Revista Española de Lingüística (RSEL) 37, 2008, pp. 247-000
(ISSN xxxx-xxxx)

EXTRACCIÓN DE CONTEXTOS DEFINITORIOS HACIA LA ELABORACIÓN EN CORPUS ESPECIALIZADOS: DE UNA HERRAMIENTA DE AYUDA TERMINOGRÁFICA

Rodrigo Alarcón - Carme Bach - Gerardo Sierra
UPF - IULATERM, UPF - GIL, UNAM

RESUMEN

Uno de los objetivos principales del trabajo terminográfico es la identificación de conocimiento sobre los términos que aparecen en textos especializados. Para confeccionar diccionarios, glosarios u ontologías, los terminógrafos suelen buscar definiciones sobre los términos que pretenden definir. La búsqueda de definiciones se puede hacer a partir de corpus especializados, donde normalmente aparecen en contextos definitorios, es decir, en fragmentos de texto donde un autor explícitamente define el término en cuestión. Hoy en día hay un interés creciente por automatizar este proceso, basado en la búsqueda de patrones definitorios sobre corpus especializados anotados morfosintácticamente.

En este artículo presentamos una investigación centrada en la extracción automática de contextos definitorios. Presentamos una metodología que incluye tres procesos automáticos diferentes: la extracción de ocurrencias de patrones definitorios, el filtrado de contextos no relevantes, y la identificación de elementos constitutivos, es

ABSTRACT

One of the main goals of terminography work is the identification of knowledge about terms in specialised texts. In order to compile dictionaries, glossaries or ontologies, terminographers used to search for definitions about the terms that they are intent to define. The search for definitions can be done in specialised corpus, where they usually appear in definitional contexts, i.e. text fragments where an author explicitly defines a term. Nowadays there is a growing interest to automate this process, based on the searching for definitional patterns, and helped by morphosintactically annotated specialised corpus.

In this paper we present a research focused on the automatic extraction of definitional contexts. We present a methodology which includes three different automatic processes: the extraction of definitional pattern's occurrences, the filtering of non-relevant contexts, and the identification of constitutive elements, i.e. terms, definition and pragmatic patterns.

decir, términos, definiciones y patrones pragmáticos.

Palabras clave: terminografía, contexto definitorio, extracción de conocimiento, extracción de contextos definitorios.

Keywords: terminography, definitional context, knowledge extraction, definitional contexts extraction.

I. INTRODUCCIÓN

Un problema general de cualquier área de conocimiento es la organización y descripción de sus conceptos. La terminografía ocupa un lugar importante para la resolución de este problema, ya que se encarga, por un lado, de la elaboración de ontologías que representen la red conceptual de un área específica, y por otro lado de la elaboración de diccionarios donde se explique el significado de los términos.

Para la elaboración de diccionarios, el terminógrafo identifica en primer lugar los términos de un área especializada y en segundo lugar realiza un estudio de éstos para encontrar su significado.

El avance tecnológico en el desarrollo de herramientas que faciliten el trabajo terminográfico ha provisto al terminógrafo tanto de corpus lingüísticos especializados donde se almacena digitalmente una gran cantidad de documentos técnicos, como de sistemas para la extracción automática de términos.

Actualmente existe un creciente interés por el desarrollo de sistemas para la identificación automática de información sobre términos que sea útil para describir su significado. Diversos estudios coinciden en la idea de que en textos especializados, cuando se define un término, se suelen emplear ciertos patrones léxicos y metalingüísticos recurrentes, los cuales pueden ser reconocidos de manera automática (Pearson 1998, Meyer 2000).

Partiendo de esta idea, en este artículo se presenta una propuesta metodológica para la elaboración de un extractor de contextos definitorios (ECODE), junto con los primeros resultados obtenidos de aplicar dicha metodología sobre un corpus etiquetado morfosintácticamente. Este extractor está enfocado a la lengua española y tiene como principal campo de aplicación el ámbito terminográfico y el conocimiento especializado. Principalmente serviría para la elaboración de ontologías, es decir, bases

de datos de conocimiento léxico, glosarios o diccionarios especializados, tanto semasiológicos como onomasiológicos.

La metodología que presentamos para extraer contextos definitorios (CDs) en textos etiquetados morfosintácticamente podría extenderse, en primer lugar, a textos especializados no etiquetados, y en segundo lugar, a textos de lengua general no etiquetados, con lo que el ámbito se ampliaría hasta la búsqueda general del significado de unidades léxicas tanto especializadas como de lengua general.

En cuanto a la estructura de este artículo, se describirá en primer lugar nuestro objeto de estudio. En segundo lugar se hará una breve descripción de trabajos previos que han abordado el tema de la extracción automática de CDs. Por último se presentará la propuesta metodológica, así como las primeras aproximaciones y los resultados obtenidos hasta el momento.

II. CONTEXTOS DEFINITORIOS

En esta investigación se entenderá por «contexto definitorio» (CD) todo aquel fragmento textual de un documento especializado donde se define un término. Los CDs están formados por un término (T) y una definición (D), los cuales se encuentran conectados mediante un patrón definitorio (PD), por ejemplo verbos como *definir* o *entender*. Opcionalmente pueden incluir un patrón pragmático (PP), esto es, estructuras que aportan condiciones de uso del término o que matizan su significado, por ejemplo *en términos generales* o *en esta investigación*.

En la siguiente figura se puede observar una representación de los elementos de un CD, donde T y D junto con PD forman una unidad que puede estar modificada por el elemento optativo PP.

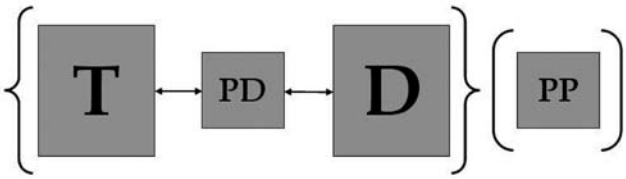

Figura 1. Estructura de un contexto definitorio.

Un ejemplo sería el siguiente, donde enmarcamos los elementos constitutivos dentro de los símbolos «<> </>»:

> <PP>De manera más formal,</PP> <T>la biología molecular</T> <PD>se ha definido como</PD> <D>una materia interdisciplinaria, que utiliza los métodos de la bioquímica, la genética y la química estructural para descubrir las bases moleculares de la forma, la función y el origen evolutivo de los seres vivos.</D>

En este contexto el término es «biología molecular»; la definición es todo lo que va desde «una materia interdisciplinaria [...]» hasta el final del enunciado; el patrón definitorio es «se ha definido como»; y el patrón pragmático es «de manera más formal», que en este caso se utiliza para indicar un matiz especial del significado del término. En seguida se describe cada elemento constitutivo de un CD.

2.1. *Término*

El término es la unidad sobre la cuál se aporta información relevante y puede tener estructuras sintácticas diferentes. El núcleo de un término generalmente será nominal, aunque no se debe descartar que en ocasiones pueda ser de otro tipo, como verbal o adjetival.

Siguiendo la clasificación propuesta por Estopà 2001, un término en este estudio corresponderá a una Unidad de Significación Especializada (USE). Estas unidades pueden ser lingüísticas y no lingüísticas. En el grupo de las USE lingüísticas encontramos USE léxicas y USE no léxicas. Las primeras pueden ser nominales, adjetivales, verbales o adverbiales, mientras que las segundas pueden ser unidades fraseológicas especializadas o combinaciones recurrentes. En el grupo de las USE no léxicas se encuentran los símbolos, nombres en latín y fórmulas.

