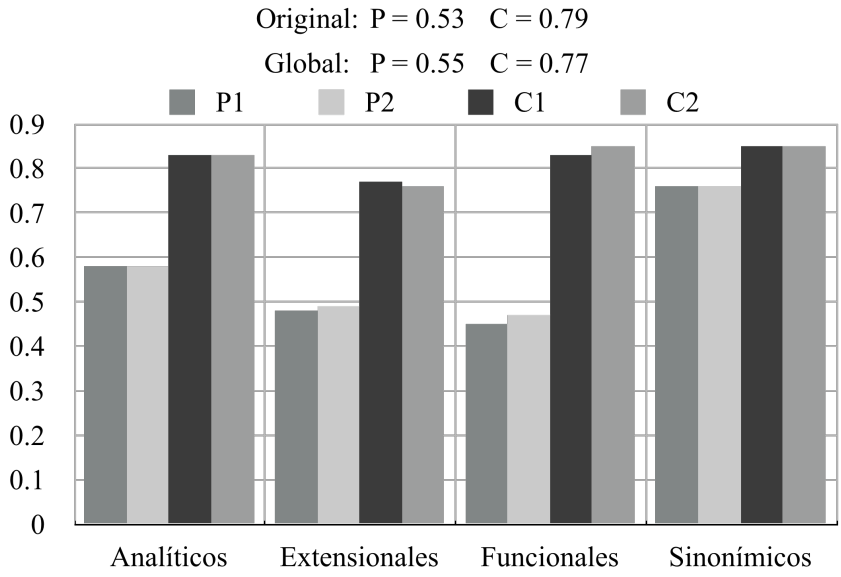

Figura 6.5. Resultados de precisión y cobertura del ECODE con combinación de restricciones 2

Finalmente, observamos cómo se estabilizan los resultados en la misma tendencia de la gramática original, pero aumentando de 0.53 a 0.55 la precisión, y disminuyendo de 0.79 a 0.77 la cobertura. En el caso de los patrones extensionales y funcionales podemos observar un aumento de la precisión. A su vez, la cobertura aumenta en los patrones funcionales pero disminuye en los extensionales. Los patrones analíticos y sinonímicos permanecen con los resultados originales, ya que se excluyeron de las restricciones.

En resumen, si bien el balance en los resultados de las restricciones es mínimo, con estos resultados pretendemos dar cuenta del comportamiento específico de cada grupo de CDs a partir de modificaciones hechas en la estructura de los patrones verbales. Pretendemos también dar un panorama de las posibilidades del sistema de adaptar la gramática de patrones verbales de acuerdo con diferentes restricciones, lo cual supone una posibilidad para el usuario final de decidir entre obtener mejores resultados de precisión o cobertura, según sus necesidades específicas.

6.6 Conclusiones

Para cerrar este apartado, en primer lugar retomaremos las siguientes consideraciones generales sobre los resultados obtenidos en la evaluación:

a) El ECODE está basado en una gramática de patrones verbales asociados a distintos tipos de definiciones. Esta gramática permite configurar diferentes opciones para crear patrones de búsqueda con diferentes niveles de restricciones, las cuales pueden darse en la raíz del verbo definitorio, en la distancia entre este verbo y sus nexos (si los tiene), o bien en el lugar donde puede aparecer el término con respecto al verbo definitorio.
b) El desarrollo de nuestro sistema lo llevamos a cabo a partir de una gramática sin restricciones, es decir, se consideraba la raíz mínima del verbo, cualquier distancia entre éste y los nexos, y el término podía aparecer en posiciones de izquierda, nexo y derecha en determinados verbos definitorios. En esta primera gramática obtuvimos un índice global de precisión de 0.53 y de cobertura de 0.79.
c) Una vez aplicadas diferentes combinaciones de restricciones, encontramos que los índices de precisión y cobertura quedaban de la siguiente forma: 0.55 y 0.77 respectivamente. Las restricciones que nos permitieron aumentar la precisión, sin bajar considerablemente la cobertura, fueron las restricciones de las raíces y de distancia en los patrones extensionales y funcionales, específicamente en aquellos que en la primera evaluación obtuvieron una precisión menor a 0.50.
d) Finalmente, con los resultados globales y particulares de precisión y cobertura nos podemos dar cuenta de que el sistema está identificando aproximadamente el 80% del total de CDs presentes en el corpus. Sin embargo, del total de lo recuperado de forma automática, el 50% aproximadamente, es ruido.

Por otro lado, tomando en cuenta el estado del arte que expusimos en el capítulo 3, en la siguiente tabla podemos ver un panorama general de los resultados del ECODE frente a otros obtenidos en

diferentes sistemas y/o metodologías para la extracción de conocimiento definitorio. Para llevar a cabo esta comparación tomamos en consideración aquellas referencias que hubieran empleado los índices de precisión y cobertura para llevar a cabo su propia evaluación.

