Propuesta recibida el 1 de noviembre 2017 y aceptada para publicacin el 31 de diciembre 2017.

# Extracción automática de definiciones analíticas y relaciones semánticas de hiponimia-hiperonimia con un sistema basado en patrones lingüísticos

**Automatic extraction of analytical definitions and hyponymy-hypernymy relations with a pattern-based system**

M. Alejandro Dorantes Cruz
Universidad Nacional Autónoma de México
mdorantescr@iingen.unam.mx

Alejandro Pimentel Alarcón
Universidad Nacional Autónoma de México
apimentala@iingen.unam.mx

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

Gemma Bel-Enguix
Universidad Nacional Autónoma de México
gbele@iingen.unam.mx

Claudio Molina
Escuela Nacional de Antropología e Historia
claudio.molina.salinas@enah.edu.mx

**Resumen**

En el presente trabajo se muestra parte de un proyecto en curso centrada en el diseño de un autómata lexicográfico. El objetivo principal de la investigación es la extracción de definiciones analíticas y relaciones semánticas de términos con datos tomados directamente de internet. Presentamos dos de las capacidades del sistema: la extracción de definiciones analíticas y de hiperónimos. La metodología consiste principalmente en la búsqueda automática de esta información con patrones construidos manualmente basados en la estructura léxica de definiciones analíticas en lenguaje natural.

Con este desarrollo, ha sido posible mejorar la precisión reportada en el estado del arte. Se ha conseguido una precisión de 92.5 % para la tarea de extracción de definiciones analíticas y de las relaciones de hiperonimia.

**Palabras clave**

extracción automática de definiciones, hipónimia-hiperonimia, patrones lingüísticos

**Abstract**

This work is part of an ongoing project that is focused on the design of a lexicographic automaton. The main objective of the research is the extraction of analytical definitions and semantic relations of terms with data taken exclusively from internet. We present two of the abilities of the system: the extraction of a) analytical definitions and b) hypernyms. The methodology consists of the automatic search of that information with manually-built patterns based on the lexical structure of analytic definitions in natural language.

This method has improved the precision reported in the state of the art. Se have reached a precision of 92.5 % for the extraction of analytical definitions and for the hypernymy relations.

**Keywords**

automatic extraction of definitions, hypernymy-hyponymy, linguistic patterns

## 1 Introducción

La generación de definiciones lexicográficas es un área de poco o nulo desarrollo dentro del procesamiento del lenguaje natural (PLN). En cambio, el crecimiento exponencial de los nuevos conceptos en ciencia y la tecnología, junto con la especialización del conocimiento, hacen de los diccionarios elementos cruciales para todos aquellos profesionales que, en un momento dado, se alejan de su campo habitual de trabajo.

En algunos ámbitos, existe la idea de que internet hace innecesaria la existencia de diccionarios. En efecto, es de destacar los beneficios de tiene la web como corpus, ya que se encuentran gran número de datos provenientes de múltiples fuentes. Pero en realidad, aunque la web ofrece una gran cantidad de información, no es fácil encontrar herramientas que la estructuren, y la conviertan en conocimiento especializado.

DOI: 10.21814/lm.9.2.257

Al mismo tiempo, si bien existen otros sistemas como wordnet, o enciclopedias, en algunas ocasiones estos se encuentran desactualizados si se comparan con la información que día a día se registra en internet.

Por todo ello, encontrar formas automáticas para extraer y procesar desde la www los elementos constituyentes de una definición y conectarlos se ha convertido en un desafío para la lingüística computacional actual.

La lexicografía aún no ha conseguido un método para automatizar la creación de definiciones. Para lograr esta meta se requiere una estructura base modelada con patrones, nutrida con información suficiente y estadísticamente pertinente.

En este artículo se sostiene que es posible generar definiciones analíticas de forma automatizada partiendo de un conjunto de candidatos a definiciones, en su mayoría aceptables, y de la identificación de sus hiperónimos.