Creemos que en el estudio de CD con fines de su extracción automática no se debe descartar la posibilidad de que los términos correspondan a la categoría de USEs no léxicas. Dependiendo del área especializada es común que en ocasiones lo que se defina esté más relacionado con fórmulas o elementos que, si bien no siguen patrones morfosintácticos comunes a los términos, sí representan una unidad de conocimiento especializada.

2.2. *Definición*

La definición en un CD corresponde a la información relevante que se aporta sobre un término y que puede ayudar para su comprensión. La definición es también una unidad especializada en tanto que provee el significado de un término especializado, por lo cual estás unidades están relacionadas con un área de conocimiento particular.

Meyer 2001 establece una tipología de distintos tipos de definiciones que se pueden encontrar en un CD. Partiendo de un modelo aristotélico,[1] propone que las definiciones presentes en contextos ricos en conocimiento (*Knowledge-rich Contexts* = KRCs) son de dos tipos distintos:

> A) KRCs definitorios *(Defining KRCs)*. Son el tipo más común y presentan la fórmula antes mencionada de una definición aristotélica: *Definición = Género próximo + Diferencia específica,* que en la fórmula de Meyer está dada como *X = Y + características distintivas*.
>
> B) KRCs explicativos *(Explanatory KRCs)*. Son aquellos donde sólo se proporciona información sobre el término, excluyendo la clase general a la cual pertenece. En la fórmula de Meyer se representa como X ⊃ *características*, donde el símbolo ⊃ significa que el elemento X debe tener, o por lo general tiene, una o varias características conceptuales.

El primer tipo se considera el más completo, ya que en él se detalla la clase general a la cual pertenece el término y además se detallan las características que lo distinguen de otros términos de su misma clase. En el segundo tipo la información sobre el término sólo permite crear una clasificación de éste a partir de la relación conceptual que establece con otros términos de su misma clase. Este tipo de definiciones sirven por lo general para encontrar relaciones conceptuales específicas como hiponimia, meronimia, sinonimia, por citar algunas.

---

[1] El cual sigue la fórmula: *X = genus + diferencia*, donde *X* es el término, *genus* es la categoría general a la cual pertenece dicho término, y *diferencia* es lo que distingue la categoría general del término que se define.

2.3. *Patrones definitorios*

En un CD los términos y las definiciones están ligados mediante un patrón definitorio. Los patrones definitorios pueden estar formados por elementos tipográficos o sintácticos, y ambos se utilizan para conectar el término con su definición.

Para este estudio se ha considerado que los patrones definitorios pueden ser patrones tipográficos definitorios (PTD), o bien patrones sintácticos definitorios (PSD), los cuales a su vez pueden ser patrones verbales definitorios (PVD) o marcadores reformulativos definitorios (MRD).

Los PTD cuando funcionan como conectores entre términos y definiciones son signos de puntuación (dos puntos, viñetas, guiones, etc.). Cuando se utilizan para resaltar la presencia de un término suelen ser marcas tipográficas o bien la propia tipografía del texto, por ejemplo el uso de comillas, subrayado, negrita, cursiva, etc.

Los PVD utilizan verbos metalingüísticos como *definir* o *denominar*, o bien verbos comunes al lenguaje general que pueden funcionar a nivel definitorio como *ser* o *conocer*. En un estudio previo (Alarcón 2003) se clasificaron los patrones verbales definitorios de acuerdo con su estructura en dos grupos: «simples» y «compuestos».

a) Los patrones verbales definitorios simples (PVDS) incluyen un verbo que se presenta de forma simple, sin ninguna otra partícula gramatical que los acompañe: X *significa* Y; Y *denominado* X, (donde X representa el término e Y la definición).

b) Los patrones verbales definitorios compuestos (PVDC) incluyen además del verbo ciertas partículas gramaticales, como adverbios, preposiciones o pronombres, y crean estructuras sintácticas compuestas: X *se define como* Y; X *sirve para* Y. Estas partículas las denominamos «nexos» (NX) y sirven para delimitar la estructura de un PVDC.

Los MRD, a grandes rasgos, son estructuras sintácticas que se encuentran relacionadas con un proceso también metalingüístico que en el caso de los CDs sirve para explicar el propio lenguaje, como señala Bach 2005, p. 2:

> La reformulación es un proceso de reinterpretación textual, mediante el cual un locutor determinado retoma algún elemento discursivo anterior para presentarlo de otra forma y con una función discursiva determinada.

En el grupo de marcadores reformulativos definitorios encontramos estructuras como *por ejemplo*, *es decir* y *esto es*.

2.4. *Patrones pragmáticos*

En los CDs se puede encontrar, además de la definición, otro tipo de información relevante para entender al término dentro del contexto en el cual aparece. Esta información está en relación con la introducción del término en el texto especializado, sus condiciones de uso, modificación y alcance (Rodríguez 1999). Este tipo de patrones se denominan «patrones pragmáticos» (PP) y pertenecen a un paradigma estructural amplio ya que su composición puede variar de acuerdo con formas estructurales o estilísticas utilizadas por cada autor. No obstante, encontramos patrones recurrentes, por ejemplo: adverbios y frases adverbiales *(usualmente*, *de manera general)*, frases prepositivas *(desde el punto de vista genético)*, o palabras simples *(definición*, *concepto*, *término)*.

III. ESTADO DE LA CUESTIÓN

El estudio de la extracción automática de CDs ha sido abordado desde una perspectiva teórico-descriptiva que ha dado paso al desarrollo de aplicaciones concretas para diferentes lenguas.

3.1. *Estudios teórico-descriptivos*

Uno de los estudios teórico-descriptivos más importantes es el trabajo de Pearson 1998, en el que se describe el comportamiento de los términos en el contexto real en el que aparecen y donde se menciona que, cuando un autor define un término, suele recurrir a patrones tipográficos para resaltar visualmente la presencia del término y/o la definición, y a patrones léxicos y metalingüísticos para conectar los dos elementos anteriores mediante estructuras sintácticas.

Esta última idea fue reforzada por el estudio de Meyer 2001, quien sostiene que en un texto especializado los patrones definitorios que conectan los términos con su definición pueden también introducir claves que permitan reconocer automáticamente el tipo de definición presente en los CDs, así como elaborar automáticamente una red conceptual.

En este sentido, y partiendo del estudio de los distintos verbos que pueden encontrarse en distintas relaciones conceptuales, en el estudio de Feliu 2004 se ha propuesto una tipología para la clasificación de dichos verbos y relaciones conceptuales con el fin de poder identificar relaciones conceptuales.

En el estudio de Bach 2005, referente a marcadores reformulativos, se ha propuesto una metodología que consiste en buscar automáticamente las ocurrencias de dichos marcadores en un corpus especializado, para, en conjunto con un sistema de identificación de términos, poder encontrar de manera semi-automática aquellos contextos donde se presente un proceso de reformulación textual útil para encontrar información definitoria.

El trabajo de Rodríguez 1999 detalla las «Operaciones Metalingüísticas Explícitas» (OMEs), que son operaciones comunicativas especializadas donde se puede localizar, entre otro tipo de informaciones, la definición del término o bien información sobre su origen o direcciones de uso.

Estos trabajos comparten la idea de buscar patrones recurrentes como punto de inicio en la búsqueda de información relevante sobre términos. Los patrones pueden englobarse en patrones tipográficos y patrones léxicos. Los primeros hacen referencia a la tipografía de un texto o a signos de puntuación, mientras que los segundos se refieren a verbos metalingüísticos, marcadores reformulativos o estructuras semántico-pragmáticas.