Referencia	Idioma	P	C
DEFINDER	in	86.95%	75.47%
Malaisé	fr	55%	39.3%
Sánchez y Márquez	es	97.44%	100%
Storrer y Wellinghoff	al	34%	70%
MOP	in	0.97	0.79
		0.94	0.81
LT4eL	bu	22.5%	8.9%
	ch	22.3%	46%
	pol	23.3%	32%
	por	0.14	0.86
GlossExtractor	in	0.87	0.86
ECODE	es	0.53	0.79
		0.55	0.77

Tabla 6.13. Comparación de resultados de precisión y cobertura del ECODE con otros sistemas de extracción de información definitoria

De la tabla anterior podemos destacar los siguientes puntos:

a) Observamos, en primer lugar, los diferentes idiomas en los que se han empleado metodologías para la extracción de definiciones o CDs. Entre ellos encontramos el trabajo de Sánchez y Márquez en español, quienes reportan una precisión de 97.44% y una cobertura de 100%. Aquí debemos tomar en cuenta que estos resultados los reportan en un experimento donde se utilizaron patrones verbales generados a partir de un sólo verbo (*entender*), y que la evaluación se llevó a cabo sobre un corpus que contenía 38 CDs.
b) Por otro lado, en francés encontramos el trabajo de Malaisé con 55% de precisión y 39.3% de cobertura. En este caso, los índices del ECODE resultan mayores aunque debe también tomarse en cuenta el escenario de la evaluación y el tamaño del corpus empleado en ambos casos.
c) En portugués están los resultados del proyecto LT4eL, donde se reporta una cobertura de 0.86, pero una precisión de 0.14.

d) En inglés, DEFINDER y GlossExtractor son capaces de recuperar aproximadamente el 87% de precisión y arriba del 75% de cobertura, siendo el caso de GlossExtractor el que reporta los resultados más altos de cobertura para este idioma.
e) Por su parte, el sistema MOP, también para el inglés, reporta dos evaluaciones donde los mejores resultados se obtuvieron en la precisión, con índices por encima de 0.90.
f) Por último, en alemán se reporta una precisión de 30% y una cobertura de 70%, mientras que en lenguas eslavas se obtuvo una precisión por encima del 20% y una cobertura cuyo valor más alto fue de 46%.

Ahora bien, como ya señalamos en el capítulo 3, las distintas metodologías de evaluación así como el tamaño de los corpus utilizados para dicho fin hace que sea complicado establecer una comparación plena entre los resultados obtenidos.

Si nos guiamos únicamente por los índices presentados en la tabla anterior, encontramos que nuestro sistema podría compararse en precisión a los resultados obtenidos por Malaisé, y estaría por encima de aquellos presentados para el alemán, portugués y las lenguas eslavas.

De la misma forma, la cobertura superaría los resultados del francés, alemán, checo y polaco, y se situaría casi a la par de los resultados presentados por el MOP. No obstante, debemos tomar en cuenta que en nuestra evaluación participaron tres evaluadores a quienes se les asignó una parte del corpus de evaluación, pero los resultados no se compararon entre sí para tratar de determinar un conjunto de CDs seleccionado por más de una sola persona, como fue el caso de DEFINDER.

Por otro lado, nuestro corpus estuvo formado por textos en el área de medicina y no por diferentes áreas como en el caso del MOP. Lo anterior puede suponer que los resultados del ECODE varíen de acuerdo con el área de aplicación, tomando en cuenta que los diferentes verbos pueden tener diferentes comportamientos en dominios particulares.

Por último, es importante notar que los resultados de nuestra evaluación favorecen la cobertura, ya que el corpus que utilizamos

para esta etapa se conformó a partir de la búsqueda de ocurrencias que tuvieran algún lema definitorio y un nexo para los patrones que lo permitieran.

En el siguiente y último capítulo retomaremos las ideas generales de esta tesis, presentaremos las principales aportaciones y concluiremos con algunas consideraciones a tomar en cuenta respecto al trabajo futuro de nuestra investigación.

7. Conclusiones

En este capítulo presentamos las conclusiones de esta tesis. Para ello comenzaremos con una recapitulación general de lo expuesto donde retomaremos también algunas ideas expuestas en las hipótesis que presentamos en el primer capítulo (7.1). Posteriormente abordaremos lo que consideramos como las principales aportaciones de esta investigación (7.2). Por último, describiremos los puntos relacionados con el trabajo futuro (7.3).

7.1 Recapitulación

De manera sintetizada, en esta investigación hemos abordado el tema de la extracción automática de CDs en español, a partir de la descripción de la metodología seguida para el desarrollo de un sistema basado en patrones lingüísticos que trabaja sobre textos etiquetados morfosintácticamente. Asimismo, hemos presentado una evaluación de dicho sistema, con la cual hemos podido obtener un panorama de diferentes aspectos de su desempeño. Ahora bien, a continuación listamos una serie de consideraciones que creemos importante resaltar a modo de recapitulación de cada capítulo de esta tesis.

En el capítulo 2 presentamos algunos conceptos necesarios para comprender el planteamiento de la tesis, principalmente relacionados con el lenguaje especializado y la práctica terminográfica. Específicamente, abordamos el tema del lenguaje especializado junto con los conceptos de término y definición, y la relación de éstos en el ámbito de la terminología y la terminografía. También explicamos algunas consideraciones sobre la práctica actual de la terminografía, específicamente relacionadas con el uso de metodologías computacionales. Posteriormente, explicamos de manera general los conceptos de contexto definitorio y relación semántica, para finalmente detallar algunos datos relacionados con la aparición de definiciones en textos especializados.