Existen distintas técnicas para extraer candidatos a definiciones (Pearson, 1998; Meyer, 2001; Alarcón, 2009) e hiperónimos (Hearst, 1992; Ortega, 2007). En ambos casos los sistemas existentes tienen un recall aceptable, pero su precisión es muy baja, o insuficiente para asegurar el éxito en la generación de definiciones. Por ello, esta propuesta se enfoca a mejorar la precisión, aunque el recall pudiera verse afectado.

El artículo tiene la siguiente estructura: en la Sección 2, se describe una tipología de las definiciones en lexicografía computacional, prestando particular atención al tipo de definición analítica, que es la que interesa en esta investigación; más tarde (Sección 3), se señalan los antecedentes en la literatura de la extracción de definiciones y la extracción de relaciones léxicas de hiperonimia-hiponimia, sea en trabajos teóricos previos o en investigaciones aplicadas puntuales; en la Sección 4 se describe la arquitectura y metodología seguida para la extracción de definiciones analíticas y relaciones semánticas de hiperonimia-hiponimia en estos mismos contextos; por último (Secciones 5 y 6), se ofrecen algunos resultados, se presenta una evaluación y se dan algunas conclusiones generales, así como algunas líneas futuras de investigación.

## 2 Definiciones en lexicología computacional

En este apartado se hace un repaso de cuáles son los elementos constitutivos de una definición y se introduce una tipología estas. Posteriormente, se presenta la noción de contexto definitorio (CD) y de patrón definitorio (PD).

### Tipología de las definiciones

Aristóteles (según señala Smith (2007)) indica que una definición consta de dos partes: el genus, conocido también como *kind* o *family*, mismo que indica qué tipo de cosa es el definendum (elemento que está siendo definido), y la *differentia*, que especifica o hace único al *definendum*. De estos dos elementos (*genus* y *differentia*) se puede asumir que el genus sea también un término más general o hiperónimo, mientras que la *differentia* es mayormente una predicación en la que se enumeran las características diferenciadoras o propias del término.

Con base en las implicaciones de la propuesta aristotélica relacionada con esta particularidad, se ofrece en la literatura al respecto (Sierra et al., 2008; Aguilar, 2009) una clasificación en cuatro tipos de definiciones: analíticas, sinonímicas, funcionales y extensionales.

Las *definiciones* analíticas son el tipo más prototípico considerando el modelo aristotélico que se ha descrito hasta ahora. Según lo explicado por Aguilar (2009), una definición de tipo analítico aporta un conocimiento inherente al término definido, aunque también suelen aportar características no esenciales o características adquiridas accidentalmente.

La *definición sinonímica* no ofrece diferencia específica, sino únicamente género próximo (Dorantes, 2016). Por ejemplo, para la entrada "Estado financiero" se propone la definición "Estado de situación financiera"; por su parte, "Impuesto" se define como "Gravamen, arancel".

Una *definición funcional* ofrece, por medio de la diferencia específica, una eplicación que aclara la función, utilidad o fin de lo referido por la entrada (Aguilar, 2009). Por ejemplo: "Estado financiero" se explica como algo que "Sirve para calcular la utilidad o pérdida neta que generará el proyecto hecho a los estados de resultados" y "Servicio financiero" es algo que "Sirve para mejorar la calidad de vida y el desarrollo de los hogares".

Las *definiciones extensionales*, por su parte, enumeran las partes o componentes que forman al término definido, por ejemplo: "Estado financiero" / "Se compone de la cuenta de resultados y el balance"; "Impuesto" / "Se compone del objeto, el sujeto, la base y la tasa o la tarifa".

Nuestro sistema se basa en las definiciones analíticas porque son el tipo más arquetípico dentro de lexicografía. Como afirma Lara (1997), "la mayor parte de la definición lexicográfica y enciclopédica contemporánea (...) se rige en mayor o menor grado por la teoría aristotélica".