3.2. *Investigaciones aplicadas*

Existen investigaciones aplicadas que han partido de los estudios teórico-descriptivos para elaborar metodologías de extracción automática de CDs. Entre estas investigaciones se encuentran sistemas con distintas finalidades:[2]

[2] Es importante señalar que en principio, el desarrollo de estas aplicaciones ha sido enfocado a lengua inglesa, siendo reciente el intento de elaborar sistemas para otras lenguas, entre ellas la lengua española.

a) el reconocimiento automático de definiciones en textos médicos (Klavans y Muresan 2000), y en textos jurídicos (Sánchez y Márquez 2005);

b) la identificación automática de definiciones en sistemas de pregunta respuesta (Saggion 2004);

c) la extracción automática de información metalingüística para terminología (Rodríguez 2004);

d) la elaboración automática de ontologías (Malaisé 2005).

Las investigaciones aplicadas tienen como finalidad la extracción automática de información relevante sobre términos. Al igual que en los estudios teórico-descriptivos, la finalidad específica de cada autor es distinta aunque compartan ciertas ideas. La principal de ellas es que la búsqueda automática de las ocurrencias de patrones léxicos y metalingüísticos puede ser un buen punto de inicio para encontrar términos y definiciones.

Comparten también la idea de que en la búsqueda de patrones se obtendrá ruido (contextos donde no se aporta información relevante sobre un término) que podría ser filtrado automáticamente, y la idea de que una vez identificadas las ocurrencias donde posiblemente se presente información sobre un término, es necesario identificar cuál es dicho término y cuál es su definición.

En cuanto a la evaluación, todos toman como referencia los índices de precisión y cobertura *(precision and recall)* para comprobar que toda la información extraída automáticamente haya sido únicamente información relevante y que toda la información relevante haya sido extraída.

Cabe señalar que existen otros estudios que siguen por lo general las mismas líneas metodológicas que los anteriores. Alguno de estos trabajos son por ejemplo, una aplicación relacionada con el estudio teórico-descriptivo de Meyer, desarrollada por Davidson 1997; una investigación relacionada con la detección automática y la anotación de definiciones de términos especializados en corpus lingüísticos en alemán (Storrer y Wellinghoff 2006); o bien la propuesta y descripción de un primer acercamiento para la detección automática de relaciones conceptuales entre dos términos en textos especializados (Feliu y otros 2006).

IV. PROPUESTA METODOLÓGICA

Como se ha señalado anteriormente, la principal finalidad de un extractor de CDs sería facilitar la búsqueda de información relevante sobre términos, siendo la base de este extractor la búsqueda de ocurrencias de patrones definitorios. Un extractor que sólo obtuviera las ocurrencias de

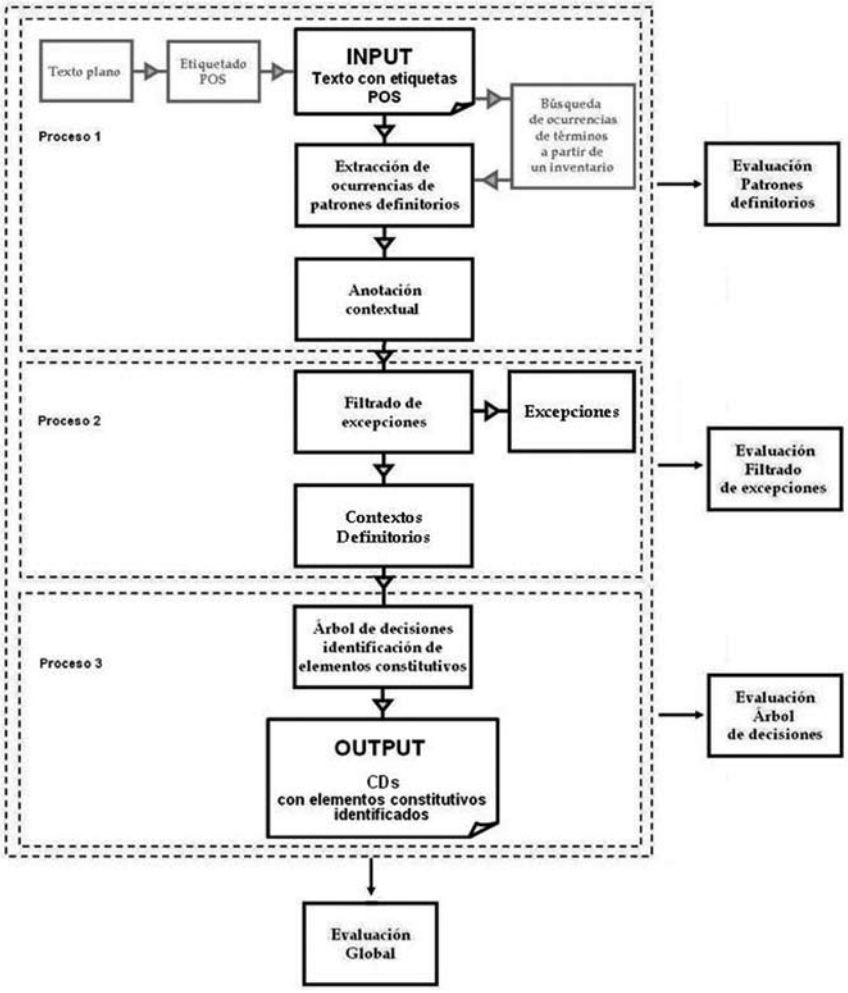

Figura 2. Esquema general del ECODE.

dichos patrones ya sería una buena herramienta de ayuda en las distintas tareas terminográficas. Sin embargo, el análisis manual de estas ocurrencias supondría todavía un esfuerzo que podría simplificarse mediante un extractor que incluyera un procesamiento automático de las ocurrencias. La metodología que aquí se propone incluye no sólo la extracción de ocurrencias de patrones definitorios, sino también el filtrado automático de excepciones (es decir, contextos no relevantes), así como la identificación automática de los elementos constitutivos de un CD. Esta metodología se representa mediante la figura número 2:

4.1. *Extracción automática de ocurrencias de patrones definitorios*

Para este trabajo se ha tomado como punto de partida el Corpus Técnico del IULA y su interfaz de búsqueda bwanaNet.[3] Este corpus está formado por documentos especializados en español, catalán, inglés, francés y alemán en las áreas del derecho, genoma, economía, medio ambiente, medicina, informática y lenguaje general. Además, el corpus está etiquetado con POS[4] mediante el estándar EAGLES[5] para representar los distintos tipos de palabra y sus características específicas.

Como parte de la aplicación de la metodología aquí propuesta, por ahora se han hecho pruebas con patrones verbales definitorios que incluyen los verbos: *concebir*, *definir*, *entender* e *identificar*. Seleccionamos estos verbos con la intención de representar la divergencia de enunciados que pueden recuperarse con verbos que tienen un matiz claramente más definitorio, como *concebir* o *definir*, frente a enunciados que se pueden emplear en una gran variedad de enunciados distintos,[6] como los recuperados con *entender* o *identificar*.

En un estudio anterior (Alarcón y Sierra 2003) se encontró que estos verbos pueden constituir los siguientes patrones verbales definitorios, donde:

---

[3] http://bwananet.iula.upf.edu/bwananetla.es.htm.