Por su parte, en el capítulo 3 revisamos algunos estudios relacionados con el tema de la extracción de definiciones y de CDs. Tratamos ciertas consideraciones sobre el concepto de contexto definitorio y la extracción de información, así como algunas

diferencias que creemos importante resaltar entre la extracción de contextos definitorios frente a la extracción de relaciones semánticas, con el fin de delimitar los estudios relacionados con nuestra investigación. Asimismo, explicamos los trabajos relacionados con el estado del arte y presentamos un análisis contrastivo de las metodologías seguidas en dichos trabajos. Observamos que los estudios sobre la extracción de CDs parten de la búsqueda de patrones definitorios, los cuales son considerados como puntos clave para reconocer fragmentos textuales con información definitoria. También observamos que uno de los tipos de patrones definitorios más recurrentes en la extracción de definiciones son los patrones verbales, y que éstos suelen asociarse a distintos tipos de información semántica. Otro punto a resaltar de este apartado es la síntesis de las metodologías de los estudios tratados, las cuales suelen recurrir no sólo a la búsqueda de patrones definitorios, sino también a la exclusión o filtro de contextos no relevantes, así como a la detección de los elementos constitutivos de los candidatos a CDs, es decir, los términos y las definiciones. Finalmente, observamos los tipos de metodologías empleadas para la evaluación de los sistemas, las cuales suelen recurrir a los índices de precisión y cobertura para llevar a cabo dicha evaluación.

En el capítulo 4 profundizamos en los conceptos de contexto y contexto definitorio, desde el punto de vista del trabajo terminográfico, como entrada para describir un análisis lingüístico de CDs en textos especializados en español. En este análisis describimos las principales características de dichos contextos, específicamente los elementos constitutivos que suelen incluir: términos, definiciones, patrones definitorios y patrones pragmáticos. Detallamos asimismo los tipos de patrones que suelen emplearse para conectar al término con su definición, así como para resaltar visualmente la presencia de estos dos elementos. En el caso de los patrones pragmáticos, explicamos que suelen utilizarse para especificar detalles sobre el significado de los términos, su alcance o sus contextos de uso. Vimos además que, en español, los CDs suelen seguir una serie de patrones contextuales que están relacionados con el patrón definitorio que se emplea, y que tienen relación con la posición que ocupa el término y la definición respecto al patrón que los conecta, lo cual constituye información que debe tomarse en cuenta para su extracción automática. Adicionalmente, propusimos una tipología semántica de CDs

basada en los tipos de patrones verbales definitorios que utilizamos en el desarrollo de nuestra investigación.

El desarrollo del sistema de extracción automática de CDs, denominado ECODE, lo presentamos en el capítulo 5. Expusimos que este sistema está basado en reglas lingüísticas e incorpora diferentes procesos para extraer contextos donde se aporta información definitoria sobre términos. Este sistema está compuesto por diferentes módulos, de los cuales, el primero consiste en buscar ocurrencias de patrones verbales definitorios, el segundo aplica un filtro de contextos no relevantes, el tercero se encarga de identificar los términos y definiciones en los candidatos a CDs, y el último de ellos realiza un análisis de los resultados obtenidos para tratar de encontrar los mejores resultados. Observamos que el núcleo del sistema son los patrones verbales, los cuales se representan mediante una gramática que incluye varios parámetros indispensables para el funcionamiento del sistema. Estos parámetros son el lema y la raíz del verbo definitorio, los nexos gramaticales que pueden acompañar al verbo, la distancia de palabras posibles entre el verbo y sus nexos, las posiciones del término que permite cada patrón verbal, así como el tipo de definición al que se asocia dichos patrones.

Por último, en el capítulo 6 presentamos la evaluación del sistema. Para ello explicamos la metodología que seguimos en la conformación del corpus de evaluación, las métricas utilizadas y los parámetros generales de la metodología. Señalamos los índices de precisión y cobertura como las medidas utilizadas en la evaluación, y explicamos los resultados de cada proceso de manera general y específica. En el primer caso, presentamos los resultados globales de precisión y cobertura del ECODE, mientras que en el segundo caso presentamos los resultados por cada tipo de patrón verbal, es decir, los patrones analíticos, extensionales, funcionales y sinonímicos. Adicionalmente, presentamos los resultados de cada uno de los patrones verbales de cada grupo, con lo cual pudimos tener un panorama más específico del desempeño del sistema.

Ahora bien, en cuanto a las hipótesis planteadas en esta investigación, en las siguientes líneas abordamos algunas consideraciones al respecto.

1. Los candidatos a CDs pueden ser extraídos automáticamente a partir de la búsqueda de las ocurrencias de patrones definitorios.

- Hemos observado a lo largo de nuestra investigación que, efectivamente, la extracción de CDs es posible a partir de la búsqueda de patrones definitorios.
- En esta búsqueda hemos observado también que existen algunas consideraciones importantes, como las restricciones que deben imponerse a cada patrón con el fin de obtener resultados más precisos.
- También hemos observado que los patrones verbales definitorios pueden asociarse a distintos tipos de información semántica.

2. Es posible filtrar automáticamente excepciones en las ocurrencias encontradas.

- Hemos comprobado igualmente que el filtro automático de excepciones es posible mediante una gramática donde se establecen las reglas léxico-sintácticas que el sistema identifica en candidatos para excluirlos como CDs.

