Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos

Gerardo Sierra
Universidad Nacional Autónoma de México
gsierram@ii.unam.mx

Resumen

La extracción automática de definiciones a partir de textos de especialidad es una tarea cada vez más demandante para diferentes aplicaciones del Procesamiento de Lenguaje Natural, tales como lexicografía computacional, extracción de información, semántica computacional, sistemas pregunta-respuesta, minería de textos, Web semántica y aprendizaje automático. Este artículo presenta un panorama de los trabajos realizados en el Grupo de Ingeniería Lingüística en el tema, desde los aspectos teóricos, la revisión del estado del arte, los estudios lingüísticos sobre definiciones y contextos definitorios, la metodología para la extracción automática y hasta diversas aplicaciones.

1. Introducción

Este artículo constituye una síntesis de la investigación realizada en el Grupo de Ingeniería Lingüística del Instituto de Ingeniería, UNAM, referente a la extracción automática de contextos definitorios en textos de especialidad en español, mediante el reconocimiento y análisis de patrones lingüísticos.

Esta investigación surge como parte de un proyecto central, que constituye la metodología para la creación de diccionarios onomasiológicos [23]. Para construir un diccionario de esta naturaleza, se debe contar con una base de conocimientos léxica lo suficientemente rica que contenga una diversidad de definiciones para cada uno de los términos que se están buscando. Para la obtención de dichas definiciones, además de las disponibles en los diccionarios, se puede acudir a los textos de especialidad, tales como artículos, reportes, tesis, etc., en donde los autores introducen los términos y, por tanto, proporcionan su definición. En este sentido, la investigación está orientada a la extracción automática de estas unidades del discurso utilizadas en los textos de especialidad donde se introduce un término y su definición, lo que aquí denominamos *contextos definitorios* (CDs).

Como parte del estado del arte, en la sección 2 tendremos una introducción a la extracción de información, en particular de la información terminológica y conceptual, campos donde se contextualiza esta investigación. Veremos las bases para el desarrollo de los sistemas de extracción de esta información y los trabajos realizados en la materia. Con ello concretaremos, en la sección 3, el concepto de CD para la terminología y específicamente para su extracción.

En la parte más descriptiva, iremos viendo que la base para la extracción radica en los patrones definitorios, los cuales detallaremos en la sección 4. En la sección 5 nos concretaremos en la definición, su tipología y el papel que juegan las predicaciones verbales en los tipos de definición. En la sección 6 continuaremos precisando los CDs sobre su extensión como unidad discursiva.

Con todos estos elementos, ya en la sección 7 veremos la metodología desarrollada para nuestro extractor de CDs. El extractor tiene como entrada una lista de patrones verbales definitorios y como salida los CDs clasificados por los tipos de definición. En la sección 8 veremos otra forma de agrupar los CDs por sus características semánticas, con lo que se mejora su extracción.

Como parte más aplicada, en la sección 9 describiremos el corpus utilizado a lo largo de nuestras investigaciones y la evaluación de algunos resultados obtenidos. En la sección 10 encontraremos como un resultado concreto la descripción del Corpus de Contextos Definitorios. Luego seguiremos, en la sección 11, con tres aplicaciones específicas de la utilización de CDs dentro del Grupo de Ingeniería Lingüística, entre las que encontramos el banco de conocimientos léxico para el diccionario onomasiológico y un sistema que aplica los resultados de esta investigación para realizar búsquedas en línea.

Cabe mencionar que la investigación en su conjunto forma parte de varios proyectos que han culminado en la publicación de algunos artículos y en diversas tesis, desde licenciatura hasta doctorado, en las áreas de lingüística, ingeniería de la computación y lingüística computacional.

Tendremos un reconocimiento a quienes han contribuido con sus estudios particulares y, para finalizar, las referencias, tanto las publicadas a lo largo de la investigación, como las que han sido base para el desarrollo de la misma.

2. La extracción de información terminológica y conceptual

Una de las áreas que dentro de la inteligencia artificial ha tenido un gran desarrollo en los últimos años, es la que se refiere al diseño de sistemas automáticos de extracción de información (EI). Este proceso, como señala Wilks [83], puede ser visto como el núcleo principal de las actuales tecnologías del lenguaje, de ahí que resulte necesario contar con sistemas de cómputo capaces de buscar, localizar y brindar información relevante de cualquier tipo a un usuario.

Se puede definir entonces a la EI como un proceso por el cual un sistema de cómputo busca de manera selectiva una serie de estructuras o combinaciones de datos, los cuales se encuentran, de manera explícita o implícita, dentro de un conjunto de textos. El resultado de lo anterior es la obtención de información específica que proporciona un conocimiento asociado a tales estructuras o combinaciones de datos [32].

En paralelo con la EI, se ha venido desarrollando un área de investigación enfocada al diseño de sistemas de cómputo capaces de generar y administrar conocimiento obtenido a partir de datos, tal área ha sido denominada *ingeniería del conocimiento* (IC). Una de los objetivos centrales de la IC, es la elaboración de *bases de conocimiento* (BC), las cuales funcionen como un repositorio organizado de información relevante susceptible de proporcionar conocimiento específico a un usuario sobre algún hecho dado.

Entre los aspectos que ha tomado gran relevancia dentro de la IC, cabe señalar la creación de sistemas de *extracción de información terminológica* y *conceptual* (EITC), proyectados para la elaboración de ontologías y diccionarios electrónicos. La generación de estos recursos es uno de los campos más relevantes en los cuales se ha aplicado la IC, en colaboración con otras disciplinas tales como la terminología y la lingüística [36].

La EITC puede ser definida como un conjunto de métodos y recursos tecnológicos orientados a la búsqueda, localización, almacenamiento y administración de términos y conceptos obtenidos de bases de textos relacionadas con un área de especialidad (ingeniería, computación, administración de empresas, periodismo, etc.). La información que se genera a partir de estas bases de texto permite diseñar glosarios, vocabularios y diccionarios electrónicos, herramientas para la traducción automática, sistemas de clasificación e indexación de textos, desarrollo de sistemas expertos y apoyo para labores terminológicas, y otros [36]. Por ello, de acuerdo con Jacquemin y Bourigault [45], se puede ver a la EICT como un área de investigación y aplicación particular y sumamente productiva de la EI.

2.1 Extracción de términos y de conceptos

Si bien existen métodos que han dado buenos resultados para los procesos de extracción terminológica [14, 30, 43], en el caso de la extracción de conceptos resulta un reto más complejo, debido sobre todo a la riqueza de relaciones que se dan a la hora de expresarlos en lengua natural. Las diferencias entre extracción terminológica y conceptual son en gran medida consecuencia de un cambio de paradigma entre una visión de índole normativa sostenida en el modelo propuesto por los diccionarios elaborados bajo los criterios de autoridades académicas [41] y una postura que tome en cuenta aspectos comunicativos y cognitivos subyacentes en la configuración de conceptos [28].

Para tener una distinción pertinente entre términos y conceptos, tomemos en cuenta que el término es una unidad de significación especializada, la cual cuenta con rasgos léxicos particulares (nombres, adjetivos, verbos o adverbios), capacidad referencial y nominativa concreta, así como un significado especializado en un dominio concreto [28, 35].

En contraste, un concepto puede ser visto como una unidad de conocimiento abstracto, la cual contiene una serie de rasgos, características o atributos propios de un objeto, un evento o una relación, con el fin de situarlo dentro del mundo [76]. Al nivel del lenguaje natural, esta unidad es representada por una definición [72, 84]. La definición, de acuerdo con la explicación tradicional de Aristóteles [56], se constituye a partir de dos elementos básicos: un *género próximo* y una *diferencia específica*. El género próximo o *genus* se entiende como un descriptor que hace referencia a la clase a la cual pertenece un objeto o evento, y la diferencia específica o *differentia* son la serie de rasgos propios que distinguen a dicho objeto o evento de los respecto a otros agrupados en su misma clase. En un nivel lingüístico, el género próximo se manifiesta, en el nivel sintáctico, a partir de unidades nominales tales como cuantificadores, determinantes o demostrativos; por su parte, la diferencia específica sería introducida por oraciones subordinadas compuestas

por frases nominales, frases adjetivas o frases prepositivas.

En el caso de los términos, su formación sintáctica implica sobre todo el uso de frases nominales, en particular nombres y adjetivos [43 44, 75], y en algunos casos, construcciones verbales en una función nominativa [49]. En las siguientes secciones veremos de qué manera estos rasgos lingüísticos marcan procesos de reconocimiento y extracción diferentes para términos y definiciones.

2.2 Marco teórico de EITC

Para el desarrollo de los sistemas de EITC es importante tomar en cuenta mecanismos de reconocimiento y extracción de información con determinadas características. En paralelo, de acuerdo con Jacquemin y Bourigault [45], deben considerarse los siguientes aspectos:

- Recopilación, organización y administración de corpus lingüísticos etiquetados, de modo que pueda reconocerse de manera automática ciertos patrones de datos (p.e., asociar términos con frases nominales).
- Implementación de bases de conocimiento léxico, las cuales permiten almacenar, administrar y suministrar conocimientos obtenidos del lenguaje natural, a partir de textos especializados.
- Diseño de sistemas de búsqueda, a partir del uso de lenguajes de programación lo suficientemente robustos como para hacer eficientes los procesos de reconocimiento de datos, la validación de los mismos y la adquisición de conocimiento.
- Empleo de métodos estadísticos, de modo que pueda evaluarse la eficacia o ineficacia de los sistemas de búsqueda de un modo formal. De este modo, el uso de esta clase de recursos es esencial para lograr un sistema de extracción óptimo y potente.
- Aplicación de modelos lingüísticos, los cuales brinden un marco teórico de interpretación pertinente para describir los patrones del lenguaje natural a buscar, así como su formalización de modo que puedan ser comprensibles para cualquier sistema de cómputo diseñado para esta clase de tareas.

Con base en estos puntos, es común que los sistemas de extracción de términos en corpus utilicen un método de aprendizaje automático que consiste en tomar en cuenta los patrones estructurales característicos que conforman tales términos [43]. Después de hacer una primera corrida en un conjunto de textos, con base en estos patrones previamente introducidos, los sistemas localizan y presentan una serie de candidatos posibles. Al final, el conjunto de candidatos es validado de forma manual por un grupo de expertos sobre el área de conocimiento a la cual pertenecen los textos, con miras a determinar cuán exitosa o no fue el proceso ejecutado por dicho sistema.

2.3 Trabajos en EITC

Un tipo de EITC en particular se ha enfocado a obtener información para la organización conceptual de unidades de conocimiento especializadas, así como para la descripción de sus significados. Este tipo de información terminológica suele denominarse *conocimiento definitorio* [7] y 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 [55]. Cabe distinguir dos tipos particulares de extracción automática de conocimiento definitorio.

Por un lado, la extracción de relaciones semánticas (p.e., hiperonimia, hiponimia, holonimia, meronimia, sinonimia, etc.), que en un principio se enfocaron en las definiciones obtenidas de diccionarios en formato electrónico [31, 67]. Posteriormente, buscaron extraer dichas relaciones de corpus lingüísticos tomando en cuenta patrones léxicos sintácticos [42] y luego mediante conceptos formales y el grado de subsunción [37].

Por otro lado, la extracción de contextos definitorios (CDs), con la cual no solo se permite recuperar relaciones semánticas específicas [20], sino también descripciones generales acerca del significado de los términos, y que pueden servir en la elaboración de diversos tipos de recursos terminológicos. A diferencia de la extracción de relaciones léxicas, la de contextos definitorios se realiza únicamente sobre corpus lingüísticos, no solo a partir de patrones léxicos sintácticos, sino de patrones tipográficos y pragmáticos, como veremos más adelante.

El estudio de Alarcón [7] presenta un estado del arte de la extracción de contextos definitorios, a la vez que realiza un análisis contrastivo de diez trabajos en este campo.

- Los trabajos de Rebeyrolle [68, 69] para el francés, que describen una metodología para la extracción de CDs a partir de patrones morfo-sintácticos y que presentan algunas consideraciones sobre la introducción de definiciones en textos de especialidad y el diseño de patrones para su extracción automática.
- El sistema DEFINDER desarrollado por Muresan y Klavans [58] para el inglés, con el fin de extraer definiciones de textos en-línea en el área de medicina mediante la búsqueda de patrones léxicos y tipográficos, en conjunto con una gramática de estados finitos.