Es importante recordar que una definición analítica proporciona dos tipos de conocimiento:

- *Genus*: este conocimiento indica a qué clase o grupo pertenece el término de entrada por medio de un término más general o hiperónimo.
- *Differentia*: este conocimiento especifica qué hace que el término de entrada sea único del resto de los elementos en su clase de pertenencia.

Las definiciones analíticas resultan muy interesantes porque "género y diferencia se convierten en condiciones necesarias y suficientes para (el) reconocimiento de todo objeto" (Lara, 1997, pg. 208).

### Contexto definitorio y patrón definitorio

Autores como Alarcón et al. (2008) señalan que un contexto definitorio (CD) es todo fragmento de tamaño indeterminado dentro de un documento en donde se describe clara y precisamente la definición de un término. Estos autores afirman que los CDs están formados por un término y una definición que se encuentran relacionados entre sí por sintagmas como "se define" o "se entiende como" entre otros, también conocidos como patrones definitorios (PDs).

Por un lado, el término es uno de los elementos constitutivos, no accesorio, del contexto definitorio y es el único elemento sobre el cual se introduce información relevante en el contexto (Alarcón et al., 2008); mientras que la definición es un elemento constitutivo del CD que contiene la información relevante que se aporta sobre el término. Esta definición constituye una explicación del término (Sager, 1993, pg. 67).

El sistema que presentamos incorpora el *nexus differentia* (ND) (Dorantes, 2016), una expansión de los patrones de CDs cuya ventaja es la división de una definición analítica en su término genérico (*genus*) y diferencia específica (*differentia*). Este elemento, según Dorantes (2016), es esencial para la estructura de las definiciones analíticas en español porque contiene una regularidad basada en la revisión de diccionarios, así como una gran cantidad de candidatos a definiciones. El autor señala que, aunque la *differentia* ya se había propuesto como un elemento de la definición analítica, no hay trabajos que muestren las características lingüísticas de la partícula o la forma en que dicha partícula introduce la diferencia específica.

Por último, conviene referir que un patrón definitorio (PD) es un elemento lingüístico que relaciona al término y a su definición, dándole a la predicación existente entre ellos una orientación de tipo sinonímica. En la literatura, se identifica que los PD en español podrían conformarse por verbos que, siguiendo a Rodríguez (1999), se denominan verbos metalingüísticos (definir, denominar, describir...), aunque autores como Alarcón (2006) señalan que es posible que verbos con una semántica distinta también funcionan como PD (ser, conocer o identificar) y cambien su comportamiento argumental a una predicación metalingüística.

Los PD, dependiendo del tipo de complemento que requieran, podrían determinar el tipo de orientación de la definición según la clasificación que ya se ha explicado anteriormente.

Un patrón verbal analítico establece una relación de predicación entre el término y su definición, en la que la segunda remite o describe características inherentes o adquiridas del primero. Según se desprende del estudio de Aguilar (2009), los verbos que orientan este tipo de definiciones son "referir", "representar", "significar" y "ser".

Todo lo anterior constituye un marco de referencia que contiene los elementos teóricos constituyentes de las definiciones analíticas. Sin embargo, es indispensable que inmediatamente se traten algunos aspectos complementarios relacionados con la extracción de definiciones y a la determinación de los *genus* de estas. Estos se desarrollan a continuación.

## 3 Trabajos previos

Desde la perspectiva que se ha planteado para la presente investigación, conviene referir algunos antecedentes tanto para la extracción de definiciones como para la extracción de relaciones de hiponimia-hiperonimia, mismas que a continuación se explican.

### Trabajos sobre extracción de definiciones

En el estudio de la extracción automática de definiciones se ha trabajado desde distintas perspectivas, por ejemplo, uno de los primeros estudios en el área fue el de Pearson (1998) en el que presenta el comportamiento de los contextos en los que aparece un término que se describe. Pearson afirma que cuando un autor define un término, comúnmente se utilizan patrones que llaman la atención hacia la presencia de un elemento importante sobre el que se está trabajando y dando una definición. El autor identifica, además, la aparición de patrones léxicos que conectan las definiciones con los términos.