[4] Partes de la Oración, por sus siglas en inglés (Part Of Speach).

[5] http://www.ilc.cnr.it/EAGLES96/home.html.

[6] Cabe aclarar que somos concientes de la gran diversidad de patrones verbales que pueden emplearse en CDs, tanto aquellos que incluyen verbos definitorios como aquellos de lengua general. Tal es el caso del verbo *ser*, cuyo carácter general presupone de antemano la recuperación de una mayor cantidad de ruido. Se tiene contemplado trabajar con este y otros verbos, al igual que con patrones tipográficos y marcadores reformulativos.

SE = Pronombre impersonal *se*
VAux = Verbo auxiliar
VDef_Inf = Verbo definitorio forma impersonal infinitivo
VDef_Par = Verbo definitorio forma impersonal participio
VDef_Con = Verbo definitorio forma personal vonjugado
Pron = Pronombre
NX = Nexo
.* = Cualquier palabra o conjunto de palabras

Tabla 1. Patrones verbales definitorios

Formas impersonales en infinitivo
SE (Pron) VAux VDef_Inf | VAux VDef_Inf (SE | Pron) | VDef_Inf (Pron) .* NX
Ejemplo: puede definir (se | lo) .* como
Formas impersonales en participio
(SE VAux | Vaux{1,2}) Vdef_Par + NX
Ejemplos: se ha definido .* como
Formas personales conjugadas
(SE) VDef_Con + NX
Ejemplos: se define .* como

En la tabla anterior los verbos auxiliares (VAux) pueden ser formas personales o impersonales de cualquier verbo y los elementos entre paréntesis son optativos. Considerando que entre el verbo definitorio y el nexo podrían aparecer tanto términos (Ts) como patrones pragmáticos (PPs), utilizamos el símbolo «*» para representar una distancia *n* de palabras posibles.

Para el caso de los verbos que aquí se tratan, estos patrones (PPs) se han buscado mediante la opción de búsqueda compleja de bwanaNet, y se ha delimitado la distancia entre el lema definitorio y el nexo *como* a 15 palabras. La ecuación de búsqueda general ha sido la siguiente:

[lemma=”concebir|definir|entender|identificar” & pos=”V[^G]....|H.*”] [word!= “como”] {0, 15} [word=”como”]

Con esta ecuación se obtienen ocurrencias con patrones como *concebido como*; *se ha concebido como*; *fue concebida al principio como,* etc.

Una vez obtenidas las ocurrencias de cada patrón verbal definitorio, éstas pasan por un proceso de preparación que tiene la finalidad de simplificar su procesamiento automático. Este proceso es una simple anotación automática de cada ocurrencia con unas etiquetas que se han denominado «etiquetas contextuales», las cuales parten del patrón definitorio y anotan dentro de una etiqueta todas las palabras que están a la izquierda de dicho patrón, y dentro de otra etiqueta todas las palabras que aparecen a su derecha. La anotación tiene como finalidad establecer fronteras que ayuden en el proceso de identificar automáticamente las diversas posiciones que pueden ocupar los términos, las definiciones y los patrones pragmáticos en un CD.

Para ello se ha desarrollado una secuencia de comandos *(script)* en Perl[7] que asigna las siguientes etiquetas al patrón verbal definitorio, dependiendo de si la forma verbal es impersonal y se encuentra en infinitivo o participio, o bien si es una forma personal conjugada:

<pvd-inf> </pvd-inf>	Forma impersonal en infinitivo
<pvd-par> </pvd-par>	Forma impersonal en participio
<pvd-con> </pvd-con>	Forma personal conjugada

Asimismo, todo lo que aparece a la izquierda del patrón definitorio se anota con: «<izq></izq>», y todo lo que aparece a la derecha del patrón definitorio es anotado con: «<der></der>». En el caso de que haya un nexo, como el adverbio *como* en ciertos PVD, se anota con «<nexo></nexo>» todo lo que aparece entre el verbo definitorio y dicho nexo. Un ejemplo es el siguiente:

> <izq>El metabolismo</izq> <pvd-inf>puede definirse</pvd-inf> <nexo>en términos generales como</nexo> <der>la suma de todos los procesos químicos (y físicos) implicados:</der>.

[7] Se escogió este lenguaje de programación por su capacidad para integrar la búsqueda de expresiones regulares, las cuales son una base fundamental en todos los procesos de la metodología aquí propuesta.

4.2. *Filtrado de contextos no relevantes*

Una vez extraídas y anotadas las ocurrencias de patrones verbales definitorios, el siguiente proceso es el filtrado automático de contextos donde probablemente no se define un término. Como se ha señalado anteriormente, los patrones definitorios no se emplean únicamente en enunciados donde se aporta información relevante sobre términos. En el caso de los PVDs, algunos de los verbos tienden a tener un nivel metalingüístico mucho más alto que otros, por ejemplo *definir* o *denominar* frente a *concebir* o *identificar*. A su vez, los mismos verbos con un nivel metalingüístico alto no se utilizan siempre en enunciados en la definición de un término.

En un trabajo previo (Alarcón 2006) se realizó un análisis manual para determinar qué tipo de partículas gramaticales o secuencias sintácticas podrían encontrarse recurrentemente en los casos en que un patrón verbal definitorio no funcionara como tal. Con las partículas y secuencias encontradas se elaboraron reglas para filtrar, de los contextos obtenidos mediante la búsqueda de PVDs, aquellos contextos no relevantes.

Las posiciones en que pueden aparecer las partículas o secuencias son 3: antes del patrón verbal definitorio: __PVD; entre dicho patrón y un nexo: PVD__NEXO, o bien después del nexo: NEXO__. Para cada posición hay una frontera de inicio o límite representada mediante una etiqueta contextual.

Tabla 2. Patrones regulares utilizados para el filtrado de excepciones

Posición	*Partícula o secuencia*	
_PVD	1	no | en ningún caso | tampoco </izq>
	2	para </izq>
PVD_NEXO	3	<nexo> verbo conjugado
	4	no nexo </nexo>
	5	[así | ya] nexo </nexo>
	6	[Tan | tanto] .* nexo </nexo>
	7	[más | poco | poco más] nexo </nexo>
	8	[gerundio | que (signo)] nexo </nexo>
	9	«,» nexo </nexo>
	10	verbo personal conjugado nexo </nexo>

(continúa)

(continuación)

Posición	*Partícula o secuencia*	
NEXO_	11	<der> no
	12	<der> [antes | cuan | para | si]
	13	<der> (se) verbo personal conjugado
	14	<der> adjetivo verbo
	15	<der> adjetivo signo

Para implementar estas reglas se desarrolló otro script, el cual parte de la identificación de las partículas o secuencias en una posición determinada respecto a la frontera de cada regla. Este script esta basado no sólo en el reconocimiento de una palabra específica en una posición determinada, sino también en la búsqueda de secuencias sintácticas mediante la ayuda de las etiquetas POS. Por ejemplo, en el caso número 14, donde la regla incluye un *adjetivo* seguido de un *verbo* en la primera ocurrencia de la posición derecha. Algunos ejemplos clasificados como contextos no relevantes a partir de ciertas regularidades son los siguientes:

> Regla 3:
> <izq>Ciertamente esta observación tiene una mayor fuerza cuando el número de categorías </izq> <pvd-par>definidas</pvd-par> <nexo>es pequeño, como </nexo> <der>en nuestro análisis .</der>
>
> Regla 14:
> <izq>Ahora,</izq> <pvd-con>entiendo</pvd-con> <nexo>que como</nexo> <der>profesionales debemos dar una imagen, pero si utilizamos un término y luego el contenido no corresponde a [...]</der>

4.3. *Identificación automática de elementos constitutivos*

Una vez realizado el filtrado de excepciones, el siguiente proceso de la metodología que se propone es identificar automáticamente cuál es el término, cuál es la definición, y cuál es el patrón pragmático, en el caso de que lo haya, en las ocurrencias extraídas con PVDs.