- El trabajo de Saggion [73] para el inglés, enfocado a la extracción de definiciones para sistemas de pregunta-respuesta, usando una lista de 50 patrones definitorios.
- Los trabajos de extracción semi-automática de definiciones de la herramienta CORPÓGRAFO, para el alemán, español, inglés, italiano, francés y portugués. A partir de la extracción terminológica, cada término se combina con una serie de patrones definitorios típicos y se formulan así expresiones regulares de búsqueda [64].
- El estudio de Malaisé [52] para extraer lo que denominó definiciones formales, semi-formales e informales, para el francés, a partir de patrones léxicos, de la posición que guardan los términos con los patrones definitorios y de la categoría morfosintáctica de estos últimos.
- El trabajo aplicado de Sánchez y Márquez [74] para textos jurídicos en español, con el fin de extraer definiciones mediante la identificación de patrones verbales recurrentes.
- El estudio de Rodríguez [70], para el inglés, en el que mediante lo que denomina Operaciones Metalingüísticas Explícitas (OMEs), busca extraer unidades de conocimiento especializadas a partir de la detección de fragmentos metalingüísticos en textos de especialidad.
- El trabajo de Storrer y Wellinghoff [78], para el alemán, orientado a detectar y anotar automáticamente definiciones y sus componentes principales en textos técnicos a partir de verbos definitorios y patrones basados en la valencia de dichos verbos.
- El proyecto *Language Technology for eLearning* (LT4eL), patrocinado por la Unión Europea y coordinado por la Universidad de Utrecht, Holanda, en conjunto con 11 instituciones educativas. Una parte central del proyecto se enfocó en desarrollar metodologías para la extracción automática de definiciones para el alemán, búlgaro, checo, holandés, inglés, maltés, polaco, portugués y rumano, con el fin de proporcionar herramientas de ayuda en la elaboración de glosarios [57].
- La aplicación web para el inglés, GlossExtractor, de Navigli y Velardi [59], cuya función es extraer una lista de candidatos a definiciones sobre varios tipos de documentos en Internet.

De este estado del arte, Alarcón observó una similitud en las metodologías y consiste en que todas ellas parten de patrones definitorios para el reconocimiento automático de fragmentos con información definitoria. Resulta notable la coincidencia de usar patrones sintácticos y, en particular, la preferencia de los patrones verbales frente a construcciones sintácticas que incluyen palabras metalingüísticas pero no verbos. Asimismo, resaltó la coincidencia de recurrir no sólo a la búsqueda de patrones definitorios, sino también al uso de filtros de exclusión de contextos no relevantes, así como a la búsqueda y detección de los elementos constitutivos de los candidatos a CDs, es decir, los términos y las definiciones.

3. La noción de contexto definitorio

Para describir el concepto de CD, conviene retomar el estudio de Alarcón [7] sobre algunas aproximaciones de su uso en el ámbito de la terminología, lo cual nos servirá de base para entender lo que se pretende en la extracción automática.

3.1 Aproximaciones del concepto de CD en terminología

Alarcón establece, como punto de partida, lo que De Bessé [34] entiende por *contexto* y que constituye el punto de inicio de cualquier trabajo terminográfico. El contexto es el entorno lingüístico de un término conformado por un enunciado, es decir, las palabras o frases alrededor de dicho término, y que persigue dos funciones básicas: aclarar el significado de un término e ilustrar su funcionamiento. Por tanto, los contextos constituyen un elemento esencial para la descripción de un concepto y resultan indispensables para redactar una definición.

De Bessé distingue los CDs como aquellos contextos donde se aporta información sobre los atributos de los términos. Diferencia los contextos conceptuales como aquellos que se refieren a características sobre las relaciones conceptuales de los términos, en tanto los materiales proveen instrucciones sobre el alcance de los términos y la forma en que éstos operan en un contexto determinado.

Por su parte, Auger [26] divide los enunciados definitorios dependiendo de los tipos de verbos o formas lingüístico-sintácticas que se utiliza en ellos para vincular un término con su respectiva definición. Considera los *enunciados definitorios metalingüísticos* como los elementos que refieren al mismo lenguaje y que utilizan verbos o formas del tipo *llamarse*, *significar*, *el sustantivo*, *el sintagma*, etc. Los *enunciados definitorios lingüísticos*, por otro lado, son los que no se utilizan exclusivamente para referirse al propio lenguaje y se conforman por los verbos o formas lingüístico sintácticas que utilizan elementos del tipo *equivaler a*, *compuesto por*, *características*, *atributos*, etc.

Pearson [62] realiza también un 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.

Meyer [55] propone una categorización simple y más genérica de los tipos de contextos que contienen información conceptual. Meyer define los *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.

Con todo, cabe mencionar que 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.

3.2 El CD en el ámbito de la extracción de información

Con el objetivo de establecer las bases necesarias para la extracción automática de CDs, a lo largo de nuestra investigación hemos establecido que un CD es aquel fragmento textual donde se aporta información que permite comprender el significado de un término, de manera que la información contenida en el contexto puede proporcionar datos sobre sus características y atributos, así como funciones, partes o bien relaciones de éste con otros términos.

Así, delimitamos el prototipo de CD como la estructura discursiva conformada por dos elementos mínimos: un término (T) y una definición (D), los cuales se encuentran conectados entre sí mediante un *patrón definitorio* (PD). Además, los CDs pueden 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). En el ejemplo 1 observamos la definición del término *logística*.

(Ej. 1) <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>.

Vemos que para conectar el término con la descripción de sus características distintivas (*el arte militar que estudia el movimiento…)*, el autor recurre al patrón definitorio que corresponde a la estructura *se define como*. Asimismo, podemos observar el patrón pragmático, *tradicionalmente*, que en este caso indica un matiz especial sobre el significado del término.

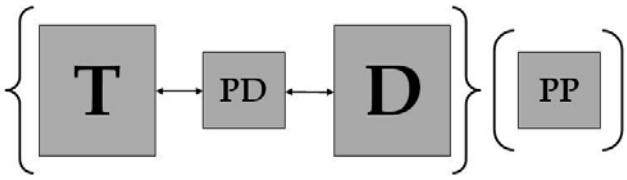

Figura 1: Estructura de un contexto definitorio

En resumen, representamos la estructura de los CDs con el esquema de la Figura 1, en donde los elementos mínimos constitutivos son el término T y la definición D junto con el patrón definitorio PD, que como unidad puede estar modificada por el elemento optativo PP.

3.2.1 Clasificación de CDs

Con base en amplias observaciones sobre la ocurrencia de CDs en diferentes tipos de textos, hemos realizado una clasificación de CDs, tomando en cuenta la presencia o ausencia de una serie de claves tipográficas y sintácticas recurrentes que se utilizan para conectar al término con la información definitoria que se introduce sobre ellos [18, 22].

CDs tipográficos. Los contextos más simples son aquellos que contienen sólo marcas tipográficas para unir al término con la definición, o bien cuya misma tipografía textual se usa para resaltar cualquiera de estos elementos. Este tipo de CDs ocurre tradicionalmente en diccionarios y glosarios, aunque, como refieren Pearson y Meyer, también es común encontrarlos en textos especializados.

(Ej. 2) Diseño: Desarrollo de configuraciones para la resolución de algún problema en base y sujetándose a sus restricciones.

(Ej. 3) *IMPACTOS AGREGADOS SOCIALES ¶* Los que impactan a la sociedad, produciendo, por ejemplo, la perturbación de las relaciones familiares.

En el ejemplo 2, el término *diseño* se presenta en negritas y se liga a la definición, en cursivas, mediante *dos puntos*. En el ejemplo 3, 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 ¶.

CDs sintácticos. Otro tipo de CDs igualmente simples son aquellos en donde el término se une a la definición mediante una estructura sintáctica, generalmente una frase verbal, aunque también es común encontrar marcadores reformulativos. En estos casos no se incluye ningún tipo de marca tipográfica para resaltar los elementos constitutivos de los CDs.

> **(Ej. 4)** 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.
>
> **(Ej. 5)** 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.

En estos dos ejemplos notamos que si bien no se recurre a la tipografía textual para resaltar la presencia del término o la definición, sí se utilizan otros patrones. El ejemplo 4 es prototípico del uso del verbo *ser* más un determinante, lo que se conoce como relación ISA, que aquí se usa para expresar la definición del término *operador logístico*. En el ejemplo 5 tenemos un caso en que para definir el término, *ramal*, se utiliza una estructura sintáctica formada por el verbo *definir* más el adverbio *como*.

CDs mixtos. Este tipo de patrones son una combinación de los dos anteriores, ya que se emplea una frase verbal o un marcador reformulativo como conector entre el término y la definición, pero además se resalta tipográficamente la presencia de cualquiera de estos dos elementos.

> **(Ej. 6) La energía primaria**, por definición, es aquel recurso energético que no ha sufrido transformación alguna, con excepción de su extracción.

En el ejemplo 6 observamos una estructura más sólida que en los ejemplos anteriores, pues aquí se utilizan elementos que permiten resaltar visual y gramaticalmente la presencia de un contexto con información definitoria.

CDs complejos. Estos representan los casos donde en un CD se definen dos o más términos.

> **(Ej. 7)** Por lo anterior, se llegan a distinguir dos tipos de sistemas interactuantes, responsables por la mayor parte de la problemática de desastres: el *afectable* y el *perturbador*. El primero, denominado **SA,** se define como el sistema donde pueden materializarse los desastres debido a la perturbación al que está expuesto; en términos generales, está integrado por la sociedad y los componentes que necesita para su subsistencia, incluyendo el medio ambiente; mientras que, en el contexto particular, puede ser una ciudad u obra civil. El otro, denominado **SP**, responsable por la perturbación, se define como el sistema capaz de producir calamidades, tales como sismos, incendios, explosiones, inundaciones y contaminación.

En 7 se muestra un tipo de casos que, si bien no ocurren en un gran porcentaje con respecto a los demás, nos permiten ver la complejidad de formas en que pueden introducirse CDs en textos especializados. En el párrafo podemos hallar dos términos: sistema afectable (SA) y sistema perturbador (SP), los cuales aunque no aparecen explícitos, se encuentran resaltados en cursivas y en negritas. Asimismo, encontramos la presencia de estructuras sintácticas que nos permiten inferir los términos, la relación entre ellos y las definiciones dadas por el autor. Aquí tenemos un caso claro de referencias anafóricas, la cual veremos a detalle más adelante.

4. Tipología de patrones definitorios

Un elemento clave en el proceso para reconocer CDs de forma automática lo constituye la identificación de los patrones que se emplean para conectar al término con su definición o para resaltar visualmente su presencia dentro del texto. Entre los elementos de CDs mencionamos estos patrones, llamados *patrones definitorios*.

Encontramos dos clases generales de patrones definitorios: los tipográficos y los sintácticos. En los últimos, y de acuerdo con los elementos que se presenten en el patrón, podemos encontrar patrones verbales y/o marcadores reformulativos. Recordemos que, con base en la clasificación de CDs, estos patrones no son excluyentes, puesto que pueden darse por separado o en conjunto.

4.1 Patrones tipográficos

La tipografía de un texto es un recurso que sirve como ayuda visual para identificar fácilmente los elementos importantes y diferenciarlos del resto del texto común. En muchos casos, los términos tienden a ser frecuentemente resaltados. Muchas veces ocurre que la definición también se encuentra señalizada con algún elemento tipográfico o con alguna tipografía específica. En este sentido, los patrones tipográficos se utilizan ya sea para resaltar a los elementos constitutivos mínimos de los CDs o bien para conectar dichos elementos.

> **(Ej. 8) Desastre**. *Perturbación de la actividad normal que ocasiona pérdidas o daños extensos o graves*.

(Ej. 9) MITIGACION: Disminuir los efectos de los impactos de las calamidades.

(Ej. 10) *Calamidad* ¶ Acontecimiento que puede impactar al sistema afectable y transformar su estado normal o deficiente en un estado de desastre.

En estos ejemplos, todos los términos están resaltados, ya sea en negritas, mayúscula o cursiva. En 8 y 9, el término se une a la definición a partir de un signo de puntuación, mientras que en 10 la definición aparece después de un salto de párrafo. En este último ejemplo, además de estar el término en cursivas, su presencia se hace más notoria por el hecho de aparecer en un párrafo anterior a modo de título.

Alarcón [6, 9, 24] encontró que las tipografías textuales más recurrentes para resaltar los elementos constitutivos mínimos de los CD son: 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, encontró que los más usados son dos puntos, punto y guión, o punto y seguido.

4.2 Patrones sintácticos

Un camino para extraer CDs de manera automática en textos de especialidad consiste en identificar las estructuras sintácticas recurrentes tanto de los elementos mínimos constitutivos como de los conectores que unen a estos dos elementos. Alarcón [7] describe dos patrones sintácticos que sirven para conectar el término con su definición. Cuando dichos conectores tienen como núcleo un verbo, tenemos entonces un *patrón verbal definitorio* (PVD). Cuando se emplean otro tipo de formas sintácticas cuya finalidad es establecer una reformulación de una idea o concepto, y que se utilizan para esclarecer el significado de un término, tenemos *marcadores reformulativos*.

4.2.1 Patrones verbales definitorios

En CDs suelen utilizarse construcciones sintácticas verbales para unir a un término con su definición, a la vez de referir atributos y características conceptuales de dicho término [2]. Algunos de estos verbos son comúnmente considerados como verbos *metalingüísticos*, esto es, se emplean para referirse al propio lenguaje, como ocurre con *definir*, *entender* o *denominar*. También encontramos verbos muy comunes que podría decirse son de lengua general, empleados en diferentes situaciones comunicativas no solo definitorias, como los verbos *ser* y *considerar*.