3. Se pueden identificar automáticamente los elementos constitutivos de los candidatos a CDs.

- Observamos también que los términos y las definiciones en los candidatos pueden identificarse automáticamente a partir de expresiones regulares y reglas contextuales que toman como base una gramática de patrones verbales definitorios para realizar dicho proceso.

4. Se pueden clasificar automáticamente los resultados obtenidos a partir de una serie de reglas lingüísticas para determinar los mejores candidatos a CDs.

- Hemos visto que esta clasificación es viable y que pueden identificarse mejores candidatos a partir de reglas lingüísticas.

7.2 Aportaciones

El sistema que hemos presentado en esta tesis constituye un esfuerzo por incorporar una metodología para la extracción automática de información definitoria sobre términos especializados en español.

Específicamente, el ECODE es un sistema basado en reglas lingüísticas, cuyo núcleo es la búsqueda de patrones verbales definitorios especificados en una gramática y asociados a distintos tipos de definiciones. En este sentido, cabe resaltar las siguientes ventajas de nuestra metodología:

- La gramática de patrones verbales definitorios puede ser modificada según los intereses del usuario.
- Es posible añadir nuevos patrones verbales no considerados previamente, así como asociarlos a nuevos tipos de definiciones.
- Como vimos a lo largo de la evaluación, la gramática permite además generar diferentes combinaciones para tratar de mejorar los resultados tanto de precisión como de cobertura, lo cual estaría igualmente relacionado con los intereses del usuario.

Asimismo, el ECODE incorpora una metodología para excluir contextos no relevantes que puedan ser recuperados por los patrones verbales de la gramática que señalamos anteriormente. Algunas de las ventajas de esta metodología son las siguientes:

- Al igual que los patrones verbales, las reglas de filtro se especifican mediante una gramática que puede ser modificada fácilmente para que el usuario agregue o altere las reglas ya especificadas.
- Las restricciones pueden ser en mayor o menor medida más complejas, ya que la metodología permite la búsqueda no sólo de palabras claves relacionadas con los filtros, sino combinaciones sintácticas que se pueden especificar mediante las etiquetas de parte de la oración del corpus de entrada.
- En este último sentido, tanto la gramática de patrones verbales como la gramática de filtro recurren a expresiones regulares

de ciertos tipos de palabras que se especifican independientemente, con lo cual es posible la aplicación del sistema en otros corpus con diferentes anotaciones POS a partir de modificaciones mínimas.

Por otro lado, el ECODE incorpora una metodología para la detección de los elementos constitutivos de los CDs. Esta identificación contiene principalmente las siguientes ventajas:

- Las reglas para identificar términos y definiciones están basadas en patrones contextuales y especificadas en la gramática de patrones verbales. En este caso, los patrones contextuales se combinan a su vez con expresiones regulares para detectar los posibles términos y definiciones de cada candidato. Con ello, es posible entonces encontrar términos simples y términos compuestos, en el sentido de su estructura sintáctica, sin la necesidad de recurrir a un extractor terminológico. No obstante, la implementación de un sistema para extraer términos, o bien la conjunción con uno de ellos, sería una mejora sustancial para el ECODE.

Finalmente, en el desarrollo de la metodología del ECODE incorporamos un proceso para reorganizar los resultados y tratar de identificar aquellos que probablemente sean mejores candidatos a CDs. Este proceso se realizó mediante la implementación de reglas lingüísticas. Creemos que la mayor ventaja de este proceso es la siguiente:

- El hecho de incorporar un análisis automático de los resultados permite tener un conjunto estructurado de acuerdo con la probabilidad de que cada candidato sea o no un CD. Esto da al usuario una mayor facilidad a la hora de discernir manualmente si lo extraído automáticamente sirve en el contexto de su trabajo o es útil para sus intereses.

7.3 Trabajo futuro

A partir de la investigación que hemos presentado en esta tesis se apuntan diferentes líneas de trabajo futuro que podrían tomarse para mejorar el sistema.

Desde el punto de vista de un análisis lingüístico, resultaría de utilidad ahondar en el estudio de CDs en español, con la finalidad de aportar bases teóricas sólidas sobre su realización en textos especializados. Estas bases podrían arrojar información útil para mejorar los algoritmos de extracción automática de información definitoria.

Por su parte, los resultados del ECODE podrían mejorarse mediante la revisión de los procesos de filtro de contextos no relevantes y el árbol de decisión para identificar los elementos constitutivos. Creemos que en estos dos procesos sería de gran utilidad implementar técnicas de aprendizaje automático con la finalidad de incorporar nuevas reglas que no hayamos considerado previamente, lo cual podría mejorar los resultados globales del algoritmo.

Asimismo, es necesario profundizar en la metodología del ranking de CDs. Creemos que este proceso resulta útil para identificar resultados de mayor calidad y como un filtro extra de contextos no relevantes. En este caso, consideramos importante tomar en cuenta que la metodología implementada en este proceso permita al usuario modificar los parámetros del ranking, esto es, que pueda decidir en qué punto establecer un umbral para obtener datos más o menos específicos de acuerdo con sus intereses.