Por otra parte, Meyer (2001) refuerza estas ideas y encuentra que los patrones definitorios también proveen elementos clave para la identificación del tipo de definiciones aplicadas a los términos. Así, la principal motivación para trabajar con contextos definitorios surge de la necesidad de obtener conocimiento semántico de los términos que aparecen dentro de diferentes áreas de especialidad.

Uno de los proyectos mexicanos que trabaja en la extracción de contextos definitorios es el corpus CORCODE en español (Sierra et al., 2006), de uso público a través de internet.[1]

Este enfoque supone la creación de metodologías para la extracción automática de definiciones, por ejemplo, Klavans & Muresan (2001) se enfocan específicamente en métodos para la extracción de términos y definiciones en textos médicos con su sistema Definder. Este sistema basado en reglas trabaja de manera excepcional; al enfocarse en la búsqueda de terminología muy especializada, se reporta una precisión del 87 %.

Por otra parte, Espinosa-Anke et al. (2016) desarrollan DefExt, una herramienta capaz de extraer definiciones a partir de un corpus utilizando un enfoque semi-supervisado. En esta misma línea, el sistema utiliza etiquetado de partes de la oración y un ordenamiento de importancia de los documentos de un corpus; los autores reportan una precisión general de 50 %.

Adicionalmente, han surgido también buscadores que trabajan utilizando el internet en lugar de un corpus estático: GlossExtractor, desarrollado por Velardi et al. (2008) recupera información de la Web, este último se enfoca en glosarios y documentos especializados dentro de internet, a partir de los cuales extrae las definiciones de un listado de términos predefinidos. Este sistema está basado en inglés, utilizan un enfoque de aprendizaje por computadora sobre diccionarios y etiquetado automático de partes de la oración; reportan una precisión del 73.5 % sobre las definiciones extraídas de internet.

Un trabajo importante que se ha desarrollado en México es el buscador ECODE (Alarcón et al., 2008), sistema capaz de extraer contextos definitorios a partir de búsquedas en la red. El enfoque que se utiliza en ECODE consiste en la separación de la tarea en dos módulos: el primero se encarga de las búsquedas en línea, se utilizan 15 patrones rodeando a un término en una búsqueda textual exacta y el segundo módulo filtra resultados que no contienen definiciones mediante el uso de árboles de decisión y etiquetas sintácticas. En este trabajo se presentan de forma separada los estudios y resultados que se hicieron y obtuvieron para cada uno de los tipos de las definiciones. Para el caso de las definiciones analíticas, obtienen una precisión del 58 % y un recall del 83 %.

[1] En `http://www.corpus.unam.mx/corcode`.

### Trabajos sobre extracción de hiperónimos

Las relaciones de hiponimia-hiperonimia son un tipo de relación léxico-semántica que es de vital importancia cuando se quiere estructurar construcciones lingüísticas como la definición analítica.

La extracción automática de relaciones semánticas es un tema clásico en lexicografía computacional, que se ha abordado principalmente utilizando diferentes enfoques basados en:

**Diccionarios:** las técnicas apoyadas por diccionarios (Calzolari, 1984; Alshawi, 1987; Richardson et al., 1993) son óptimas para descubrir relaciones hiponimia-hiperonimia y son capaces de alcanzar una gran precisión. Pero necesitan textos estructurados, que no siempre están disponibles. En algunos trabajos como el de Calzolari (1984), se llega a obtener una precisión de hasta el 90 %. En cambio, tienen la desventaja de que no se toman en cuenta términos específicos de un dominio, pues los diccionarios llegan a abarcar varios dominios.

En la actualidad existen otras herramientas que pueden presentar información de forma similar a un diccionario clásico. Por ejemplo Wikcionario, Wikipedia o Wordnet. Pero algunos de ellos no ofrecen la información como la que se quiere extraer. Wordnet sí contiene información estructurada, aunque el uso de la versión en español tiene muchos inconvenientes. Además, estos recursos y otros de parecidas características, no se actualizan con la asiduidad que permitiría un uso fiable para estos objetivos.