Ocurren dos tipos de construcciones sintácticas verbales: En la más sencilla sólo se emplea un verbo de manera aislada, como *entendemos* o *definimos*. En la más compleja se recurre a una serie de partículas gramaticales, siendo de las más comunes el pronombre impersonal *se* en posición proclítica o enclítica en relación con el verbo definitorio, las preposiciones *a* o *por*, y el adverbio *como*. Algunas de las construcciones con estas partículas podrían ser: *se entiende por*, *se denomina a*, *definirse como*, etc.

(Ej. 11) En este sentido, el estado de un sistema se define como[1] 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.

(Ej. 12) Se denomina "equipo de salud" a todo el personal del hospital que tiene una función directa o indirecta para el paciente.

(Ej. 13) 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 observamos que se introduce información definitoria a partir de los verbos *definir*, *denominar* y *ser*. Asimismo, la ocurrencia del pronombre *se* para los dos primeros verbos, *definir* y *denominar*, y el adverbio *como* y la preposición *a* para formar los patrones *se define como* y *se denomina a*. En el ejemplo 13, tenemos la combinación *ser* + *un*, estructura prototípica para definir un término.

4.2.2 Marcadores reformulativos

Al mismo nivel sintáctico e igualmente útiles para desarrollar una metodología de extracción automática de CDs, existe otro tipo de conectores que no consta de un núcleo verbal, pero que igualmente sirve para conectar al término con su respectiva definición. Este tipo de conectores o patrones sintácticos, que denominamos como *marcadores reformulativos*, conforman un proceso de reformulación en el que se explica el significado de un término a partir de estructuras sintácticas no verbales y, en el caso de los CDs, sirven para referirse a los términos como elementos del propio lenguaje.

Estos marcadores permiten retomar elemento de un discurso para presentarlo de otra forma, garantizan la cohesión textual y puntualizan el significado de algunos enunciados presentados anteriormente [27].

En el grupo de marcadores reformulativos podemos encontrar, entre otras estructuras: *por*

[1] Para distinguir de la tipografía original de los ejemplos, a partir de entonces utilizaré el subrayado para resaltar la parte de texto de interés.

ejemplo, *es decir*, *esto es*, *en otras palabras*, *dicho de otra manera*.

(Ej. 14) *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.

(Ej. 15) El índice secundario es a menudo un índice denso, es decir, contiene todos los valores posibles de la clave primaria.

En 14 se utiliza el marcador *esto es* como conector entre el término *pronóstico de daños* con la definición. En 15 tenemos una reformulación para explicar que el término *índice secundario* implica que *contiene todos los valores posibles de clave primaria*.

4.3 Patrones pragmáticos

En textos especializados es común encontrar, además de la definición, otro tipo de información relevante para entender al término dentro del contexto en el cual aparece. Esta información describe el uso de los términos y manifiesta explícitamente las condiciones de uso o de alcance de dicho término, como son el ámbito temático, la ubicación geográfica, las instituciones que utilizan el término, el nivel de especialidad, o la frecuencia de uso, entre otras características pragmáticas [29].

Este tipo de patrones, que denominamos *patrones pragmáticos* (PPR), son muy útiles, junto con los patrones verbales, para identificar un posible CD dentro del texto cuando no existen patrones tipográficos. También 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.

Este tipo de patrones, que denominamos *patrones pragmáticos* (PPR), los dividimos en tres clases generales: 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*.

(Ej. 16) Inicialmente, Rosch definió el prototipo como el ejemplar que mejor se reconoce, el más representativo y distintivo de una categoría (…)

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*.

(Ej. 17) Por ejemplo, la unidad de longitud – el metro - se definió en 1889 como la longitud de una determinada barra de platino iridiado (…)

Por último, los instruccionales consisten en 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 cuál es su alcance en un contexto determinado.

(Ej. 18) Desde el punto de vista genético, el desarrollo puede definir se como «un proceso regulado de crecimiento y diferenciación resultante de (…)

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. Con todo, podemos decir que las estructuras más recurrentes están conformadas 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*).

5. Análisis lingüístico de definiciones

El objetivo de extraer CDs es tener un repositorio de términos y sus correspondientes definiciones debidamente agrupadas según el tipo de información definitoria, lo que constituye la tipología de definiciones. Posteriormente veremos que esta tipología va íntimamente ligada con el patrón verbal definitorio, el cual presenta una estructura sintáctica precisa.

5.1 Tipología de definiciones

Nuestra tipología de definiciones identificables en CDs se sustenta en el modelo analítico [3, 24], en el hecho de que se haga explícito cuál es el género próximo y/o la diferencia específica, como se observa de la figura 2.

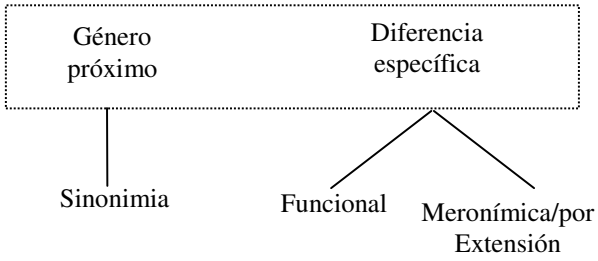

Figura 2: Tipología de definiciones

A partir de la relación observada entre la presencia y/o ausencia del género próximo y diferencia específica, así como entre el tipo de predicación que introduce y asocia a la definición con un término, se observan cuatro tipos de definiciones básicas con los siguientes rasgos:

- **Definición analítica o aristotélica:** se da una definición de este tipo cuando la predicación verbal introduce de manera explícita tanto el género próximo como la diferencia específica. El género próximo puede ser representado en forma de frase nominal, mientras que la diferencia específica puede expresarse en forma de algún tipo de frase (p.e., prepositiva, adjetiva o adverbial), o de oración subordinada introducida por alguna partícula de relativo (que/la cual/el cual/cuyo, quien, etc.). Por ejemplo: *Un algoritmo* es *un conjunto de instrucciones que se ocupa para una computadora*.
- **Definición sinonímica:** se da cuando la predicación introduce una definición que únicamente hace explícito el género próximo, sin considerar ningún tipo de diferencia específica, por lo que se establece una equivalencia conceptual con el término que es definido, p.e., *un maremoto* equivale a *un tsunami*.
- **Definición funcional:** se da cuando se reconoce únicamente la presencia explícita de la diferencia específica, la cual describe como rasgo distintivo de un objeto su función en un contexto dado. Por ejemplo: *una computadora* sirve para *procesar problemas y resultados lógicos, matemáticos y/o estadísticos*.
- **Definición extensional:** se presenta cuando la predicación introduce una definición en donde se explicita la diferencia específica (sin mencionar el género próximo). La clase de información conceptual asociada a estas definiciones puede ser de dos tipos: a) enumeración de las partes o componentes que integran un objeto; b) listado de todos aquellos objetos que conforman un conjunto. Por ejemplo: *una computadora* cuenta con *un hardware, un software, así como una serie de unidades periféricas.*

Definición	Verbo	Adverbio o preposición	Unidades nominales	Predicación
Analítica	Referir Representar Ser Significar	*A*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria
	Caracterizar Comprender Concebir Conocer Considerar Definir Describir Entender Identificar Visualizar	*Como* *Por*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Secundaria
Sinonímica	Denominar Equivaler Llamar Nombrar Ser	*También* *A* *Igual a* *Similar a*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria
Funcional	Emplear (se) Encargar Funcionar Ocupar Permitir Servir Usar Utilizar	*De* *Para*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria
Extensional	Componer Comprender Consistir Constar Contar Constituir Contener Incluir Integrar Es/son parte Es / son + **:** (dos puntos)	*De* *Por* *Con*	*Artículos indefinidos* *Artículos definidos* *Determinantes* *Cuantificadores*	Primaria

Tabla 1: Predicaciones verbales en CDs

5.2 Sintaxis de las predicaciones verbales

En el estudio de Aguilar [1, 4, 5] se observó que las cuatro clases de definiciones anteriores mantienen estrecha relación con el verbo definitorio. Así, en función del verbo que opere como núcleo de una predicación, existe un patrón sintáctico en donde el género próximo y la diferencia específica se sitúan en posiciones de sujeto, objeto o predicado. Por ejemplo, en relación con el verbo *ser*, se observa un patrón sujeto + predicado, en donde el sujeto representa al término a definir y el predicado introduce la definición:

(Ej. 19) (Un algoritmo)$_{Suj}$ es (un conjunto de instrucciones para una computadora)$_{Pred}$

(Ej. 20) (Turing)$_{Suj}$ define (algoritmo)$_{Obj}$ como (un conjunto de instrucciones para una computadora)$_{Pred}$

Aguilar analizó que los patrones predicativos que funcionan como conectores entre términos y definiciones en CDs muestran una constante frecuencia de uso a la hora de introducir el término y su definición.

En un plano general, existe una secuencia de organización sintáctica entre término, verbo y definición que se establece mediante el patrón predicativo: el término puede ocupar la posición de sujeto u objeto, el verbo como núcleo de la predicación, en tanto la definición es introducida por el predicado asociado al sujeto.

En un plano particular, los verbos que operan como núcleos de las predicaciones establecen una relación con la definición expresada por el predicado, de tal suerte que el verbo puede influir en la selección del tipo de definición que es introducida en un CD.

En un nivel de construcción sintáctica de un CD, una predicación organiza en qué posiciones pueden situarse el término y la definición en torno al verbo que opera como núcleo de dicha predicación. Entrando en mayores detalles, en este nivel se dan clases de secuencias de organización:

Una secuencia del tipo término + verbo + definición, en donde el término equivale al sujeto, el verbo funge como núcleo, y la definición es representada por el predicado que se asocia al sujeto, p. e.: *un error de programación es un fallo en la semántica de un programa*.

Una secuencia del tipo autor + término + verbo + definición, en donde el sujeto indica quién es el autor o los autores de una definición, el término equivale al objeto de la predicación, el verbo opera como núcleo, y la definición es introducida por el predicado asociado al objeto, p. e.: *Turing definió la inteligencia artificial como aquella inteligencia exhibida por artefactos creados por humanos*.

5.3 Variaciones tipológicas

En nuestra tipología de definiciones tenemos cuatro tipos. En la analítica se expresan los caracteres genéricos así como los diferenciales de una cosa, es decir, el género próximo y la diferencia específica. La extensional expresa las partes y componentes o el tamaño del término que se define. La funcional expresa la función, utilidad o el fin con el que se utiliza el concepto representado por el término en el CD. Tanto la extensional como la funcional tienen como rasgo característico compartido el carecer de género próximo.

Sin embargo, sucede en realidad que algunos CDs con definición de tipo analítica pueden tener como diferencia específica la extensión o la función del término que se define, ya que de esta manera la extensión o la funcionalidad de algún objeto permite tener un conocimiento más amplio del objeto que se está definiendo.

Sánchez [17] observó que la diferencia específica que expresa la función de un término definido puede ser introducida por una preposición o por el uso de sintagmas preposicionales. En particular, estudió la funcionalidad de un término introducida por la frase preposicional o patrón sintáctico *para* + infinitivo.

(Ej. 21) Así, un molino de viento es un artefacto útil para captar y aprovechar parte de esta energía

Del ejemplo 21 observamos que el término *molino de viento* tiene la función de *captar y aprovechar parte de esta energía*.

En términos generales, la preposición es una partícula o elemento sintáctico utilizado para establecer un tipo de relación entre un elemento A y un elemento B, donde A y B pueden ser oraciones o segmentos de una oración [38, 51, 61]. En función de la relación que establece con los términos que une, la preposición *para* es de tipo nocional [38], esto es, como su nombre lo indica, incluyen nociones como causa, finalidad, destinatario, instrumento, compañía, modo, etc.

En su investigación, Sánchez observó que la preposición *para* seguida de un verbo en infinitivo aporta, en un alto porcentaje, funcionalidad del término que se define, salvo las siguientes excepciones:

Caso 1.- El patrón *para* + *infinitivo* se encuentra fuera del CD. Como veremos en la sección 6.2, la extensión de un CD puede acabar antes del punto, por lo que hay que tomar en cuenta las reglas de delimitación para asegurar que el patrón se encuentra dentro de los límites del CD. Por ejemplo:

(Ej. 22) La máquina virtual es un ordenador con una pila sencilla; los programas están estructurados para permitir que los clientes verifiquen la existencia de referencias ilegales ni errores gramaticales en el código descargado

En 22 observamos que el CD termina en el punto y coma, por lo que el patrón no aporta información funcional al término *máquina virtual*, sino a *programas*.

Caso 2.- El patrón se encuentra alejado del término o del género próximo o del término por una sucesión en más de dos grados de sintagmas preposicionales (sp). Por ejemplo:

(Ej. 23) La impresora es el órgano típico $(\text{de salida})_{sp1}$ $(\text{de información})_{sp2}$ $(\text{del ordenador})_{sp3}$ para ser utilizada en la empresa

Caso 3.- Se encuentra separado el término o la diferencia específica del patrón mediante la introducción de una oración relativa; p.e.:

(Ej. 24) Un "analizador sintáctico" es un programa con el <u>que</u> se pueden comprobar series de caracteres <u>para ver</u> si son fórmulas bien formadas de un lenguaje dado.