Es necesario aclarar que, dependiendo del patrón definitorio, los términos y las definiciones pueden ocupar un lugar específico en los CDs.

Por ejemplo, los patrones definitorios tipográficos generalmente presentarían el término en la posición izquierda y la definición en la posición derecha: *T : D*, mientras que los patrones verbales definitorios presentarían otras posiciones recurrentes para T y D: *T se define como D* o *D es denominado T*.

Las distintas posibilidades respecto a las posiciones en las que pueden aparecer los elementos constitutivos dependiendo del patrón definitorio se han denominado «patrones contextuales». En el caso de los PVDs y dependiendo del verbo que se utilice para conectar al término con su definición, el número de distintas posiciones aumenta considerablemente, como en el caso de los PVDC que siguen el patrón *se define como*. En estos casos, T y D pueden aparecer a izquierda o derecha, además T puede aparecer entre el patrón verbal definitorio y el nexo, en el caso de que lo haya. A su vez, los elementos pragmáticos añaden un mayor número de combinaciones posibles.

Para identificar los elementos constitutivos se ha desarrollado un último script a partir de los patrones y etiquetas contextuales. Al igual que en el filtrado de excepciones, las etiquetas contextuales referentes a las posiciones de izquierda, nexo y derecha (<izq>, <nx> y <der>) se utilizaron como fronteras para delimitar las instrucciones del proceso automático de identificación. Además se han establecido expresiones regulares para representar las estructuras sintácticas de los elementos constitutivos.

La lista de las expresiones regulares que se han utilizado para representar un término,[8] una definición y un patrón pragmático son las siguientes:

Término:	FRON (Det.) + N + Adj. {0,2} .* FRON
Patrón pragmático:	FRON (signo) (Prep | Adv) .* (signo) FRON
Definición:	FRON Det. + N .* FRON

Donde:

Det.	= determinante
N	= nombre
Adj.	= adjetivo

[8] Por ahora se ha considerado que los términos sean únicamente unidades de carácter nominal, pero en un trabajo futuro se tiene contemplada la inclusión de expresiones regulares para representar unidades de carácter verbal en forma impersonal.

Prep. = preposición
Adv. = adverbio
FRON = frontera
.* = cualquier palabra o conjunto de palabras

En esta etapa, el procesamiento automático está fuertemente relacionado con la toma de decisiones para determinar las distintas posiciones en que pueden aparecer los elementos constitutivos en los candidatos a CDs. Para resolver este problema se desarrolló un árbol de decisiones que determina mediante inferencias lógicas las distintas posibilidades de aparición de los términos, definiciones y patrones pragmáticos.

Según Moreno y otros 1994, p. 49: «un árbol de decisión es una representación posible de los procesos de decisión involucrados en tareas inductivas de clasificación». Los árboles de decisiones son funciones de clasificación que están estructuradas como un árbol: tienen *nodos*, *ramas*, y *hojas*. Los nodos son decisiones tomadas a partir de atributos representados por las ramas y las hojas son elementos clasificados.

En el árbol desarrollado, las ramas en un primer nivel son las posiciones en las que pueden aparecer los elementos constitutivos, es decir izquierda, derecha y opcionalmente nexo; en un segundo nivel son las expresiones regulares para identificar cada elemento constitutivo. Los nodos corresponden a las decisiones tomadas a partir de los atributos de cada rama y están relacionados entre sí a nivel horizontal por inferencias del tipo *IF* 'si', *IF NOT* 'si no', y a nivel vertical por inferencias del tipo *THEN* 'entonces'. Por último, las hojas son las distintas posiciones una vez asignadas a un elemento constitutivo. Este árbol se implementó mediante otro script también en Perl.

En seguida se presenta un ejemplo de las inferencias que sigue el árbol de decisiones para determinar a qué elemento o elementos constitutivos corresponde la información presente en la posición izquierda.

Observamos en primer lugar que las decisiones parten del reconocimiento de las expresiones regulares de término, patrón pragmático o definición. Específicamente, con las inferencias 1 y 2 se puede determinar que la posición izquierda equivale a un término, o a un término y un patrón pragmático, que se distingue por una frontera como un signo de puntuación, mientras que la posición derecha equivale a una definición. En cambio, con la tercera inferencia se puede determinar que la posición iz-

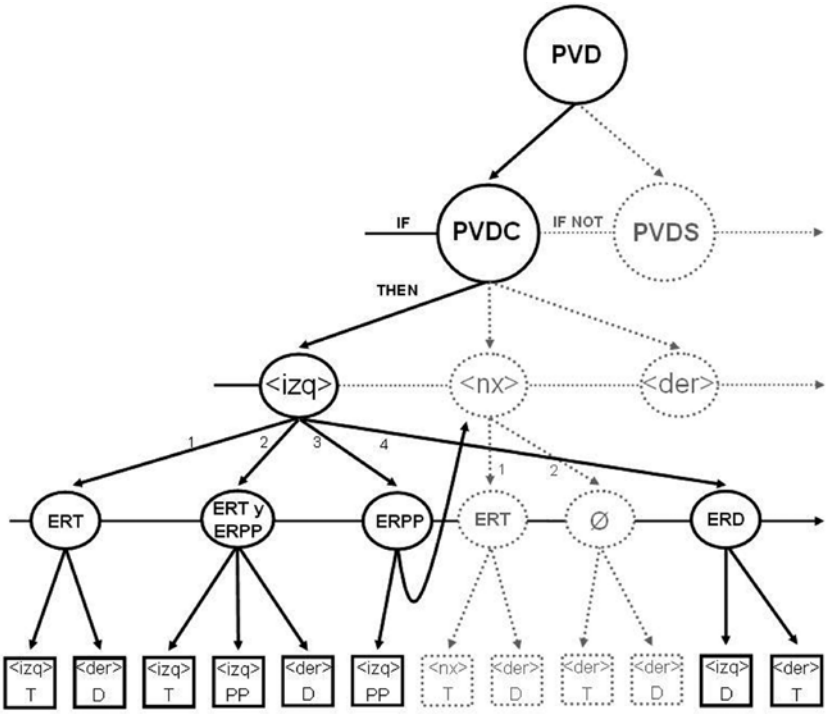

Figura 3. Árbol de decisiones para la posición izquierda.

quierda equivale únicamente a un patrón pragmático, y para saber en qué posición se encuentran el término y la definición se recurre a algunas inferencias de la posición de nexo. Así, si esta posición incluye únicamente una expresión regular de término, la posición nexo será el término y la posición derecha será la definición. Si la posición de nexo no incluye ninguna expresión regular correspondiente a un elemento constitutivo, entonces el término y la definición estarán en la posición derecha y podrán ser reconocidos a partir de una frontera como un signo de puntuación. Con la inferencia número 4 se encuentra a la definición en posición izquierda y al término en posición derecha. Esta última inferencia, al igual que la número 1, recurren a las inferencias de la posición de nexo para determinar si existe o no un patrón pragmático en esa posición.