**Agrupamiento (o co-ocurrencias):** los métodos desarrollados para el enfoque de agrupamiento (Pereira et al., 1993; Riloff & Shepherd, 1997; Caraballo, 1999; Cimiano et al., 2004; Widdows, 2003; Mititelu Barbu, 2006) son capaces de encontrar hiperónimos incluso cuando no están explícitamente en el corpus de búsqueda. Sin embargo, necesitan textos muy amplios para obtener buenos resultados. Una de las ventajas que ofrece el método de agrupamiento radica en que es una alternativa que permite encontrar este tipo de relaciones semánticas, aun cuando no están explícitamente

en el corpus de búsqueda; dentro de sus desventajas está que no ofrece buenos resultados con textos pequeños (Ortega, 2007).

**Patrones:** las técnicas basadas en patrones léxicos simples manuales fueron desarrolladas por primera vez por Hearst (1992). Muchos autores han seguido este modelo (Ravichandran et al., 2004; Cederberg & Widdows, 2003; Pantel & Ravichandran, 2004; Oakes, 2005). Varios trabajos integran aprendizaje de máquina para optimizar los sistemas (Snow et al., 2004; Pasca, 2004; Pantel & Pennacchiotti, 2006; Pantel & Ravichandran, 2004; Bunescu & Mooney, 2007). Una de las principales desventajas de este método es la necesidad de corpus muy grandes para poder reportar resultados del 85 % en precisión.

Aunque la mayoría de la literatura está relacionada con el inglés, existen propuestas en varios idiomas que siguen métodos similares.

En español, Acosta et al. (2010) utiliza la teoría de prototipos y el aprendizaje de máquina para extraer las relaciones de un corpus, y luego calcula los pares hipónimo /hiperónimo. El trabajo de Ortega (2007) también se encuentra en el marco de los enfoques basados en patrones.

## 4 Metodología

En este apartado se explica cómo se llevan a cabo las tareas de extracción de definiciones y la obtención de las relaciones semánticas de hiponimia-hiperonimia. Creemos pertinente destacar que, con la intención de que el procesamiento fuera más ligero, el sistema no lleva a cabo ningún tipo de etiquetado gramatical, morfológico o sintáctico.

### Arquitectura del sistema

El sistema presentado trabaja en tres etapas: en la primera se hace una serie de búsquedas con la finalidad de extraer de la web candidatos a contextos definitorios; en la segunda, y una vez que se tienen dichos contextos definitorios, se refina la búsqueda utilizando un elemento que forma parte constituyente de la definición analítica, el nexus differentia (mismo que se describe a continuación); por último, una vez que se han extraído candidatos a definiciones más precisas, se hace posible que de ellas se puedan obtener hiperónimos, pues estos quedan delimitados entre los patrones definitorios y los patrones diferenciales. A continuación se esquematizan a detalle las tres etapas anteriores.

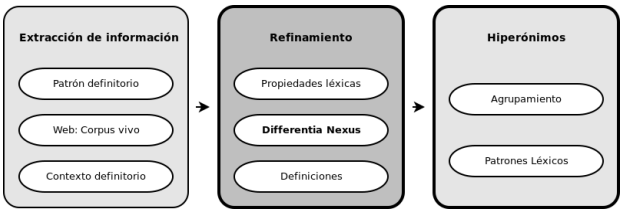


Figura 1: Etapas de la extracción de definiciones.

En la Figura 1 se describen gráficamente las etapas de este desarrollo. Los pormenores de cada una de ellas se describirán detalladamente en los apartados siguientes (4.2 Recuperación de información, 4.3 Extracción de la definición y *nexus differentia*, y 4.4 Extracción de hiperónimos).

### Recuperación de información

La meta de esta etapa es la extracción de contextos definitorios de la web. En este estadio se genera un corpus de candidatos a definiciones, con base en algunos patrones definitorios del español (ser, definir y concebir). Este corpus tiene poca precisión, pero un alto recall, lo cual es perfecto para nuestro sistema, pues a mayor información mejores resultados.