Caso 4.- Existe un elemento que cambia la funcionalidad del patrón, ya sea modificándolo, mediante un adverbio de negación, o bien negándolo, mediante un adjetivo con carga semántica negativa; por ejemplo:

(Ej. 25) El problema de la luz es que es una mezcla de varias frecuencias, y por tanto <u>poco útil para ser</u> empleada como medio de comunicación, excepto si usamos una luz monofrecuencia obtenido por medio del láser y un conductor conocido como fibra óptica

Así, el estudio de Sánchez nos permitió observar que la tipología propuesta de definiciones es flexible y que el patrón sintáctico *para + infinitivo*, inserto en la definición de un CD de tipo analítico, aporta información de funcionalidad del término que se define. Asimismo, definió algunas reglas de exclusión a este patrón, con lo que es posible mejorar la extracción automática de CDs.

6. La extensión de un CD

Hemos visto los elementos constitutivos de los CDs y la forma en que se construyen. Sin embargo, falta mencionar la extensión de los CDs, esto es, los límites de inicio y finalización del fragmento textual que contiene la definición completa de un término. Como unidades discursivas, tienen estructuras distintas, sin tener un número de palabras fijo y con elementos que pueden presentarse en diferente orden.

En un principio puede considerarse el párrafo como la estructura textual para establecer la extensión de un CD, pero como veremos a continuación, cabe la posibilidad que la extensión vaya más allá de un párrafo, o bien que en el mismo párrafo exista más de un CD.

6.1 Las anáforas en la expansión de CDs

Un tema interesante, pero complejo tratándose de CDs, es la forma en que las relaciones anafóricas intervienen para su extracción. La anáfora es comúnmente el término que se emplea para hacer referencia a algo que anteriormente ya fue mencionado, y considera cualquier expresión, palabra o frase que recupera algo previamente enunciado. El análisis de relaciones anafóricas juegan un papel determinante para la obtención completa de un CD.

(Ej. **26)** <u>Este</u> consta de un banco de capacitores sumergidos en aceite en un recipiente de porcelana y conectados en serie (…)

En efecto, en (26) vemos un pronombre demostrativo en representación del término del contexto. Si solo extrajéramos este CD incompleto sería imposible determinar a qué término corresponde la definición: "un banco de capacitores sumergidos en aceite en un recipiente de porcelana (…)". Con base en lo anterior, es evidente la necesidad de una *extensión* de este tipo de casos. Por *extensión*, entendemos el tamaño del fragmento textual que contiene el CD completo, con término y definición, mientras que por *expansión* se comprenderá la pertinencia de acudir al documento de origen del contexto con el objetivo de verificar la extensión del CD.

Con la finalidad de resolver este problema, primero es necesaria la identificación de los tipos de relaciones anafóricas que operan con CDs. Con este fin, Benítez [13] realizó un estudio profundo donde se describen de manera completa relaciones anafóricas presentes. En dicho estudio Benítez encontró que, principalmente, son cuatro las expresiones más frecuentes en CDs.

En el primer grupo se encuentran algunos pronombres demostrativos (esto, aquellos, esta), personales (lo, le), relativos (la cual, lo cual, que) e impersonales (el primero). La frecuencia de esta clase de expresiones no es muy alta, pero son las más comunes en la ocurrencia de candidatos incompletos, como puede verse en el ejemplo 27, ya que la expresión apunta a un antecedente omitido en la extracción automática.

(Ej. 27) Esto es lo que se entiende por enfoque genético de la medicina o "medicina genética".

El segundo grupo abarca los sintagmas nominales con determinante demostrativo, los cuales son expresiones con valor anafórico porque refieren a una parte anterior en el texto. Por ejemplo:

(Ej. 28) <u>Estos elementos</u> son parte constitutiva de los compuestos que forman la base material para la vida (…)

El tercer grupo lo conforman las expresiones mixtas (pronombres y sintagmas nominales con demostrativo) en las que se muestran cadenas de anáforas o anáforas muy cerca de otras, es decir, que las cadenas de referencia se manifiestan con pronombres y sintagmas nominales que se encuentran en una relación anafórica.

(Ej. 29) <u>Esta concepción</u> es <u>lo</u> que se conoce con el nombre de materialismo histórico.

En 29 se oberva cómo la expresión anafórica, representada por el pronombre *lo* hace referencia al sintagma nominal con demostrativo *esta concepción,* que a su vez tiene como referente al verdadero término de la definición.

El último grupo está constituido por las expresiones ligadas a una entidad previamente enunciada, las cuales pueden ser sintagmas nominales, elipsis o marcadores discursivos.

(Ej. 30) El primer grupo es típico de los buques rápidos y consiste en olas de gran periodo, que sufren poca dispersión al alejarse del barco (…)

Una vez llevada a cabo la observación del corpus y después de realizar la clasificación de los elementos más frecuentes en las relaciones anafóricas, Benítez realizó el diseño de etiquetas XML para la identificación de relaciones anafóricas siguiendo los patrones de formación de las etiquetas ya establecidas para el CORCODE.

6.2 La delimitación del CD

Para la conformación de un sistema de extracción conceptual, es importante tener en cuenta que no todos los contextos definitorios son iguales, esto es, que no todos los CDs tienen una misma estructura en la que comienzan con el término y terminan en el primer punto después de la definición.

Para reconocer automáticamente la extensión de un CD dentro de un texto se tomó en cuenta un criterio básico inicial, que consiste en delimitar un contexto en el primer punto. Si bien este criterio es funcional en gran medida, no siempre obtiene buenos resultados como se muestra a continuación.

(Ej. 31) La "acción" es entendida como la conducta intencionada proyectada por el agente; en cambio el "acto" es definido como la acción cumplida.

En 31 podemos ver que la definición del término "acción" acaba antes del primer punto y antes de la introducción del término, "acto".

Con la finalidad de evitar información que no sea parte del CD y así mejorar el sistema de extracción, se requiere del planteamiento de reglas lingüísticas que permitan delimitar definiciones automáticamente cuando éstas terminan antes del primer punto.

Hernández [15] realizó un estudio para delimitar contextos en definiciones de tipo analítico; es decir, con género próximo y diferencia específica, debido a que cada tipo de definición requiere de un propio estudio y reglas particulares. En su investigación, observó y analizó dos tipos de patrones lingüísticos de delimitación.

6.2.1 Patrones que rompen con la definición

Un primer tipo de patrones lingüísticos que delimitan un CD tienen la característica de que lo que viene después del patrón rompe por completo con lo que se estaba expresando en la definición sobre el término, esto es, marcan la introducción de un nuevo término o foco dentro del discurso, el cual ya no pertenece al CD. Cinco de los patrones encontrados son:

Patrón 1.- Por tanto/por lo tanto. Este marcador discursivo, considerado como conector consecutivo [53], introduce una consecuencia o una conclusión en el elemento siguiente. Como podemos observar en (32), la intensión que se introduce en la definición se ve concluida con el enunciado posterior al patrón *por tanto*.

(Ej. 32) Finalmente, no debemos olvidar que el <T>dengue</T> <PVD>es</PVD> <D>un virus que puede replicarse en células de mamífero y en células de mosquito,</D> por tanto, los aspectos antes descritos para células de humano pueden también estar operando en el mosquito vector.

Patrón 2.- Sin embargo + FN. Este patrón se encuentra constituido por un marcador discursivo de tipo conectivo contra-argumentativo [53], pues vincula dos miembros, de tal modo que el segundo se presenta como supresor o atenuador de alguna conclusión que se pueda obtener del primero.

(Ej. 33) <PP>En general</PP>, el <T>ácido nucleico </T> <PVD> es </PVD> <D> una molécula única de hélice simple o doble</D>; sin embargo, ciertos virus tienen el material genético segmentado en dos o más partes.

Patrón 3.- En cambio + FN. Este patrón compuesto por un marcador conector contra-argumentativo seguido por una frase nominal muestra un contraste entre los términos que se definen. En el ejemplo 34 los términos "adenina" y "citosina" son contrapuestos semánticamente a través del marcador que funciona como conector.

(Ej. 34) La <T>adenina</T> y la <T>guanina </T> <PVD> son </PVD> <D>bases púricas </D>, en cambio la citosina y la timina son bases pirimidínicas.

Patrón 4.- Mientras que + FN. Con el marcador contra-argumentativo *mientras que* se oponen dos enunciados distintos y en nuestro caso los elementos contrapuestos son CDs. En 35, el patrón (mientras que + FN) se encarga de definir hasta dónde llega el primer CD cuyo término es "hiperalgesia primaria".

(Ej. 35) La <T>hiperalgesia primaria</T> <PVD> se concibe como </PVD> <D> el aumento de la respuesta al estímulo doloroso en la región de la lesión </D>, mientras que la hiperalgesia secundaria es aquella que se extiende para áreas adyacentes.

Patrón 5.- (En tanto/en tanto que) + FN. El marcador conector contra-argumentativo *en tanto* se encuentra funcionando de la misma forma que *en cambio* y *mientras que* cuando les sigue una frase nominal y están cerca de contextos definitorios en ámbitos de especialidad.

6.2.2 Patrones que continúan con información relevante

El segundo tipo de es aquel en donde la información que sigue a la regla o patrón lingüístico es pertinente para el CD, ya que amplía, reformula o explica la información definitoria del mismo término, pero ya no constituye ninguna de las partes formales de la definición analítica. El beneficio que aportan estos patrones consiste en que la información que se introduce puede ser parte o no del CD, según las necesidades y propósitos del sistema de extracción. Hay que tomar en cuenta que, al aportar información enriquecedora para el CD, no pueden ser considerados como patrones de delimitación como tal, sino más bien como indicadores del final de la diferencia específica.

(Ej. 36) La <T> adolescencia </T> <PVD> es definida como</PVD> <D>una etapa del ciclo vital entre la niñez y la adultez, que se inicia por los cambios puberales</D> y se caracteriza por profundas transformaciones biológicas, psicológicas y sociales, muchas de ellas generadoras de crisis, conflictos y contradicciones, pero esencialmente positivos.

En 36 podemos ver que el término "adolescencia" tiene dos definiciones. En la primera es definida como "una etapa del ciclo vital entre la niñez y la adultez, que se inicia por los cambios puberales" y en la segunda es caracterizada por "profundas transformaciones biológicas, psicológicas y sociales, muchas de ellas generadoras de crisis, conflictos y contradicciones, pero esencialmente positivos". El patrón delimita la extensión de la primera definición, aunque lo que viene después sigue siendo relevante para el CD y debe por tanto tomarse en cuenta.

Entre los patrones que podemos encontrar de este tipo tenemos: *por ejemplo*, *como por ejemplo*, *tal como*, *o sea*, *es decir*, *y*+PVD.

7. La extracción automática de CDs

El objetivo que perseguimos con el análisis previo es lograr la extracción automática de CDs en español a partir de textos de especialidad. Gracias al conocimiento lingüístico de la conformación de CDs nos fue entonces posible desarrollar la metodología pertinente.

Nuestra metodología para extraer CDs está basada en reglas lingüísticas y consiste en la búsqueda automática de ocurrencias de patrones definitorios, específicamente PVDs [10, 18, 19]. El Extractor de Contextos Definitorios (ECODE), que fue desarrollado por Alarcón [7], abarca un procesamiento automático de los candidatos a CDs: primeramente, un filtro de contextos no relevantes, esto es, aquellos contextos donde, a pesar de tener un PVD, no se define un término; luego, la identificación de los elementos constitutivos del CD, es decir el término y la definición; finalmente, una ponderación de resultados para determinar cuáles son los mejores CDs propuestos por el sistema.

Para obtener CDs se debe tener como entrada un corpus anotado con etiquetas de partes de la oración (POS). De ahí, el proceso general consiste en tres pasos: la extracción de candidatos, el análisis de candidatos y la evaluación de los resultados.

7.1 Extracción de candidatos

El proceso principal del ECODE lo constituye la extracción de candidatos, la cual requiere una gramática de PVD que contiene una serie de parámetros:

- Los verbos definitorios a buscar junto con los nexos que los acompañan, ya que un verbo puede estar acompañado o no de diferentes nexos para expresar definiciones de varios tipos. Por ejemplo, el verbo *conocer* asociado con el nexo *como* obtendrá por resultado CDs del tipo analítico, en tanto asociado con el nexo *también* nos dará un CD del tipo sinonímico.
- Las restricciones verbales referentes al tiempo y a la persona gramatical, ya que la información definitoria a recuperar depende del tiempo, de la forma verbal o de la persona gramatical para cada verbo. Como puede verse en los ejemplos 37 y 38, el verbo *definir* en primera persona de plural nos traerá información definitoria, pero no así el verbo *contar*.

(Ej. 37) La radiación provoca mutaciones, que definimos antes como cambios en la secuencia de las bases del ADN.

(Ej. 38) Cómo se va a regular la aplicación de estudios de escrutinio conforme contemos con el conocimiento de dichos genes?

- Los patrones contextuales, esto es, la delimitación de las posiciones en las que podrían aparecer el término y la definición respecto al verbo definitorio. Este parámetro es crucial para el ECODE, pues posteriormente será utilizado para identificar los elementos constitutivos de cada CD. Entre algunos de los patrones contextuales tenemos: T+PVD+D, VD+T+NX y PVD+T+D, como se observa en los siguientes tres ejemplos, respectivamente.