Finalmente, sería de gran utilidad incorporar en la metodología del ECODE nuevos tipos de patrones definitorios, como patrones tipográficos y marcadores reformulativos, los cuales podrían servir como puntos de inicio para la recuperación de candidatos a CDs.

Por otro lado, en el estudio de la extracción automática de CDs nos hemos encontrado con ciertos problemas que por sí mismos han conformado líneas de investigación paralelas, las cuales han sido abordadas a diferentes niveles, principalmente como investigaciones a nivel de tesis por otros miembros del Grupo de Ingeniería Lingüística. Consideramos importante hacer mención de estas líneas y tomarlas en cuenta como parte del trabajo futuro.

Uno de los problemas al que nos hemos enfrentado es la identificación de referencias anafóricas cuando el término es presentado en párrafos anteriores a donde se define y se sustituye mediante una anáfora. En este sentido, se ha desarrollado una

investigación relacionada con el estudio de la realización de anáforas en CDs (Benítez 2008).

Otra línea de investigación que se ha trabajado es el análisis de la extensión de las definiciones, ya que en algunos casos el sistema identifica como tal a ciertas partes del contexto que ya no aportan información definitoria. En este caso, se desarrolló un estudio lingüístico de las definiciones analíticas con la finalidad de desarrollar reglas para delimitar automáticamente su extensión (Hernández 2009).

Asimismo, las definiciones analíticas se estudiaron con el fin de considerar patrones que introduzcan otro tipo de información definitoria, ya que, en ocasiones, en este tipo de definiciones se incluye además otra relación semántica específica, por ejemplo funcional o extensional (Sánchez 2009).

Por último, en el marco del desarrollo de un sistema para la extracción de definiciones en la Web, se desarrolló un estudio para el agrupamiento semántico de CDs, lo cual tiene como finalidad clasificar automáticamente grupos de definiciones parecidas para un mismo término (Molina 2009).

Creemos importante que el trabajo futuro de nuestra investigación tome en cuenta el desarrollo y la continuidad de estos trabajos, los cuales en conjunto pueden aportar mejoras relevantes en el tema de la extracción automática de CDs.

Bibliografía

Aguilar, C.; Alarcón, R.; Rodríguez, C. y Sierra, G. (2006). "Reconocimiento y clasificación de patrones verbales definitorios en corpus especializados". En Cabré, M. T.; Estopà, R. y Tebé, C. (eds.) *La Terminología en el Siglo XXI: contribución a la cultura de la paz, la diversidad y la sostenibilidad*. Barcelona, Institut Universitari de Lingüística Aplicada, Documenta Universitaria. 259-268.

Aguilar, C. (2009). *Análisis lingüístico de definiciones en contextos definitorios*. México, Universidad Nacional Autónoma de México. [Tesis de doctorado]

Alarcón, R. (2003). *Análisis lingüístico de contextos definitorios en textos de especialidad.* México, Universidad Nacional Autónoma de México. [Tesis de licenciatura]

Alarcón, R. (2006). *Extracción automática de contextos definitorios en textos especializados. Propuesta para la elaboración de un ECCODE (Extractor de Candidatos a Contextos Definitorios)*. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Proyecto de tesis]

Alarcón, R. y Sierra, G. (2003). "El rol de las predicaciones verbales en la extracción automática de conceptos". *Estudios de Lingüística Aplicada* 38. 129-144.

Alshawi, H. (1987). "Processing Dictionary Definitions with Phrasal Pattern Hierarchies". *Computational Linguistics* 13 (3-4). 195-202.

Amsler, R. (1981). "A Taxonomy for English Nouns and Verbs". En *Proceedings of the 19th Annual Meeting of the Association for Computational Linguistics*. California, 29 de junio al 1 de julio. 133-138.

Arntz, R. y Picht, H. (1995). *Introducción a la terminología*. Madrid, Pirámide. [Tit. Orig.: *Einführung in die*

Terminologiearbeit, 1989; traducción de Amelia de Irrazazábal, *et al*.]

Auger, A. (1997). *Regérage des énoncé d'intér*êt *définitoire dans les bases de données textuelles*. Neuchâtel, Universidad de Neuchâtel. [Tesis de doctorado]

Auger, A. y Barrière, C. (2008). "Pattern-based Approaches to Semantic Relation Extraction. A state of the art". *Terminology* 14 (1). 1-19.

Austin, J. L. (1962). *How to do things with words*. Oxford, Oxford University Press.

Bach, C. (2005). "Los marcadores de reformulación como localizadores de zonas discursivas relevantes en el discurso especializado". En *Debate Terminológico* 1. [En línea]. RITerm. http://www.riterm.net/revista/n_1/bach.pdf

Bach, C.; Saurí, R.; Vivaldi, J. y Cabré, M. T. (1997) *El Corpus de l'IULA: descripció*. Informe 17. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Benítez, V. (2008). *Anáforas en la expansión de contextos definitorios: una propuesta de etiquetado*. México, Universidad Nacional Autónoma de México. [Tesis de licenciatura]

Berland, M. y Charniak, E. (1999). "Finding Parts in Very Large Corpora". En actas del *37th Annual Meeting of the Association for Computational Linguistics, ACL'99*. Maryland, 20-26 de junio. 54-64.