Así, si tomamos el siguiente ejemplo:

<izq>En sus comienzos</izq> <pv-con>se definió</pv-con> <nexo>la psicología como </nexo> <der> «la descripción y la explicación de los estados de conciencia» (Ladd 1887).</der></s>.

Una vez identificado que el patrón verbal definitorio PVD corresponde a un PVDC (se definió como), se encontraría entonces que la posición izquierda:

1. NO está ocupada únicamente por una ERT
2. NO está ocupada por una ERT y una ERPP
3. SÍ está ocupada únicamente por una ERPP

Por lo tanto, la posición izquierda corresponde a un PP (en sus comienzos), y para identificar el término y la definición se recurre a las inferencias de la posición de nexo, con lo que se encuentra que en este caso:

1. *SÍ* está ocupado únicamente por una ERT.

Por lo que la posición de nexo corresponde a un término (la psicología) y la posición derecha corresponde a una definición («la descripción y la explicación de los estados de conciencia» [...]), quedando los elementos constitutivos anotados de la siguiente forma:

Término	=	psicología
Definición	=	«la descripción y la explicación de los estados de la conciencia» (Ladd 1887).
P. Verbal	=	se define como
P. Pragmática	=	En sus comienzos

V. EVALUACIÓN DE LOS RESULTADOS

Para evaluar los resultados obtenidos hasta ahora se utilizaron los índices de precisión y cobertura. En este estudio, dichos índices se entienden de la siguiente forma:

La precisión es una medida para determinar cuánta información, de la extraída automáticamente, corresponde a información «relevante». La cobertura es una medida para determinar cuánta información «relevante»

del INPUT se extrajo automáticamente. Los índices se determinan con las siguientes fórmulas:

$$\text{Precisión} = \frac{\text{\# Total de CDs válidos extraídos automáticamente}}{\text{\# Total de posibles CDs extraídos automáticamente}}$$

$$\text{Cobertura} = \frac{\text{\# Total de CDs válidos extraídos automáticamente}}{\text{\# Total de CDs en el INPUT}}$$

Para determinar el índice de precisión se debe saber cuántos CDs se extraen de forma automática, mientras que para determinar el índice de cobertura se debe conocer el número total de CDs en el INPUT de entrada, lo que supone un análisis manual previo. En los resultados de estos índices un número cercano al 1 indica que se han obtenido mejores resultados y por lo general suelen tener resultados inversos: si el número de precisión es alto, el número de cobertura será bajo y viceversa.

Como se ha visto en la figura 2, cada uno de los procesos consta de un sistema de evaluación propio, aparte de la evaluación global final. A continuación se expone el resultado obtenido para cada uno de los procesos llevados a cabo hasta ahora con los verbos definitorios con que se ha trabajado.

5.1. *Resultado y evaluación de la extracción de ocurrencias de patrones verbales definitorios*

Para adquirir una muestra representativa de ocurrencias de los patrones definitorios, se han tomado aleatoriamente 250 ocurrencias de cada patrón que incluyen ejemplos de todos los subdominios del Corpus Técnico del IULA. Los resultados de cada verbo se han analizado manualmente con la intención de encontrar contextos que realmente sean definitorios. En la siguiente tabla se presenta el número total de ocurrencias de cada verbo definitorio y el número total de CDs encontrados:

Tabla 3. Total de ocurrencias de los patrones verbales definitorios

Verbo definitorio	*Ocurrencias*	*CDs*
Concebir	120	74
Definir	250	192

(continúa)

(continuación)

Verbo definitorio	*Ocurrencias*	*CDs*
Entender	264	76
Identificar	250	59

Para evaluar la efectividad de los patrones buscados se utilizó el índice de precisión de manera aislada, ya que para utilizar el índice de cobertura se debería saber la cantidad total de CDs que se encuentran en el corpus de estudio. El índice de precisión corresponde en este caso al número total de CDs extraídos mediante la búsqueda de PVDs, sobre el total de ocurrencias recuperadas automáticamente. Así, para el índice de precisión se obtienen los siguientes resultados:

Tabla 4. Precisión de los patrones verbales definitorios

Verbo definitorio	*Precisión*
Concebir	0.6166
Definir	0.768
Entender	0.2878
Identificar	0.236

Se observa que los verbos que pueden funcionar en mayor medida como conectores entre un término y una definición, esto es, *concebir* y *definir*, recuperan efectivamente una mayor cantidad de CDs. Por su parte, los verbos *entender* e *identificar* sólo recuperan CDs en una cantidad inferior al 30 % del total recuperado, lo cual supone que recuperan una cantidad de ruido mayor.

5.2. *Resultado y evaluación del filtrado automático de excepciones*

En esta etapa se determina el índice de precisión dividiendo el número de CDs válidos extraídos automáticamente sobre el total de posibles CDs extraídos automáticamente. La cobertura se determina dividiendo el número total de CDs válidos extraídos automáticamente sobre el número total de CDs presentes en las ocurrencias extraídas automáticamente y

detectadas previamente de forma manual (durante el primer proceso). En estos casos los posibles CDs son las ocurrencias restantes una vez que se ha realizado el filtrado automático de contextos no definitorios.

Tabla 5. Resultados de Precisión y Cobertura

Verbo definitorio	*Precisión*	*Cobertura*
Concebir	0.7115	0.9866
Definir	0.8495	0.9896
Entender	0.3619	0.95
Identificar	0.3189	0.9076

Se observa que los índices de cobertura son superiores a 0.9, lo cual es un indicio de que algunos CDs se han filtrado como una excepción. Por su parte, se puede observar que los índices de precisión son buenos para los casos de *concebir* y *definir*, mientras que para los casos de *entender* e *identificar* dichos índices bajan notablemente.

Esto quiere decir que en el proceso de filtrar contextos no relevantes se filtran correctamente las excepciones, aunque algunas de ellas se escapan al script implementado, de forma que por el momento se deben detectar manualmente. De un total de 470 excepciones, se filtran automáticamente 146, lo cual indica que se puede detectar aproximadamente el 30 % de los contextos que no funcionan a un nivel definitorio.

Si se comparan los resultados iniciales de precisión (obtenidos mediante la evaluación de los PVDs) con los resultados obtenidos con esta misma medida una vez que se han aplicado las reglas de excepciones, se encuentran los siguientes resultados:

Tabla 6. Comparación entre precisión del proceso 1 y proceso 2

Verbo definitorio	*Precisión (proceso 1)*	*Precisión (proceso 2)*
Concebir	0.6166	0.7115
Definir	0.768	0.8495
Entender	0.2878	0.3619
Identificar	0.236	0.3189

Se observa que el índice de precisión mejora los resultados, aunque aún es necesaria una revisión y un refinamiento detallado de las reglas con la intención de filtrar más contextos que no funcionen como definitorios y para tratar de no filtrar CDs válidos.

5.3. *Resultado y evaluación de la identificación automática de elementos constitutivos.*

Con el script desarrollado para el proceso de identificación de los elementos constitutivos se pueden identificar correctamente contextos donde se presenta el término en la posición de nexo y algún patrón pragmático en la posición izquierda, por ejemplo:

Término = imitación.
Definición = el aprendizaje de un gesto a partir de la observación de su ejecución; sigue vigente ese significado en la actual investigación psicológica.
P. Verbal = definía como.
P. Pragmático = A principios de l siglo xx, Edward Thorndike.
Completo = <izq>A principios del siglo xx, Edward Thorndike</izq> <pvd-con>definía</pvd-con> <nexo>la imitación como</nexo> <der>el aprendizaje de un gesto a partir de la observación de su ejecución ; sigue vigente ese significado en la actual investigación psicológica.</der>.