(Ej. 39) <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>

(Ej. 40) <PVD> Se ha <VD> definido </VD> <T>el genotipo</T><NX>como<NX> </PVD> <D> la constitución genética del individuo en un locus. </D>

(Ej. 41) <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 del tracto GI.</D>

- Restricciones de distancia entre el verbo y su nexo, pues entre ambos puede aparecer desde un adverbio o un término simple, hasta unidades más complejas como sería un término compuesto más una frase adverbial. Este parámetro debe analizarse con cuidado, pues de lo contrario puede causar que el extractor traiga mucho ruido. En el ejemplo 42 podemos observar un CD con una distancia de 8 palabras, mientras que en 42 tenemos que inclusive se rompe el CD entre el verbo definitorio y el nexo.

(Ej. 42) En 1977, Oshimura et al describieron las deleciones del brazo largo del cromosoma 6 como una anomalía recurrente en leucemias.

(Ej. 43) La clasificación de las distrofias musculares ha ido evolucionando con el tiempo: desde finales del 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 a tener (...)

7.2 Análisis de candidatos

Una vez extraídos los candidatos a partir de los PVD y el empleo de la gramática, el análisis de los CDs incluye dos procesos principales: el primero consiste en eliminar los contextos no relevantes mediante reglas de filtrado y, el segundo, en la identificación de los elementos constitutivos.

El filtro de contextos no relevantes se basa en una serie de reglas lingüísticas y contextuales para determinar los casos en los que es probable que un patrón verbal no esté introduciendo información definitoria. Mientras existen verbos de carácter prototípicamente definitorios, otros se utilizan en una gran variedad de situaciones. Por ello, entre las reglas de filtrado, Alarcón [7, 8, 11] propuso una lista de restricciones basadas en ciertas partículas gramaticales (principalmente preposiciones, adverbios, pronombres y verbos en forma conjugada), y en la posición en las que pueden aparecer dichas partículas adyacentes o dentro del PVD. Por ejemplo:

(Ej. 44) <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.

En 44 tenemos un contexto no relevante debido a la partícula *así* inmediatamente anterior al nexo *como*.

Ya con los candidatos que no fueron filtrados como excepciones, el siguiente paso consiste en la identificación de los elementos constitutivos, esto es, el término y la definición. Para ello, se utiliza un árbol de decisión (Fig. 3) que recurre igualmente a la gramática de patrones verbales. El árbol de decisión, a través de inferencias lógicas, asocia una serie de patrones contextuales para cada verbo definitorio, de forma que dichos patrones indican las posiciones en que puede aparecer el término y la definición.

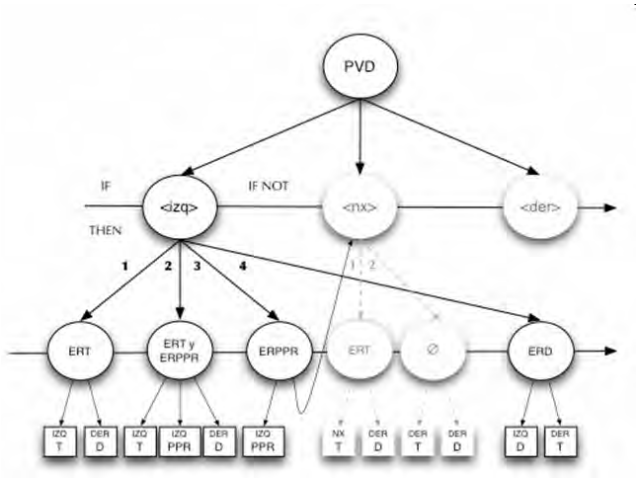

Figura 3: Árbol de decisión para el ECODE

De ahí el procedimiento busca asegurar, mediante el empleo de expresiones regulares, si el elemento se ajusta a la expresión de término, de definición o de patrón pragmático.

(Ej. 45) <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>

En el patrón contextual mostrado en el ejemplo 45 tenemos el PVD *ha sido concebido como,* además de una parte a la izquierda y otra a la derecha. La gramática de patrones verbales señala que, para el verbo concebir, el término puede encontrarse en la posición izquierda pero no en la posición derecha del verbo definitorio. Se asegura

que el término se encuentra en la posición izquierda por una expresión regular de término que pide la presencia de frontera *<IZQ>* seguida de un determinante, un sustantivo y todo lo que esté después hasta la siguiente frontera que es la etiqueta de cierre de la posición izquierda (*</IZQ>*). Continuando el análisis, con una expresión regular de patrón pragmático se identifica *en términos generales* porque empieza con una coma seguida de una preposición y hasta la frontera de cierre. Para el caso de la definición se tiene su expresión regular formada por determinante más sustantivo, delimitado por las fronteras de inicio y de cierre.

7.3 Ponderación de resultados

El tercero y último proceso del ECODE busca evaluar los CDs que resultan después del filtrado de excepciones de contextos no relevantes, y en particular los elementos constitutivos, para ponderar los mejores CDs según la estructura del contexto recuperado automáticamente. Se utiliza una serie de reglas heurísticas que comparan las estructuras sintácticas de los elementos etiquetados como término y definición con sus estructuras prototípicas. Se asigna un valor a cada elemento y un valor global a partir de las combinaciones encontradas. Los contextos que pasen de un umbral determinado serán los que el ECODE arroje como buenos CDs.

Si bien este último proceso permite obtener en primer lugar los mejores candidatos a CDs, cabe advertir que existen riesgos. Ya que este proceso de ponderación se basa en las estructuras sintácticas de los elementos que se van a ponderar, se puede asignar un valor equívoco en caso de también el etiquetado POS contenga errores. Además, las expresiones regulares de término y definición también pueden traernos elementos que no lo son pero que cumplen estructuralmente con las reglas de los buenos candidatos. Entre algunos casos erróneos, Alarcón llegó a encontrar los siguientes términos: *repetitivo*, *una vez*, *la molestia de la ropa interior teñida*.

8. Agrupamiento de CDs

El ECODE proporciona finalmente una lista de CDs asignados a alguno de los tipos de definiciones: analítica, extensional, funcional o sinonímica, y organizada según la probabilidad de que sean en mayor o menor medida mejores CDs.

Además de la clasificación de los CD por su tipo de definición, también pueden ser agrupados según sus características semánticas. Esto es, se pueden agrupar los CD polisémicos por sus diferentes significados o incluso por las características descritas en su definición. En el primer caso, tenemos por ejemplo el término virus, del cual se pueden tener por un lado los CDs correspondientes al área de informática y por otro lado los correspondientes al área de medicina o de biología. En el segundo caso, podemos encontrar por ejemplo los CDs con definiciones analíticas para el término gen, que por un lado lo describen como la unidad de la herencia y por otro como una secuencia de ADN.

Por esta razón, Molina [16] se ha dedicado a la tarea de desarrollar un algoritmo para poder llevar a cabo el agrupamiento automático de CDs según su significado, de tal forma que los resultados de la búsqueda de un término polisémico sean presentados mediante una clasificación semántica.

La ventaja más importante del algoritmo de agrupamiento es ser independiente del idioma, no requiere de ningún tipo de anotación lingüística, como etiquetas POS, tampoco requiere de un conjunto de entrenamiento previo, ni es necesario indicar el número de grupos a generar y, finalmente, a diferencia de otros algoritmos similares como *lingo* [60], el algoritmo aquí descrito es fácilmente configurable, pues depende únicamente de un parámetro: el valor de corte por distancia.

Para la realización del algoritmo de agrupamiento semántico se toman como base los resultados del sistema de extracción de contextos definitorios *ECODE*, el cual entrega un archivo de salida con CDs clasificados según el tipo de definición.

El algoritmo lleva a cabo tres grandes etapas. Dentro de la primera, el texto es procesado hasta llegar a la representación vectorial usando diversas técnicas de procesamiento del lenguaje natural. En la segunda, se calcula la distancia entre cada vector utilizando la matriz de energía textual y, en la última etapa, se aplica el agrupamiento jerárquico con el método de vecino más lejano.

De forma general, los pasos que realiza el algoritmo son los siguientes:

8.1 Preprocesamiento

Con la finalidad de reducir el tamaño del espacio vectorial, se procesa cada archivo en tres etapas: la transformación de signos de puntuación y diacríticos; la eliminación de palabras que no son de contenido y el truncamiento de las palabras.

El primer contacto del texto con el algoritmo será a través de un archivo en texto plano, sin etiquetas, ni ningún tipo de marcaje como XML y HTML. La primera etapa de preprocesamiento consiste en unificar la diversidad de los símbolos gráficos. Por ejemplo, con la intensión de reducir la diversidad de símbolos u, ü y ú, se unificarán bajo el símbolo u.

En la segunda etapa, los textos son filtrados con una lista de paro (stop list), lo cual reduce en gran medida el tamaño del diccionario generado por la colección. También, son eliminados todos los patrones verbales definitorios junto con el término, pues estos elementos aparecen en todas las definiciones de la colección y, por tanto, no contribuyen a constituir un criterio de agrupamiento.

La última transformación consiste en truncar las palabras mediante el algoritmo de Portero [65]. La intensión de esta transformación es unificar en un solo símbolo aquellas palabras que poseen en la misma raíz y que están relacionadas semánticamente. Por ejemplo, las palabras *vivo* y *viviente* se unifican bajo el mismo símbolo *viv*.

8.2 Construcción del espacio vectorial

En esta etapa se construye el espacio vectorial generado por las definiciones, es decir, una matriz concebida como un arreglo de vectores que representan documentos. Un documento es una cadena de longitud arbitraria pero finita de símbolos gráficos denominados entidades léxicas (EL). Entendemos como EL aquella que puede ser representada mediante un símbolo o la unión de varios de ellos. Así, la palabra *manzana* puede representar una EL, o bien una frase como *Estados Unidos Mexicanos*. Asimismo, una EL puede ser un símbolo ininteligible como *Viv* o *A4*. De esta manera, una colección es un conjunto de documentos y un diccionario es una lista de ELs únicas que aparecen en una colección.

8.3 Cálculo de la energía textual

Una vez generada una matriz binaria surge necesidad de comparar definiciones a partir de su representación vectorial. Para esto, es necesario tener un mecanismo de comparación entre textos que funcione como criterio para determinar los grupos semánticos. Con esta finalidad se optó por derivar una medida de distancia a partir de la matriz de *energía textual* propuesta por Fernández, San Juan y Torres Moreno [39]. Esta técnica resulta funcional porque fue concebida desde sus inicios como una aproximación teórica para ponderar las relaciones de significado en textos.

La distancia entre los vectores a partir de la matriz de energía textual se calcula a partir de la siguiente fórmula:

$$DistEner = \frac{máx(D_{ener}) - D_{ener}}{máx(D_{ener})}$$

DistEner es un arreglo que contiene la distancia entre cualesquiera dos documentos *i*,*j* de la colección dado que *E* es una matriz simétrica, esto es, $e_{ji}=e_{ij}$. De esta forma, hemos calculado la distancia entre documentos a partir de la matriz de energía textual. Tenemos, ahora, la posibilidad de utilizar un algoritmo de agrupamiento para generar una estructura de grupos utilizando esta distancia como criterio.

8.4 Agrupamiento de definiciones

Una vez que la proximidad entre textos es calculada, son generados los grupos por medio de un algoritmo jerárquico aglomerativo simple. Un algoritmo de tipo jerárquico ofrece la ventaja de que no requiere que el número de grupos sea especificado previamente.

El método utilizado para comparar los grupos en el algoritmo jerárquico es el método del vecino más lejano (*complete linkage*). Es preferible este método porque genera grupos pequeños, cohesivos y bien delimitados, brindando la posibilidad de mejorar la precisión de los grupos.

El criterio para determinar el número de grupos generados es un valor umbral de corte por distancia. Con dicho valor es posible indicar el valor máximo de distancia que puede haber entre dos grupos. Por ejemplo, si determinados que el valor umbral de corte por distancia es 0.1 significa que aquellos grupos cuya distancia es mayor a 0.1 nos son unificados. Además el algoritmo de agrupamiento jerárquico genera un dendograma que permite calcular coeficientes de comparación entre agrupamientos y representar gráficamente los resultados obtenidos de cada ejecución del módulo.

9. Resultados y evaluación

Hasta ahora he mostrado una síntesis de los estudios que se han realizado en el Grupo de Ingeniería Lingüística para analizar los CDs, clasificarlos, reconocer y precisar sus elementos constitutivos, delimitar su extensión, extraerlos y agruparlos. Hemos visto la metodología de cada uno de estos estudios, pero conviene ahora tener una síntesis de los resultados obtenidos y su evaluación.