Boguraev, B. y Pustejovsky, J. (1996). "Issues in Text-Based Lexicon Acquisition". En Boguraev, B. y Pustejovsky, J. (eds.) *Corpus Processing for Lexical Acquisition*. Cambridge, MIT Press. 3-17.

Borg, C. (2007). "Discovering Grammar Rules for Automatic Extraction of Definitions". En *Doctoral Consortium at the 8th EUROLAN Summer School*. Iasi, 30 de julio a 2 de agosto 2.

Borg, C.; Rosner, M. y Pace, G. (2007). "Towards Automatic Extraction of Definitions". En *Proceedings of the Computer Science Annual Workshop (CSAW'2007). Malta, 5 a 6 de noviembre.* 52-65.

Bourigault, D.; Jacquemin, C. y L'Homme, M. C. (2001). (eds.) *Recent Advances in Computational Terminology*. Ámsterdam, John Benjamins.

Bowker, L. y Pearson, J. (2002). *Working with Specialized Language. A Practical Guide to Using Corpora*. Londres, Routledge.

Cabré, M. T. (1993). *La terminología. Teoría, metodología, aplicaciones*. Barcelona, Editorial Antártica / Empúries.

Cabré, M. T. (1999). *La terminología: representación y comunicación. Elementos para una teoría de base comunicativa y otros artículos*. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Cabré, M. T.; Domènech, M.; Morel, J. y Rodríguez, C. (2001). "Las características del conocimiento especializado y la relación con el conocimiento general". En Cabré, M. T. y Feliu, J. (eds.) *La Terminología Científico – Técnica: Reconocimiento, análisis y extracción de información formal y semántica*. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. 173-186.

Calzolari, N. y Picchi, E. (1988). "Acquisition of Semantic Information from an On-Line Dictionary". En *12th International Conference on Computational Linguistics*, *Coling'88*. Budapest, 22-27 de agosto. 87-92.

Condamines, A. (2002). "Corpus Analysis and Conceptual Relation Patterns". *Terminology* 8 (1). 144-162.

Condamines, A. y Rebeyrolle, J. (2001). "Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CTKB)". En Bourigault, D.; Jacquemin, C. y L'Homme, M. C. (eds.) *Recent Advances in Computational Terminology*. Ámsterdam, John Benjamins. 127-148.

Cucchiarelli, A.; Navigli, R.; Neri, F. y Velardi, P. (2004) "Automatic generation of glosses in the OntoLearn System". En *Proceedings of the 4th International Conference on Language Resources and Evaluation LREC2004*.. Lisboa, 26 a 30 de mayo. 1293-1296.

Davidson, L. (1997). Knowledge Extraction Technology for Terminology [en línea]. Ontario, Universidad de Ottawa. [Tesis de Maestría] http://aix1.uottawa.ca/~etithese/ldav/ldav_index.htm

De Bessé, B. (1991). "Le Contexte Terminographique". *Meta* 26 (1). 111-120.

Degórski, Ł.; Marcińczuk, M. y Przepiórkowski, A.; (2008). "Definition Extraction Using a Sequential Combination of Baseline Grammars and Machine Learning Classifiers". En *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*. Marruecos, 28 a 30 de mayo.

Del Gaudio, R. y Branco, A. (2007). "Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach". En *Proceedings of the Workshop Text Mining and Applications (TeMA). 13th Portuguese Conference on Artificial Intelligence (EPIA 2007)*. 3 a 7 de diciembre. 659-670.

Dolan, W.; Vanderwende, L. y Richardson, S. (1993). "Automatically Deriving Structured Knowledge Bases From on-Line Dictionaries". En *Proceedings of the 1st Conference of the Pacific Association for Computational Linguistics* (*Pacling'93*). [En línea]. Vancouver, 21-24 de abril. 5-14. http://research.microsoft.com/research/pubs/view.aspx?tr_id=102

Feliu, J. (2004). *Relaciones conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Tesis de doctorado]

Fellbaum, C. (ed.) (1998). *WordNet: An electronic lexical database*. MIT Press.

García de Quesada, M. (2001). "Estructura definicional terminográfica en el subdominio de la oncología clínica". *Estudios de Lingüística Española (ELiES)* 14. [En línea]. RedIRIS. http://elies.rediris.es/elies14/index.html#indice

Hearst, M. (1992). "Automatic Acquisition of Hyponyms from Large Text Corpora". En *Proceedings of the 14th International Conference on Computational Linguistics, Coling'92*. Nantes, 23-28 de agosto. 539-545.

Hernández, C. (2009). *Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática*. México, Universidad Nacional Autónoma de México. [Tesis de licenciatura]

Iftene, A.; Trandabăț, D. y Pistol, I. (2007). "Grammar-Based Automatic Extraction of Definitions. Applications for Romanian". En *Proceedings of the Workshop Natural Language Processing and Knowledge Representation for eLearning Environments. International Conference Recent Advances in Natural Language Processing (RANLP'2007)*. Borovets, 27 a 29 de septiembre. 19-25.

ISO 704. (2000)(E). *Terminology work – Principles and methods*.

ISO 10241. (1992). *International terminology standards – Preparation and layout*.

ISO 12620. (1999). *Computer applications in terminology. Data categories*.