Se pueden clasificar también contextos que incluyen un término en la posición izquierda y un patrón pragmático en la posición de nexo:

Término = metro.
Definición = la longitud de una determinada barra de platino iridiado mantenida en unas condiciones fijas.
P. Verbal = se definió como.
P. Pragmático = en 1889.
Completo = <izq>Por ejemplo , la unidad de longitud —el metro—</izq> <pvd-con>se definió</pvd-con> <nexo>en 1889 como</nexo> <der>la longitud de una determinada barra de platino iridiado mantenida en unas condiciones fijas.</der>.

O bien contextos que incluyen un término en la posición izquierda, nexo o derecha, por ejemplo:

Término = máquinas dedicadas.
Definición = ordenadores de terminal de trabajo *(Workstations*.
P. Verbal = están concebidas como.
Completo = <izq>Las máquinas dedicadas</izq> <pvd-par> están concebidas </pvd-par> <nexo>como</nexo> <der>ordenadores de terminal de trabajo *(Workstations)*.</der>.

Término = gen.
Definición = una unidad transcripcional, incluyendo sus regiones reguladoras asociadas.
P. Verbal = se entiende como.
Completo = <izq>Ya se ha hecho mención de que el propio concepto de gen ha ido cambiando a medida que ha progresado el conocimiento, pero en la mayoría de los casos</izq> <pvd-con>se entiende</pvd-con> <nexo>como</nexo> <der>gen una unidad transcripcional, incluyendo sus regiones reguladoras asociadas.</der>.

Por otro lado, los autores de textos especializados no suelen emplear constantemente un término sino que a veces utilizan referencias anafóricas para referirse a él. En el extractor que aquí se propone no se excluye la posibilidad de encontrar contextos donde se sustituye el término por una referencia anafórica. Sin embargo, por el momento no se tiene contemplada la búsqueda automática de su correferente, aunque no se descarta para un futuro.

Se han identificado ya algunos casos donde el término puede ser una posible referencia anafórica y puede venir señalado por un especificador demostrativo más una parte genérica del término, o bien por un pronombre personal:

R. Anafórica = Estos agentes.
Definición = carcinógenos en animales antes de que se descubriera su capacidad de transformar células en cultivos.
P. Verbal = fueron identificados como.

Completo = <izq>Estos agentes </izq> <pvd-par>fueron identificados</pvd-par> <nexo> como </nexo> <der> carcinógenos en animales antes de que se descubriera su capacidad de transformar células en cultivos.</der>.

R. Anafórica = lo.
Definición = un sistema de depuración del agua residual a través del terreno, con posibilidad de aprovechamiento agrícola o forestal del mismo.
P. Verbal = Podríamos definir como.
Completo = <izq>NULO</izq> <pvd-inf>Podríamos definir lo< /pvd-inf> <nexo> como </nexo> <der>un sistema de depuración del agua residual a través del terreno, con posibilidad de aprovechamiento agrícola o forestal del mismo .</der>.

Por otra parte, se ha observado también que en algunos casos donde se presenta la partícula de negación *no*, también puede presentarse después la partícula *sino*, con lo cuál se introduce, entre estas dos partículas, lo que denominamos una «contra-argumentación definitoria»[9] (CA-Def). Por ejemplo:

Término = redes de colectores.
CA-Def = meros receptores pasivos de la escorrentía urbana.
Definición = parte de un sistema que incluye elementos de control y cierta capacidad de almacenamiento, de manera que es posible la laminación de las avenidas y [...].
P. Verbal = se conciben como.
P. Pragmático = desde un punto de vista cuantitativo.
Completo = <izq>Asimismo, desde un punto de vista cuantitativo, las redes de colectores no</izq> <pvd-con>se conciben</pvd-con> <nexo>como</nexo> <der>meros receptores pasivos de la escorrentía urbana sino como parte de un sistema que incluye elementos de control y cierta capacidad de almacenamiento, de manera que es posible la laminación de las avenidas y [...]</der>.

[9] Se propone este término basándonos en los principios de la Teoría de la Argumentación de Ducrot y Anscombre 1983, 1995.

Por último, cabe señalar que todos los contextos que el script no puede identificar automáticamente se agrupan bajo la etiqueta «No Clasificable» (NC).

En esta etapa se utiliza el índice de precisión con el fin de determinar la exactitud del script en el momento de identificar automáticamente cuál es el término y cuál la definición del candidato a CD. Se ha limitado por ahora la evaluación al caso de estos dos elementos constitutivos, debido principalmente a la variedad de formas sintácticas que pueden representar los patrones pragmáticos en comparación con los términos y las definiciones.

La evaluación se realizó analizando manualmente los resultados y asignando un valor distinto a los candidatos a CDs en orden descendente, tomando en cuenta los siguientes criterios:

CD3 para los candidatos donde lo clasificado automáticamente como término y definición corresponde exactamente al término y la definición del CD, por ejemplo:

Término	=	turismo.
Definición	=	la reproducción de los hábitos cotidianos en un ambiente diferente.
P. Verbal	=	ha sido concebido como.
P. Pragmática	=	en términos generales.
Completo	=	<izq>El turismo en términos generales </izq> <pvd-par>ha sido concebido</pvd-par> <nexo>como</nexo> <der>la reproducción de los hábitos cotidianos en un ambiente diferente.</der>.

CD2 para los casos en que lo clasificado automáticamente en el CD como término y definición corresponde exactamente al término y la definición del CD, pero aparece otra información no relevante, por ejemplo:

Término	=	llamada proteinuria «postural», que.
Definición	=	proteinuria transitoria o invariable en posición erecta pero no recumbente, puede ocurrir sin que haya lesiones demostrables por estudio histológico de biopsias renales; el pronóstico a largo plazo en esos sujetos al parecer es excelente.
P. Verbal	=	se define como.

P. Pragmática = en sujetos jóvenes.
Completo = <izq>Más aún, en sujetos jóvenes, la llamada proteinuria «postural», que</izq> <pvd-con>se define</pvd-con> <nexo>como</nexo> <der>proteinuria transitoria o invariable en posición erecta pero no recumbente, puede ocurrir sin que haya lesiones demostrables por estudio histológico de biopsias renales; el pronóstico a largo plazo en esos sujetos al parecer es excelente. </der>.

CD1 si lo clasificado automáticamente no corresponde a un término o una definición, pero éstos se encuentran en otra posición del CD, por ejemplo:

Término = relación entre la cantidad de fármaco en el cuerpo y su concentración en plasma.
Definición = el «volumen aparente de distribución» (VD) del medicamento.
P. Verbal = se define como.
Completo = <izq>La relación entre la cantidad de fármaco en el cuerpo y su concentración en plasma</izq> <pvd-con>se define</pvd-con> <nexo>como </nexo> <der>el «volumen aparente de distribución»(VD) del medicamento .</der>.

Finalmente, en los casos en que la información en el candidato a CD no se puede clasificar automáticamente se asigna el valor cero: CD0.

El índice de precisión se determina dividiendo el total de CDs de cada grupo sobre el total de CDs encontrados automáticamente. En la siguiente tabla se observa el índice de precisión en la identificación automática de cada uno de los grupos. En estos casos representamos los valores de manera porcentual para dar una visión general de la cantidad de CDs que se clasifican para cada grupo en relación con el total de los CDs extraídos automáticamente.