9.1 El corpus de estudio

Para los diferentes estudios que se describen en este artículo, se trató de ser consistentes en el empleo de las mismas fuentes, con lo que conformamos los corpus de experimentación, de prueba y de evaluación. Los principales corpus utilizados son los siguientes:

- El Corpus Lingüístico de Ingeniería o CLI [54]. Se trata de un corpus en español orientado al área de ingeniería y desarrollado por el Grupo de Ingeniería Lingüística. Está

conformado por documentos en texto plano (extensión *.txt*), con alrededor de 500,000 palabras (tokens). Se trata de un corpus que reúne textos especializados del área de ingeniería, tales como tesis, artículos, informes, etcétera. Una de las ventajas de este corpus es que los textos usualmente incluyen apartados, ya sea introducción, presentación o bien un capítulo específico, que funcionan como marco teórico donde se definen los términos esenciales para la comprensión del contenido.

- El Corpus Técnico del Instituto Universitario de Lingüística Aplicada (CTIULA) de la Universidad Pompeu Fabra en Barcelona [80]. Este corpus cuenta con 9,542,000 palabras en su sección dedicada al español, al cual se puede acceder a través de su herramienta de búsqueda *BwanaNet*. Está etiquetado con partes de la oración y cuenta con tres opciones de búsqueda: básica, estándar y compleja.
- El Corpus Informático en Español o CIE [50], es un corpus técnico desarrollado para las áreas de informática y ciencias de la computación, con miras a la creación e implementación de un diccionario electrónico en español. Cuenta con alrededor de 500,000 palabras, divididas en 4 sub-corpus: de la revista *PC World Latinoamérica* (PCWLAF), revista *Guía Computación*, *WindowsTI Magazine*, y entradas obtenidas de la Wikipedia en español.

Fuente	Número de CDs
CLI	238
CTIULA	1,361
CIE	562
SkE	5
Google	49

Tabla 2: Corpus de CDs

En menor medida se utilizó el Spanish Web Corpus de la herramienta *Sketch Engine* (SkE) [47], y el motor de búsqueda *Google*.

Como resultado, en total se obtuvieron en total 2,215 contextos, como muestra la tabla 2.

9.2 Evaluación

Como medidas de evaluación se han usado principalmente las tradicionales para los sistemas de recuperación y extracción de información: precisión y cobertura. Como explican Jurafsky y Martin [46], 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.

La precisión se representa entonces como la proporción del número de respuestas válidas propuestas por el sistema, del total de respuestas propuestas por el sistema. La cobertura queda como la proporción del número de respuestas válidas propuestas por el sistema, del total de respuestas del texto.

Cabe advertir que determinar las respuestas válidas resulta complicado en el caso de CDs. En el ámbito de la terminología y lexicografía resulta muchas veces un retos precisar los límites de una definición. Si bien Aguilar [1, 4] profundizó en el concepto de definición, en la práctica resulta muchas veces difícil llegar a consenso sobre el límite de la definición analítica o a precisar el género próximo de la misma.

(Ej. 46) En ecología, biomasa es el término usado para definir el volumen total de materia viva en forma de microorganismos, vegetales, animales, que soporta un ecosistema determinado.

Así, la definición del término biomasa es del tipo analítica, y como tal debe estar constituida por un género próximo y una diferencia específica. Sin embargo, es controversial precisar dónde termina el género próximo, si en total, en materia viva o en animales. Por esta razón, para la evaluación nos apoyamos en estudiantes involucrados en el área de terminología. Para resolver las dudas trabajamos en equipo y discutimos cada uno de los casos.

Como muestra de la evaluación, podemos mencionar la obtenida para el ECODE, en donde se consideró como CD cuando apareciera explícitamente el término y la definición. El corpus de evaluación quedó conformado por contextos definitorios y contextos no relevantes. Alarcón [7, 12] reporta que para la precisión dividió el número de CDs válidos propuestos por el sistema sobre el número de CDs propuestos por el sistema (1783/3309), con lo que quedó un valor de 0.53. Para la cobertura dividió el número de CDs válidos propuestos por el sistema sobre el número de CDs en el corpus (1783/2254), quedando un valor de 0.79. Esto es, se obtuvo una mejor cobertura frente a la precisión. Mientras que se recuperó el 80% de CDs presentes en el corpus, solo un poco más del 50% de lo obtenido era válido.

10. El corpus de CDs

A lo largo de la investigación hemos obtenido un acervo de CDs con lo que podemos construir el CORCODE o Corpus de Contextos Definitorios. Éste va más allá de ser un repositorio de documentos, pues constituye una herramienta valiosa para la terminología y la lexicografía, al

permitir facilitar el proceso de extracción de unidades tales como términos y definiciones.

El CORCODE es un corpus compuesto por CDs enfocados en áreas de especialidad. Actualmente puede consultarse en la página del Grupo de Ingeniería Lingüística un total de 127 CDs.[2] La interfaz de búsqueda permite realizar navegaciones a partir del tipo de término, tipo de definición, tipo verbo definitorio, de marcadores textuales definitorios (comas, dos puntos, comillas, etc.) y de los patrones pragmáticos (autoría, patrones temporales o instruccionales).

Este método de búsqueda se da a partir de un etiquetado en XML que facilita la identificación de las partes de los CDs. Estas etiquetas delimitan a cada CD de forma global, así como los elementos que los constituyen. En primera instancia, se configuró el encabezado del documento XML, que se muestra a continuación:

- Fuente. Indica la fuente original del documento (CLI, CTIULA, CIE, *Google*, SkE).
- Fecha. Indica la fecha del recopilado y del etiquetado del documento.
- Nombre. Contiene el nombre de la recopilación hallada en el documento, como puede ser "verbo definir".
- Verbo. Muestra el nombre del verbo definitorio que se analiza.
- Tipo. Se indica si el criterio de clasificación del documento es la *definición*. Estas pueden ser: analítica, funcional, extensional o sinonímica.
- Recopilador. Muestra el nombre de la persona que recopiló el documento.

El cuerpo del documento contiene los CDs etiquetados. Las etiquetas utilizadas se pueden apreciar en el siguiente cuadro.

- CD. Contexto Definitorio: Indica los elementos que constituyen al CD, dentro de ellos se encuentra el término, su definición, la predicación verbal y las relaciones de correferencia.
- TERM. Término: En su atributos se marca se trata de un término lingüístico o de uno no lingüístico (cifras, símbolos). Se toman en cuenta tres tipos de frase: *fn* (frase nominal, *fn Y fprep* (frase nominal seguida de frase prepositiva) y *fv Y fn* (frase verbal seguida de frase nominal).
- DEF. Definición: En ella se debe omitir cualquier texto complementario que de manera estricta no forme parte de dicha definición. Existen cinco tipos: *GD* (Género próximo/Diferencia específica), *FUN* (Funcional), *EXT* (Meronímica/Extensional), *Ges* (Género exclusivo) y *Sin* (Sinonímica que se marcan en los atributos.
- PVD. Patrón Verbal Definitoria: Contiene todos los componentes de un PVD, incluyendo el clítico *se*, el verbo auxiliar, el verbo definitorio y el nexo.
- VD. Verbo Definitorio: Cuenta con los atributos *lema*, *args* (marca los argumentos del verbo); *mod* (indica el modo verbal: infinitivo *inf*, gerundio *ger*, participio *part*, formas finitas o verbo conjugado *fin*).
- Semarc. Clítico *Se.* Indica su posición respecto al verbo. El atributo distingue entre *enclítico (enc)* cuando *se* es parte de la morfología verbal y está en posición final, y *proclítico (prec)* cuando el clítico está en posición preverbal.
- Vaux. Verbo Auxiliar. Contiene cualquier verbo auxiliar dentro de la PVD (p.e., se puede considerar como, se ha definido, se debe concebir como…).
- NX. Nexo: Señala la función que cumple un adverbio o preposición entre el verbo y la definición.
- MRD. Marcadores Reformulativos Definitorios: Abarcan estructuras sintácticas con la función de explicar el propio lenguaje, p.e.: es decir, por ejemplo, esto es, etc.
- MTD. Marcadores Tipográficos Definitorios: Señala cualquier signo de puntuación o marcadores tipográficos definitorios (MTD). Se distingue en dos tipos: 1) marcadores definitorios (mdef): unen a un término con su definición, sustituyendo o complementando la función de la PVD. En los atributos se señalan como mdef= dp, viñ, par, gui, cll. 2) marcadores tipográficos (mt): indicación de negritas, cursivas, subrayado y otras marcas que dan prominencia al término definido o a la definición, este caso se marca mt= neg,curs,subr,otr.
- PP. Patrones Pragmáticos: Dan información sobre el uso de los términos. Los tres patrones considerados en este rubro son: Autoría (Aut), instruccionales (Inst) y temporales (Temp).
- Cf. Correferencia: Contiene las relaciones de referencia que se dan dentro del CD. En los atributos se marca si la *Cf* se da con el término (TERM) o con cualquier otro elemento del CD que opere como referente (ORef). Se especifica si la *Cf* es una *frase nominal* (fn), *frase nominal con demostrativo* (frdem), o tiene otra estructura (otr). A partir de números se marca el *índice* de la Cf (idcf) que permite ligarla con su referente (REF).

[2] http://www.iling.unam.mx:8080/CorcodeAppV/

- Anf. Anáfora: Marca las anáforas dentro del CD. En los atributos se marca si la *Anf* se da con el término (TERM) o con cualquier otro elemento del CD que opere como referente (ORef); se especifica también el tipo de anáfora o tipo de pronombre. Igual que en el caso anterior, el *índice* para ligar con su referente, es señalado con números.
- REF. Referente: Contiene al referente (REF) o antecedente de las correferencias y a las anáforas presentes en el CD. En los atributos se señala como índice (indcf/indanf), si el término definido (TERM) es el referente o es cualquier otra entidad (ORef) del CD.

La estructura queda ilustrada jerárquicamente en la figura 4.

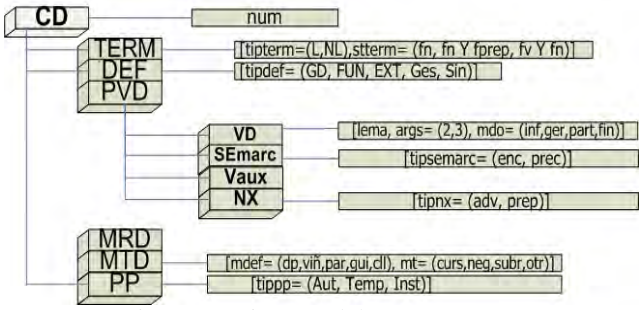

Figura 4: Etiquetas del CORCODE

11. Aplicaciones

Como hemos visto, las aplicaciones del empleo de la metodología aquí descrita para extraer CDs de textos de especialidad a partir de patrones verbales son diversas. En el grupo de Ingeniería Lingüística hemos trabajado en tres principales, las cuales describo a continuación.

11.1 Bancos de conocimiento

Como mencioné en la introducción, un aspecto relevante dentro de las investigaciones realizadas por el Grupo de Ingeniería Lingüística es el desarrollo de *bases de conocimientos léxico* (BCL) para diccionarios onomasiológicos electrónicos, las cuales incorporan de manera pertinente información lingüística, codificada en un nivel léxico, que ayuda a mejorar las consultas que hacen los usuarios. De manera general, las BCL son sistemas de bases de datos que almacenan, administran y proporcionan conocimientos obtenidos del lenguaje natural, a partir de textos tales como diccionarios, glosarios, artículos, etc. [63, 82].

El *diccionario onomasiológico* constituye un recurso léxico que permite a un usuario localizar la palabra adecuada para designar una idea que tiene en mente respecto a alguna cosa. En concreto, la intención de este diccionario es que a partir de conceptos o descripciones elaboradas por un usuario en lenguaje natural, el diccionario proporcione términos relacionados con dichas descripciones, en particular dentro de un dominio técnico o de especialización [77].

Un fenómeno que se ha observado a partir experimentos en torno a los modos de consulta onomasiológica en diccionarios electrónicos, es el amplio rango de posibilidades que tiene un usuario para codificar un concepto en una definición. Como señalan Lara [48] y Sager [71], existen diferentes métodos ofrecidos por el lenguaje natural para estructurar un concepto, más allá de la vía *Genus* y *Differentia* de la definición analítica. Se puede considerar entonces que los usuarios generan *definiciones libres*, las cuales se asocian a un término en particular; se trata de un proceso por el cual una persona, a partir de una idea, deduce la palabra que sirve para designar algo y que, en algún momento, se halla "en la punta de la lengua".

Dado que el diccionario onomasiológico arroja términos a partir de la descripción de los conceptos proporcionados por el usuario, la BCL requiere un módulo primario de adquisición de datos que concentre y amalgame la información conceptual que el usuario busca relacionar con una palabra específica. Para esto, es necesario considerar, además de la información contenida en diccionarios y enciclopedias, la información definitoria dada por los documentos de especialidad.

Por esta razón, la extracción de CDs resulta esencial, pues con la metodología mostrada se puede obtener cuatro tipos de definiciones: analíticas, funcionales, extensionales y sinonímicas. Además, todas ellas desde el punto de vista del experto que normalmente va más allá de la opinión del lexicógrafo.