Jurafsky, D. y Martin, J. (2000). *Speech and Language Processing*. *An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Nueva Jersey, Upper Saddle River Prentice.

Klavans, J. y Muresan, S. (2001). "Evaluation of DEFINDER: A System to Mine Definitions from Consumer-Oriented Medical Texts". En *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01)*. Roanoke, 24 a 28 de junio. 201-202.

Kobyliński, Ł. y Przepiórkowski, A. (2008). "Definition Extraction with Balanced Random Forests". En *6th International Conference on Natural Language Processing (GoTAL'2008)*. Gotemburgo, 25 a 27 de agosto.

Lorente, M. (2001). "Teoría e innovación en terminografía: la definición terminográfica". En Cabré, M. T. y Feliu, J. (eds.) *La Terminología Científico – Técnica: Reconocimiento, Análisis y Extracción de Información Formal y Semántica*. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. 81 – 112.

Malaisé, V. (2005). *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles á partir de corpus textuels*. Paris, Université Paris 7–Denis Diderot. [Tesis de doctorado]

Malaisé, V.; Zweigenbaum, P. y Bachimont, B. (2004). "Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie". En *Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2004)*. Fès, 19 a 21 de abril.

Malaisé. V.; Zweigenbaum, P. y Bachimont, B. (2005). "Mining Defining Contexts to Help Structuring Differential Ontologies". *Terminology* 11 (1). 21-53.

Medina, A.; Sierra, G.; Garduño, G.; Méndez, C. y Saldaña, R. (2004). "CLI: An Open Linguistic Corpus for Engineering". En *Taller de Herramientas y Recursos Lingüísticos para el*

Español y el Portugués, IX Congreso Iberoamericano de Inteligencia Artificial (IBERAMIA). Puebla, 23 de noviembre.

Meyer, I. (2001). "Extracting Knowledge-rich contexts for Terminography". En Bourigault, D.; Jacquemin, C. y L'Homme, M. C. (eds.) *Recent Advances in Computational Terminology*. Ámsterdam, John Benjamins. 278-302.

Molina, A. (2009). *Agrupamiento semántico de contextos definitorios*. México, Universidad Nacional Autónoma de México. [Tesis de Maestría]

Monachesi, P. (2007). "The LT4eL Project: Overview". [En línea]. Utrecht, Universidad de Utrecht. www.lt4el.eu/content/files/ws_prague/lt4el-prague.pdf

Montero, S. (2002). "Estructuración conceptual y formalización terminográfica de frasemas en el subdominio de la oncología". *Estudios de Lingüística Española (ELiEs)* 19. [En línea]. RedIRIS. http://elies.rediris.es/elies19/index.html

Mooney, R. (2003). "Machine Learning". En Mitkov, R. (ed.) *Oxford Handbook of Computational Linguistics*. Oxford University Press. 376-394.

Moreno, Ribas.; Armengol, V.; Béjar, A.; Belanche, M.; Cortés, U.; Gavaldá, R.; Gimeno, J.; López, I.; Martín, M. y Sánchez, M. (1994). *Aprendizaje automático*. Barcelona, Universidad Politécnica de Catalunya.

Morin, E. (1998). "PROMÉTHÉE. Un outil d'aide à l'acquisition de relations sémantiques entre termes". En *5ème Conférence annuelle sur le Traitement Automatique des Langues Naturelles* (*TALN*). París, 10-12 de junio. 172-181.

Muresan, S. y Klavans, J. (2002). "A Method for Automatically Building and Evaluating Dictionary Resources". En *Proceedings of the 3th International Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas, 29 a 31 de mayo. 231-234.

Navigli, R. y Velardi, P. (2007). "GlossExtractor: A Web Application to Automatically Create a Domain Glossary". *Lecture Notes in Computer Science* 4733. 339-349.

Pearson, J. (1998). *Terms in contexts*. Ámsterdam, John Benjamins.

Pérez, C. (2002). "Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento". *Estudios de Lingüística Española (ELiEs)* 18. [En línea]. RedIRIS. http://elies.rediris.es/elies18/index.html

Pinto, A. y Oliveira, D. (2004). Extracção de Definições no Corpógrafo [en línea]. Facultad de Letras, Universidad de Porto. http://www.linguateca.pt/documentos/OliveiraPintoOut2004.pdf

Przepiórkowski, A.; Degórski, Ł.; Spousta, M.; Simov, K.; Osenova, P.; Lemnitzer, L.; Kuboň, V. y Wójtowicz, B. (2007). "Towards the Automatic Extraction of Definitions in Slavic". En *Proceedings of the Workshop Balto-Slavonic Natural Language Processing (BSNLP'2007). 45th Annual Meeting of the Association for Computational Linguistics (ACL'2007)*. Praga, 23 a 30 de junio. 43-50.

Pustejovsky, J.; Bergler, S. y Anick, P. (1993). "Lexical Semantic Techniques for Corpus Analysis". *Computational Linguistics* 19 (2). 331-58.

Rebeyrolle, J. (2000). "Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes". En *Journées Francophones d'Ingénierie des Connaissances. Toulouse*, 10 a 12 de mayo. 105-114.

Rebeyrolle, J. y Tanguy, L. (2000). "Repérage automatique de structures linguistiques en corpus: le cas des énoncés définitoires". *Cahiers de Grammaire* 25. 153-174.