Tabla 7. Precisión de la identificación automática de los elementos constitutivos

Verbo	*CD 3*	*CD 2*	*CD 1*	*CD 0*
Concebir	64,86 %	18,91 %	12,16 %	4,05 %
Definir	64,58 %	18,22 %	10,93 %	6,25 %
Entender	51,31 %	23,68 %	9,21 %	15,78 %
Identificar	47,45 %	5,08 %	38,98 %	8,47 %

Se observa que en la mayoría de los casos lo clasificado automáticamente corresponde exactamente con un término o una definición (CD 3). En este sentido todos los verbos presentan resultados semejantes, siendo *concebir* el que obtiene un porcentaje mayor.

En un porcentaje menor lo clasificado automáticamente incluye información extra o ruido (CD 2). Sin embargo, en estos casos la información presente en las distintas posiciones se clasifica correctamente.

También en un porcentaje menor, los términos y definiciones se clasifican en posiciones inversas (CD 1), exceptuando el verbo *identificar*, cuyo porcentaje es elevado en comparación con los demás verbos.

Sólo para el caso del verbo *entender*, en más del 10 % de las ocurrencias la información contenida en las posiciones de izquierda, nexo y derecha no se puede clasificar de forma automática (CD 0). En los demás verbos la información que no se puede clasificar es menor al 6.5 % del total de los CDs procesados automáticamente.

Lo anterior supone que deberá realizarse un estudio detallado para determinar porqué razón algunos candidatos a CDs no se clasificaron automáticamente y así poder incorporar nuevas inferencias en el árbol de decisiones que permitan su clasificación.

VI. CONCLUSIONES

Si bien la extracción automática de terminología es posible hoy en día gracias a los sistemas de extracción de terminología, dichos sistemas de extracción no permiten completar de forma automatizada el trabajo del terminólogo. La obtención de términos es útil para la confección de glo-

sarios especializados pero no es suficiente para la elaboración de diccionarios con definiciones.

El artículo que se ha presentado abre una nueva vía esperanzadora hacia la automatización del trabajo del terminólogo o lexicógrafo en la medida en que un extractor de contextos definitorios a partir de corpus especializados como el que aquí se presenta, facilitará para cada contexto obtenido un término, la definición que se le asocia, así como información pragmática de cada entrada, lo que podría ser útil explotar en futuras investigaciones (por ejemplo en la posibilidad de estudiar la evolución de la terminología a lo largo del tiempo).

De momento se ha expuesto una metodología con la que elaborar una herramienta para la búsqueda automática de contextos definitorios que se ha probado a partir de cuatro patrones verbales definitorios. Somos conscientes de que queda aún bastante trabajo por hacer, el cual incluye fundamentalmente:

a) explorar todos los tipos de patrones definitorios a partir de los cuales puedan extraerse automáticamente contextos definitorios;

b) incorporar la búsqueda y filtrado de estos patrones en los scripts de cada uno de los procesos expuestos;

c) mejorar el algoritmo para la identificación automática de los elementos constitutivos de cada contexto definitorio extraído;

d) realizar una evaluación de los resultados con el fin de obtener un panorama general del funcionamiento de la metodología propuesta.

VII. BIBLIOGRAFÍA

Alarcón, R. 2003: *Análisis lingüístico de contextos definitorios en textos de especialidad*, Tesis de licenciatura, México DF, Universidad Nacional Autónoma de México.

——2006: *Primeras aproximaciones a la extracción automática de contextos definitorios*, Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

——y Sierra, G. 2003: «El rol de las predicaciones verbales en la extracción automática de conceptos», *Estudios de Lingüística Aplicada* 38, México DF, Universidad Nacional Autónoma de México-Centro de Enseñanza en Lenguas Extranjeras, pp. 129-144.

Anscombre, J.C., y otros 1995: *Théorie des topoï*, París, Kimé.

Bach, C. 2005: «Los marcadores de reformulación como localizadores de zonas discursivas relevantes en el discurso especializado», *Debate Terminológico* 1, (Revista electrónica), Riterm. [http://www.riterm.net/revista/n_1/bach.pdf]

Davidson, L. 1997: *Knowledge extraction technology for terminology*, Tesis de maestría, Ottawa, University of Ottawa.

Danells, D. 2005: *Recognizing swedish* acronyms and their definitions in biomedical literature, Gotemburgo, Department of Swedish language, Göteborg University.

Ducrot, O., y Anscombre, J.C. 1983: *L'argumentation dans la* langue, Bruselas, Mardaga (trad. esp.: *La argumentación en la lengua*, Madrid, Gredos, 1995).

Estopá, R. 2001: «Elementos lingüísticos de las unidades terminológicas para su extracción automática», en. Cabré, M.T. y Feliu, J. (eds.), *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica*, Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, pp. 67-80.

Feliu, J. 2004: *Relaciones conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*, Tesis de doctorado, Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

——, Vivaldi, J. y Cabré, M. T. 2006: «SKELETON: Specialised knowledge retrieval on the basis of terms and conceptual relations», 5*th International Conference on Language Resources and Evaluation LREC2006*, Génova, European Language Resources, pp. 2377-2382.

Klavans, J. y Muresan, S. 2000: «Evaluation of the DEFINDER system for fully automatic glossary construction», American Medical Informatics Association Symposium, Washington, pp. 324-328.

Malaisé, V. 2005: *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles á partir de corpus textuels*, Tesis de doctorado, París, UFR de Linguistique, Universite Paris 7-Denis Diderot.

Meyer, I. 2001: «Extracting Knowledge-rich contexts for Terminography», en Bourigalt, D., Jacquemin, C. y L'Homme, M. C. (eds.), *Recent advances in computational terminology*, Ámsterdam, John Benjamins, pp. 278-302.

Moreno, R., Armengol, V., Béjar, A., Belanche, M., Cortés, U. Gavaldá, R., Gimeno, J., López, I., Martín, M., y Sánchez, M. 1994: *Aprendizaje automático*, Barcelona, Universidad Politécnica de Cataluña.

Pearson, J. 1998: *Terms in context*, Ámsterdam, John Benjamins.

Rodríguez, C. 1999: *Operaciones metalingüísticas explícitas en textos de especialidad*, Trabajo de investigación, Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

——, C. 2004: «Metalinguistic information extraction for terminology», 3rd *International Workshop on Computational Terminology (CompuTerm2004)*, Génova, Coling, http://arxiv.org/ftp/cs/papers/0504/0504074.pdf.

Saggion, H. 2004: «Identifying definitions in text collections for question answering», *4th International Conference on Language Resources and Evaluation LREC2004*, Lisboa, European Language Resources, pp. 1927-1930.

Sánchez, A., y Márquez, M. 2005: «Hacia un sistema de extracción de definiciones en textos jurídicos», *I Jornada Venezolana de Investigación en Lingüística e Informática*, Venezuela, [http://alexy.sanchez.tripod.com/Documentos/ExtraccionDefinicionesArticulo.pdf]

Sarmento, L., Maia, B., y Santos, D. 2004: «The Corpógrafo - a Web-based enviroment for corpora research». En *4th International Conference on Language Resources and Evaluation LREC2004*. Lisboa, European Language Resources. pp. 449-452.

Storrer, A., y Wellinghoff, S. 2006: «Automated detection and annotation of term definitions in german text corpora». En 5th *International Conference on Language Resources and Evaluation LREC2006*. Génova, European Language Resources, pp. 2373-2376.