11.2 Extracción de relaciones léxicas

Las relaciones léxicas (RLs) son un tipo de relación producida a partir del significado que contiene una palabra [66, 79]. El contenido de significación puede configurar dos tipos de situaciones:

Por un lado, como indica Fillmore [40], el contenido léxico de una palabra puede proyectar un escenario en donde se sitúan varios elementos que cumplen determinadas funciones acordes con dicho escenario. Por ejemplo, ciertos verbos de acción como *correr* configuran un escenario donde se necesita un agente que realice la acción, con una locación donde se lleve a cabo tal acto, una trayectoria que señale la ruta a seguir, una temporalidad que indique cuándo se realizó, etc.

Por otro lado, para Cruse [33] una palabra puede fijar una serie de relaciones con otras palabras que tengan un significado cercano a ésta. Por ejemplo, en el significado de un verbo como *correr* pueden encontrarse conceptos relacionados:

jerárquicamente superiores (p.e., *correr* es un tipo de acción); con un significado similar (trotar, acelerar); o con un significado contrario (*caminar*).

En el caso concreto de los lenguajes especializados, las RLs pueden servir para representar el sistema de conceptos de un campo de conocimiento específico. Dicho sistema constituye una especie de mapa donde se establece el lugar y la situación específica de un término frente a los demás de su mismo campo de conocimiento.

El desarrollo de un sistema de conceptos contempla la necesidad de conocer el significado de los términos. En el caso específico de los CDs, como unidades textuales que ayudan a describir el significado de un término, se pueden considerar como un repertorio de relaciones léxicas. En los CDs se establece una relación específica entre el término y su definición a partir del tipo de verbo definitorio que los une. Tal es el caso de las relaciones sinonímicas que se pueden distinguir con patrones verbales definitorios como *también llamado* o *también conocido como*. En otras situaciones, los verbos pueden indicar relaciones léxicas de función o extensión. Por ejemplo, en CDs con patrones como *consiste de*, *consta de*, *formado por*, *constituido por*, denotan una relación de extensión respecto al término que se define [21].

Las RLs son fundamentales para elaborar ontologías, tesauros, terminologías y otros recursos lingüísticos similares. Contar con herramientas para la identificación automática de relaciones léxicas permitirá su implementación en sistemas de pregunta-respuesta, web semántica, minería de textos e interfaces inteligentes, por mencionar algunos ejemplos. Desarrollar métodos automáticos con esta idea en mente implica crear perfiles sofisticados para repositorios de textos, los que serán necesarios en la siguiente generación de herramientas para el descubrimiento de recursos textuales tanto en Internet como en colecciones enormes de textos.

Si bien para el inglés existen varios sistemas de RLs, para el español son contados o casi nulos, y en general se trata de adaptaciones del inglés. Ahora, contar con una metodología y una herramienta para extraer relaciones léxicas que tome en cuenta el comportamiento lingüístico real del español tiene un impacto científico de gran valor para terminólogos y lexicógrafos, a la vez que permite la creación de otros recursos computacionales para nuestra lengua.

Ahora bien, es posible plantear la extracción de RLs a partir del análisis de los patrones verbales que aparecen como elementos constitutivos en definiciones localizadas en textos especializados.

Un hecho observado a raíz de esta investigación es la existencia de una relación estrecha entre el tipo de definición y el verbo que aparece como núcleo de un patrón verbal definitorio (PVD), lo que permite postular una taxonomía de cuatro tipos de definiciones basada en el tipo de PVD que aparece en el CD:

- Analítica: aquella definición que presenta de forma explícita un género próximo y una diferencia específica, por ejemplo: *una computadora es una máquina que resuelve operaciones lógicas*, donde el género próximo al que pertenece computadora es *máquina*, y las diferencias específicas son *que resuelve operaciones lógicas*.
- Sinonímica: aquella definición que manifiesta exclusivamente un género próximo, el cual establece una relación de equivalencia o sinonimia, por ejemplo: *un ordenador se llama también computadora*.
- Extensional: aquella donde se muestra una relación meronímica que enumera las partes que conforman una entidad, por ejemplo: *una computadora se compone de software, hardware y periféricos*.
- Funcional: aquella definición que describe la función o el uso de una entidad particular, por ejemplo: *una computadora sirve para resolver problemas lógicos, matemáticos y estadísticos*.

Esta clase de patrones, así como el comportamiento que presentan cuando aparecen ligados a una clase de definición específica, ha dado pie a que diferentes autores [25, 81, 84] reconozcan en ellos distintos tipos de RLs. Siguiendo la propuesta de Cruse [33], aquí se plantea la posibilidad de reconocer en los tipos de definiciones arriba expuestos las siguientes relaciones:

- Hiponimia-Hiperonimia: Una entidad hiponímica se deriva de un hiperónimo o elemento superior, por ejemplo: *una autobiografía es un libro*.
- Sinonimia: Dos entidades que mantienen cierta equivalencia a nivel cognitivo, por ejemplo: *Una mujer policía es un policía femenino*.
- Antonimia: Dos entidades que tienen un significado opuesto, por ejemplo: alto/bajo, computadora/calculadora, encender/apagar, entre otras.
- Individuación: Aquellas entidades donde aparece un cambio de individuación. Existen dos tipos de individuación: a) cantidad/masa, es decir, una relación entre una porción o una pieza y una cierta sustancia o entidad, por ejemplo: *Una hora es una porción de tiempo*; b) miembro/grupo, que es una relación entre una entidad que puede ser inherente a un grupo o colectivo, por ejemplo: *Un policía es un miembro de la fuerza policíaca*.

Así, se puede observar que con la clasificación de las relaciones posibles entre las definiciones, los patrones verbales asociados a definiciones y el agrupamiento automático, es posible la formulación de un algoritmo para la extracción automática de relaciones léxicas y, aun mejor, de definiciones.

11.3 El sistema Describe

Una aplicación directa del ECODE es el sistema denominado Describe® para la búsqueda, clasificación y agrupamiento de definiciones en la Web. La metodología parte de utilizar robots para indexar constantemente páginas que contengan alguno de los 2 millones de términos en el área de medicina. Estas páginas constituyen nuestra base de datos inicial para la extracción de contextos definitorios. Una vez extraídos los diferentes tipos de definiciones, éstos se clasifican según su tipo y se agrupan de acuerdo con el contenido semántico que en ellos se vincula.

Describe es una aplicación de arquitectura cliente-servidor orientada a Web, compuesta por varios módulos que permiten organizar la información disponible en Internet.

Del lado del servidor, el sistema está conformado por los módulos siguientes (Fig. 5):

- Extractor: Módulo encargado de extraer de Internet candidatos a CDs.
- Etiquetador: Permite etiquetar el texto de los candidatos a CDs proporcionados por el extractor.
- ECODE: Procesa el texto etiquetado e identifica los CDs finales, clasificándolos en los tres tipos de definición.
- Agrupamiento: Agrupa los CDs de acuerdo con sus características.

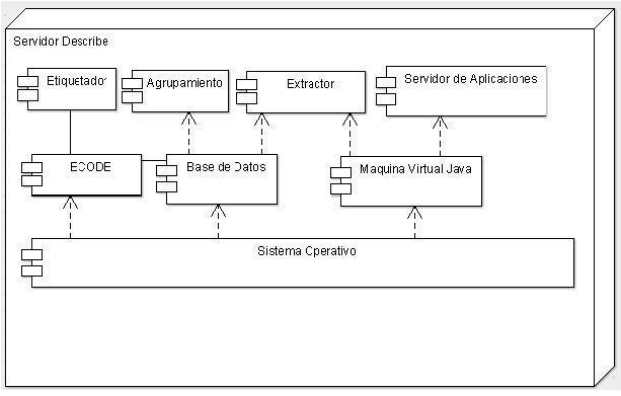

Figura 5: Diagrama del Describe

- Maquina Virtual de Java: Componente que permite ejecutar el Extractor de candidatos, independientemente de la plataforma o sistema operativo.
- Servidor de Aplicaciones: Permite al usuario interactuar con el sistema en ambiente Web.
- Sistema Operativo: Aplicación sobre la que se ejecutan todas las aplicaciones y módulos residentes en la máquina del servidor, permitiendo administrar y gestionar eficazmente sus recursos.

Dos Módulos vitales en el Describe son el ECODE y el de agrupamiento. Hemos visto en este artículo que el ECODE es un método satisfactorio para la extracción de definiciones en textos, además clasificadas en diferentes tipos. Este método, como se ha mostrado, sirve no sólo para el Describe, sino para otras aplicaciones, como la extracción de relaciones semánticas, elaboración de diccionarios semasiológicos y onomasiológicos, obtención de bases de conocimientos léxicas, etc.

El algoritmo de agrupamiento utilizado es un método novedoso que involucra una técnica adaptada de resúmenes automáticos y adecuada para fungir como medida de similitud. Este algoritmo, además de su uso para el Describe, será de utilidad para organizar los resultados (snippets) en motores de búsqueda.

El sistema Describe, de esta manera, apuesta a ser un buscador de definiciones con base en la web y será de gran utilidad tanto para especialistas como para individuos que deseen profundizar en el significado de un término especializado. Por ahora se trabaja en el área de medicina y se tiene contemplado ampliar el alcance de esta herramienta a otras áreas de conocimiento.

12. Agradecimientos

Esta investigación ha sido financiada por el Consejo Nacional de Ciencia y Tecnología, CONACYT, a través de los proyectos 46832 “Extracción de conceptos en textos de especialidad a través del reconocimiento de patrones lingüísticos y metalingüísticos”, 54616 “Análisis lingüístico de definiciones en contextos definitorios”, 82050 “Extracción de relaciones léxicas para dominios restringidos a partir de contextos definitorios en español” y de la beca doctoral CONACYT/Fundación Carolina 179210. Asimismo, bajo el patrocinio de DGAPA-UNAM, con el proyecto IN403108 “Extracción de relaciones semánticas a partir de definiciones en textos de especialidad”.

Un agradecimiento especial a los que en el marco de esta investigación realizaron estudios particulares y documentaron en su tesis, tanto a nivel de licenciatura como de maestría o doctorado: César Aguilar, Rodrigo Alarcón, Alberto Barrón, Valeria Benítez, Ariadna Hernández, Alejandro Molina y Octavio Sánchez. A Carme Bach que participó como codirectora de la tesis de doctorado de Rodrigo Alarcón. A los demás miembros del

Grupo de Ingeniería Lingüística que aportaron con su trabajo o en las discusiones: Edwin Aldana, Gabriel Castillo, Alfonso Medina, Víctor Mijangos y Carlos Rodríguez.

13. Referencias

13.1 Publicaciones del proyecto

[1] Aguilar, César. 2009. *Análisis lingüístico de definiciones en contextos definitorios*. Tesis de Doctorado, UNAM, México.

[2] Aguilar, César, Rodrigo Alarcón, Carlos Rodríguez y Gerardo Sierra. 2006. Reconocimiento y clasificación de patrones verbales definitorios en corpus especializados. En *La terminología en el siglo XXI: contribución a la cultura de la paz, la diversidad y la sostenibilidad*, editado por M. T. Cabré, R. Estopà, C. Tebé. Barcelona, IULA, Documenta Universitaria.

[3] Aguilar, César y Gerardo Sierra. 2008. Hacia una tipología de definiciones basada en el modelo analítico, *Memorias del XV Congreso Internacional ALFAL 2008*, Montevideo, Uruguay.

[4] Aguilar, César y Gerardo Sierra. 2009. Reconocimiento de definiciones asociadas a frases predicativas en contextos definitorios. *Procesamiento de Lenguaje Natural*, 43:151-158.

[5] Aguilar, César y Gerardo Sierra. 2009. A formal scope on the relations between definitions and verbal predications. *1st International Workshop on Definition Extraction*, Borovets, Bulgaria.

[6] Alarcón, Rodrigo. 2003. *Análisis de contextos definitorios en textos de especialidad*, Tesis de Licenciatura, UNAM, México.

[7] Alarcón, Rodrigo. 2009. *Extracción automática de contextos definitorios en corpus especializados*. Tesis de Doctorado, Universidad Pompeu Fabra, Barcelona.

[8] Alarcón, Rodrigo, Carme Bach C y Gerardo Sierra. 2008. Extracción de contextos definitorios en corpus especializados: Hacia una elaboración de una herramienta de ayuda terminográfica. *Revista Española de Lingüística* 37:247-278.

[9] Alarcón, Rodrigo y Gerardo Sierra. 2002. Hacia la extracción automática de conceptos. *Proc. VIII Simposio Iberoamericano de Terminología*. Red Iberoamericana de Terminología RITerm, Cartagena, Colombia.

[10] Alarcón, Rodrigo y Gerardo Sierra. 2003. El rol de las predicaciones verbales en la extracción automática de conceptos. *Estudios de Lingüística Aplicada*, 21(38):129-144.

[11] Alarcón, Rodrigo, Gerardo Sierra G y Carme Bach. 2008. ECODE: A Pattern Based Approach for Definitional Knowledge Extraction. *XIII EURALEX International Congress*, Barcelona.

[12] Alarcón, Rodrigo, Gerardo Sierra y Carme Bach. 2009. Description and Evaluation of Definition Extraction System for Spanish language. *1st International Workshop on Definition Extraction*, Borovets, Bulgaria.