Rey, A. (1995). *Essays on Terminology*. Sager, J. C. (editor y traductor). Ámsterdam, John Benjamins.

Rodríguez, C. (1999). *Operaciones Metalingüísticas Explícitas en Textos de especialidad*. Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. [Treball de Recerca]

Rodríguez, C. (2004). "Metalinguistic Information Extraction for Terminology". En *Proceedings of the 3rd International Workshop on Computational Terminology (CompuTerm2004)*. Génova, 29 de agosto. 15-22.

Rodríguez, C. (2005). *Metalinguistic Information Extraction from Specialised Texts to Enrich Computational Lexicons*. Barcelona, Departament de Traducció i Filologia, Universitat Pompeu Fabra. [Tesis de doctorado]

Sager, J. C. (1993). *Curso práctico sobre el procesamiento de la terminología*. Madrid, Pirámide. [Tit. Orig.: *A Practical Course on Terminology Processing*, 1990; traducción de Laura Chumillas Moya]

Sager, J. C. y Ndi-Kimbi, A. (1995). "The conceptual structure of terminological definitions and their linguistic realisations". *Terminology* 2 (1). 61-85.

Saggion, H. (2004). "Identifying Definitions in Text Collections for Question Answering". En *Proceedings 4th International Conference on Language Resources and Evaluation LREC2004*.. Lisboa, 26 a 30 de mayo. 1927-1930.

Sánchez, A. y Márquez, M. (2005). "Hacia un sistema de extracción de definiciones en textos jurídicos". En *Actas de la 1er Jornada Venezolana de Investigación en Lingüística e Informática*. Venezuela, 14 de Octubre. 1-10.

Sánchez, O. (2009). *La funcionalidad al interior de contextos definitorios con definiciones analíticas: el patrón sintáctico para + infinitivo*. México, Universidad Nacional Autónoma de México. [Tesis de licenciatura]

Sarmento, L.; Maia, B.; Santos, D.; Pinto, A. y Cabral, L. (2006). "Corpógrafo V3. From Terminological Aid to Semi-automatic

Knowledge Engineering". En *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*. Génova, 22 a 28 de mayo. 1502-1505.

Sierra, G.; Medina, A.; Alarcón, R. y Aguilar, C. (2003). "Towards the Extraction of Conceptual Information from Corpora". En Archer, D.; Rayson, P.; Wilson, A. y McEnery, T. (eds.) *Corpus Linguistics*. Lancaster, UCREL.

Sierra, G.; Alarcón, R.; Aguilar, C. y Bach, C. (2008). "Definitional Verbal Patterns for Semantic Relation Extraction". *Terminology* 14 (1). 74-98.

Storrer, A. y Wellinghoff, S. (2006). "Automated Detection and Annotation of Term Definitions in German Text Corpora". En *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*. Génova, 22 a 28 de mayo. 2373-2376.

Trimble, L. (1985). *English for Science and Technology. A Discourse Approach*. Cambridge, Cambridge University Press.

Velardi, P.; Navigli, R. y D'Amadio, P. (2008). "Mining the Web to Create Specialized Glossaries". *IEEE Intelligent Systems* 23 (5). 17-25.

Vivaldi, J. (2004). *Extracción de candidatos a términos mediante la combinación de estrategias heterogéneas*. Barcelona, Institut Universitari de Lingüística Aplicada (Serie Tesis; 9). [Tesis de doctorado]

Vivaldi, J. y Rodríguez, H. (2007). "Evaluation of Terms and Term Extraction Systems: A Practical Approach". *Terminology* 13 (2). 225-248.

Vossen, P. y Copestake, A. (1993). "Untangling Definition Structure into Knowledge Representation". En Briscoe, T.; Pavia, V. y Copestake, A. (eds.) *Inheritance, Defaults and the Lexicon*. Cambridge, Cambridge University Press. 246-274.

Walter, S. y Pinkal, M. (2006). "Automatic Extraction of Definitions from German Court Decisions". En *Proceedings of the Workshop on Information Extraction Beyond the Document*. *21st International Conference on Computational Linguistics (COLING'2006)*. Sydney, 22 a 23 de julio. 20–28.

Westerhout, E. y Monachesi, P. (2007). "Extraction of Dutch Definitory Contexts for eLearning Purposes". En *Computational Linguistics in the Netherlands (CLIN'2007)*. Nijmegen, 7 de diciembre.

Westerhout, E. y Monachesi, P. (2008). "Combining Pattern-Based and Machine Learning Methods to Detect Definitions for eLearning Purposes". En *Proceedings of the Workshop Natural Language Processing and Knowledge Representation for eLearning Environments. International Conference Recent Advances in Natural Language Processing (RANLP'2007)*. Borovets, 27 a 29 de septiembre.

Wright, S. y Budin, G. (eds.) (1997). *Handbook of Terminology Management: Basic Aspects of Terminology Management.* Ámsterdam, John Benjamins.

Xu, J.; Cao, Y.; Li, H. y Zhao, M. (2005). "Ranking Definitions with Supervised Learning Methods". En *Special interest tracks and posters of the 14th international conference on World Wide Web*. Nueva York, Association for Computing Machinery. 811-819.