[13] Benítez, Valeria. 2008. *Anáforas en la expansión de Contextos Definitorios: una propuesta de etiquetado.* Tesis de Licenciatura, UNAM, México.

[14] Barrón, Alberto. 2007. *Extracción automática de términos en contextos definitorios*. Tesis de Maestría, UNAM, México.

[15] Hernández, Ariadna. 2009. *Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática*. Tesis de Licenciatura, UNAM, México.

[16] Molina, Alejandro. 2009. *Agrupamiento automático de contextos definitorios*. Tesis de Maestría, UNAM, México.

[17] Sánchez, Octavio. 2009. *Análisis de relaciones léxicas en definiciones analíticas, extensionales y funcionales*. Tesis de Licenciatura, UNAM, México.

[18] Sierra, Gerardo y Rodrigo Alarcón. 2002. Identification of recurrent patterns to extract to definitory contexts. *Lecture notes in Computer Science* 2276:436-438.

[19] Sierra, Gerardo y Rodrigo Alarcón. 2003. The Role of Verbal Predications for Definitional Contexts Extraction. *TIA 2003*, Strasbourg: Université de Strasbourg.

[20] Sierra, Gerardo, Rodrigo Alarcón y César Aguilar. 2006. Extracción automática de contextos definitorios en textos especializados. *Procesamiento de Lenguaje Natural* 37:351-352.

[21] Sierra, Gerardo, Rodrigo Alarcón, César Aguilar y Carme Bach. 2008. Definitional verbal patterns for semantic relation extraction. *Terminology* 14(1):74-98.

[22] Sierra, Gerardo, Rodrigo Alarcón, Alfonso Medina, César Aguilar. 2004. Definitional contexts extraction from specialised texts. En *Practical Applications in Language and Computers*, editado por Barbara Lewandowska. Frankfurt: Peter Lang.

[23] Sierra, Gerardo, Gabriel Castillo, Antonio Reyes y Rodrigo Alarcón. 2001. Desarrollo de la Ingeniería Lingüística en la UNAM, México. *II Taller Internacional de Procesamiento Computacional del Español y Tecnologías del Lenguaje*. Jaén, España.

[24] Sierra, Gerardo, Alfonso Medina, Rodrigo Alarcón y César Aguilar. 2003. Towards the

extraction of conceptual information from corpora. *Proceedings of the Corpus Linguistics 2003 conference*, editado por Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. UCREL Technical Paper, No. 16, Lancaster University.

13.2 Bibliografía

[25] Alshawi, Hiyan. 1987. Processing Dictionary Definitions with Phrasal Pattern Hierarchies. *Computational Linguistics* 13(3-4):195-202.

[26] Auger, Alain. 1997. *Repérage des énoncé d'intér*êt *définitoire dans les bases de données textuelles*. Tesis de doctorado, Neuchâtel, Universidad de Neuchâtel.

[27] Bach, Carme. 2005. Los marcadores de reformulación como localizadores de zonas discursivas relevantes en el discurso especializado. *Debate Terminológico* 1.

[28] Cabré, Teresa. 1993. *La terminología. Teoría, metodología y aplicaciones*, Barcelona: Antártica.

[29] Cabré, Teresa. 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.

[30] Cabré, Teresa, Rosa Estopà y Jorge Vivaldi. 2001. Automatic term detection. A review of current systems. En *Recent Advances in Computational Terminology*, editado por Bourigault, D, Jaquemin, C, & L'Homme, M.C. Amsterdam: Benjamis.

[31] Calzolari, Nicoletta y Eugenio Picchi. 1988. Acquisition of Semantic Information from an On-Line Dictionary. *12th International Conference on Computational Linguistics*, *Coling'88*. Budapest.

[32] Cowie, Jim y Yorick Wilks. 2000. "Information extraction". En *Handbook of Natural Language Processing,* editado por R. Dale, H. Moisl and H. Somers. New York, Marcel Dekker.

[33] Cruse, D.A. 1986. *Lexical semantics.* Cambridge: Cambridge University Press.

[34] De Bessé, Bruno. 1991. Le Contexte Terminographique. *Meta* 26(1):111-120.

[35] Estopà, Rosa. 2001. Elementos lingüísticos de las unidades terminológicas para su extracción automática", en *La terminología científico-técnica*, editado por Cabré T, Feliu J., IULA-UPF, Barcelona.

[36] Estopà, Rosa, Jorge Vivaldi y Teresa Cabré. 1998. Sistemes d'extracció automática de candidats a terme. Estat de la qüestió. *Papers de l'IULA*, Série Informes, 22.

[37] Fajardo, Juan y Héctor Jiménez. 2003. Determinación de relaciones léxicas con base en el grado de subsunción. *Estudios de Lingüística Aplicada*, 22(38):81-87.

[38] Fernández, María del Carmen. 1999. *Las preposiciones en español. Valores y usos Construcciones Preposicionales*. Salamanca: Colegio de España.

[39] Fernández, Silvia, Eric San Juan y Juan Manuel Torres Moreno. 2008. Enertex: un sisteme basé sur l´energie textuelle. Traitement Automatique de la Langue Naturelle, Avignon.

[40] Fillmore, Charles. 1968. The case for case. En *Universals in Linguistic Theory,* Ediatod por Bach y Harms. New York: Holt, Rinehart and Winston.

[41] Haensh, Günther, Lothar Wolf, Stefan Ettinger y Reinhold Werner. 1982. *La lexicografía*, *de la lingüística teórica a la lexicografía práctica*. Madrid: Gredos.

[42] Hearst, Marti. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. *Proceedings of the 14th International Conference on Computational Linguistics, Coling'92*. Nantes.

[43] Heid, Ulrich, Susanne Jauss, Katja Krüger y Andrea Hohmann. 1996. Term Extraction with standard tools for corpus exploration". *4th International Congress on Terminology and Knowledge Engineering*, Viena.

[44] Jacquemin, Christian. 1996. A symbolic and surgical acquisition of terms through variation. En *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, editado por S. Wermter, E. Riloff y G. Scheler. Springer:Heidelberg.

[45] Jacquemin, Christian y Didier Bourigault. 2003. Term Extraction and Automatic Indexing. En *Handbook of Computational Linguistics*, editado por R. Mitkov, Oxford: Oxford University Press.

[46] Jurafsky, Daniel y James Martin. 2000. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Nueva Jersey: Upper Saddle River Prentice.

[47] Kilgarriff, Adam, Pavel Rychly, Pavel Smrz y David Tugwell. 2004. The Sketch Engine. *Proceedings of Euralex*, Lorient.

[48] Lara, Luis Fernando. 1997. *Teoría del diccionario monolingüe*, México: COLMEX.

[49] L'Homme, Marie-Claude. 2002. What can Verb and Adjectives tell us about Terms?. *Proc. Terminology and Knowledge Engineering, TKE 2002*. Nancy.

[50] L'Homme, Marie-Claude. 2005. Conception d'un dictionnaire fondamental de l'informatique

et de l'Internet : sélection des entrées, *Le langage et l'homme* 40(1):137-154.

[51] López, María López. 1972. *Problemas y métodos en el análisis de preposiciones*. Madrid: Gredos.

[52] Malaisé, Verónica. 2005. *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles á partir de corpus textuels*. Tesis de doctorado. Paris, Université Paris 7–Denis Diderot.

[53] Martín, María Antonia. 1999. Los marcadores del discurso. En *Gramática descriptiva de la lengua española,* editado por Bosque, I, Demonte, V. Madrid: Espasa.

[54] Medina, Alfonso, Gerardo Sierra, Gabriel Garduño, Carlos Méndez y Roberto Saldaña. 2004. CLI: An open Linguistic Corpus for Engineering. *Proc. Ibero-America Workshop on Artificial Intelligence,* Puebla, México.

[55] Meyer, Ingrid. 2001. Extracting a knowledge-rich contexts for terminography: A conceptual and methodological framework. En *Recent Advances in Computational Terminology,* editado por Bourigault, D.; Jaquemin, C. & L'Homme, M.C. Philadelphia: John Benjamins.

[56] Modrak, Deborah K.W. 2001. *Aristotle's Theory of Language and Meaning*, Cambridge: Cambridge University Press.

[57] Monachesi, Paola, Dan Cristea, Diane Evans, Alex Killing, Lothar Lemnitzer, Kiril Simov, Cristina Vertan. 2006. Integrating Language Technology and Semantic Web techniques in eLearning. Proc. ICL, Villach, Austria.

[58] Muresan, Smaranda y Klavans, Judith. 2002. A Method for Automatically Building and Evaluating Dictionary Resources. *Proc. 3th International Conference on Language Resources and Evaluation (LREC'02).* Las Palmas.

[59] Navigli, Roberto y Paola Velardi. 2007. GlossExtractor: A Web Application to Automatically Create a Domain Glossary. *Lecture Notes in Computer Science* 4733:339-349.

[60] Osinski, Stanis, Jerzy Stefano y Dawid Weiss. 2004. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. *Proc. Intelligent Information Systems*.

[61] Pavón, María Victoria. 1999. Clases de partículas: preposición conjunción y adverbio. En *Gramática descriptiva de la lengua española Vol 1. Sintaxis básica de las clases de palabras,* editado por Ignacio Bosque y Victoria Demonte. Madrid: Espasa.

[62] Pearson, Jennifer. 1998. *Terms in Context*, Philadelphia, John Benjamins.

[63] Pérez, Chantal. 2002. Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento, *Estudios de Lingüística Española* 18.

[64] Pinto, Ana Sofía y Oliveira, Débora. 2004. Extracção de Definições no Corpógrafo. Thecnical report. Faculdade de Letras da Universidade do Porto.

[65] Porter, Martin. 1980. An alghorithm for suffix stripping. *Readings in information retrieval,* San Francisco CA: Morgan Kaufmann Publisher Inc

[66] Pustejovsky, James. 1998. "Issues in text-based lexicon acquisition". *Corpus processing for lexical acquisition,* editado por B. Boguraev y J. Pustejovsky. Cambridge: The MIT Press.

[67] Pustejovsky, James, Sabine Bergler y Peter Anick. 1993. Lexical Semantic Techniques for Corpus Analysis. *Computational Linguistics* 19(2): 331-358.

[68] Rebeyrolle, Josette. 2000. *Forme et fonction de la définition en discours*, Tésis de doctorado, Université Toulouse-Le Mirail.

[69] Rebeyrolle, Josette y Ludovic Tanguy. 2000. Repérage automatique de structures linguistiques en corpus: le cas des énoncés définitoire. *Cahiers de Grammaire* 25:153-174.

[70] Rodríguez, Carlos. 2004. Metalinguistic Information Extraction from specialized texts to enrich computational lexicons. Tesis de Doctorado. Universitat Pompeu Fabra, Barcelona.

[71] Sager, Juan Carlos. 1990. *A Practical Course in Terminology Processing*, Philadelphia: John Benjamins.

[72] Sager, Juan Carlos. 2001. *Essays on Definitions*, Philadelphia: John Benjamins.

[73] Saggion, Horacio. 2004. Identifying Definitions in Text Collections for Question Answering. *Proc. 4th International Conference on Language Resources and Evaluation LREC2004,* Lisboa.

[74] Sánchez, A. y Melva Márquez. 2005. Hacia un sistema de extracción de definiciones en textos jurídicos. *Actas de la 1er Jornada Venezolana de Investigación en Lingüística e Informática*. Venezuela.

[75] Saurí, Roser. 1997. *Tractament Lexicogràfic dels Adjectius*, Sèries Monografies, IULA-UPF, Barcelona.

[76] Seiler, Bernhard y Wolfgang Wannenmacher. 1983. *Concept development and the development of the word meaning,* Berlin: Springer Verlag.

[77] Sierra, Gerardo y John McNaught. 2000. Design of an onomasiological search system: A concept-oriented tool for terminology. *Terminology,* 6(1): 1-34.

[78] Storrer, Angelika y Sandra Wellinghoff. 2006. Automated Detection and Annotation of Term Definitions in German Text Corpora. *Proc. 5th International Conference on Language Resources and Evaluation (LREC'06)*. Génova.

[79] Valero, Esperanza y Amparo Alcina. 2009. Linguistic realization of conceptual features in terminographic dictionary definitions. *Proc. 1st. International Workshop on Definition Extraction*. Borovets

[80] Vivaldi, Jorge. 1995. Proyectos del IULA: El corpus técnico, Simposio de Lingüística Hispánica. Instituto Cervantes y Universidad de Manchester, Manchester.

[81] Vossen, Piek y Ann Copestake. 1993. Untangling Definition Structure into Knowledge Representation. En *Inheritance, Defaults and the Lexicon.* Cambridge University Press.

[82] Walker, Donald y Robert Amsler. 1986. The Use of Machine-Readable Dictionaries in Sublanguage Analysis. En *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, Hillsdale: New Jersey.

[83] Wilks, Yorick. 1997. Information extraction as a core language technology. En *Information Extraction,* editado por M. T. Pazienza, Berlin: Springer.

[84] Wilks, Yorick, Brian Slator y Louise Guthrie 1996. *Electric Words. Dictionaries, Computers and Meaning*, MIT Press: Cambridge.