Cabe destacar que los verbos utilizados tienen la propiedad de extraer el tipo de definiciones analíticas. A diferencia de otros métodos que únicamente hacen búsquedas en Google All, este sistema también hace búsquedas en Google Scholar y Google Books para que los datos obtenidos puedan servir como el respaldo de un saber especializado.

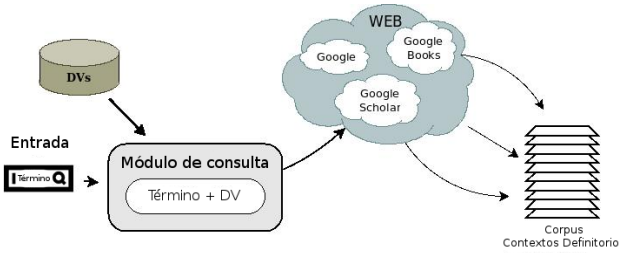


Figura 2: Extracción de Información.

En la Figura 2 pueden verse los pasos que sigue nuestro desarrollo para la extracción de contextos definitorios. Se ilustra al término siendo añadido a todos los verbos definitorios (DVs) considerados para ser buscados en cada uno de los módulos de Google. Los DVs se encuentran siempre en tercera persona del singular del presente de indicativo. La búsqueda genera un corpus de contextos definitorios que es utilizado en la siguiente etapa.

### Extracción de la definición y nexus differentia

En esta etapa se procesa la salida obtenida en el paso anterior y se suma uno de los *nexus differentia*, así como algunas características léxicas. El término *nexus differentia*, introducido por Dorantes (2016), consiste en el "pronombre relativo simple", que se ha observado ser esencial en la estructura de definiciones analíticas en español, ya que en estas existe una regularidad comprobada con base en la revisión de diccionarios, así como de una gran cantidad de candidatos a definiciones. El único "pronombre relativo simple" que se ha considerado en este trabajo es "que", ya que es el más generalizado. En posteriores etapas del trabajo, se piensa incluir otros relativos, como "el/la cual", aunque su uso en las definiciones es mucho más bajo que el del genérico "que".

En la mecánica de construcción de los patrones se siguen las siguientes reglas:

- El término a buscar debe aparecer en conjunto con alguno de los verbos definitorios, así como con un artículo que lo identifique como un sustantivo.
- No puede aparecer una palabra funcional previo a la detección del patrón antes descrito.
- Se filtran aquellas protodefiniciones que comienzan o terminan con palabras funcionales.
- Si una protodefinición contiene el mismo término que se quiere definir se descarta como definición.
- Por último, el sistema se asegura de que contenga el *nexus differentia* en una posición adecuada, es decir, después del término que se busca definir.

Existen algunas particularidades en la forma como el español caracteriza un *genus* respecto de su *differentia*. Si bien la *differentia* se había planteado anteriormente como un elemento de la definición analítica, no se habían hecho trabajos que mostraran las características lingüísticas de dicha partícula ni tampoco la forma en que dicha partícula introduce la diferencia específica.

Por su parte, entre las características léxicas podemos encontrar aquellas que explican tanto la forma como la medida estándar de un término. Suponemos que los términos se encuentran también circunscritos por uno o más elementos de una lista cerrada de palabras que reducen la posibilidad de aportar un hiperónimo al aparecer al inicio y/o al final de un contexto. Dichos elementos se agregan a una búsqueda en la que se fuerza su aparición, con lo que se consigue una definición que contiene un genus con su *differentia*. Lo anterior logra un aumento de la precisión en la extracción de definiciones. El proceso que sigue el desarrollo se ilustra en la Figura 3. La entrada del sistema es un corpus de contextos definitorios y la salida son las definiciones analíticas y, en una etapa intermedia, la suma de las características léxicas y el nexus differentia que permiten hacer el filtrado y ofrecer definiciones analíticas precisas.

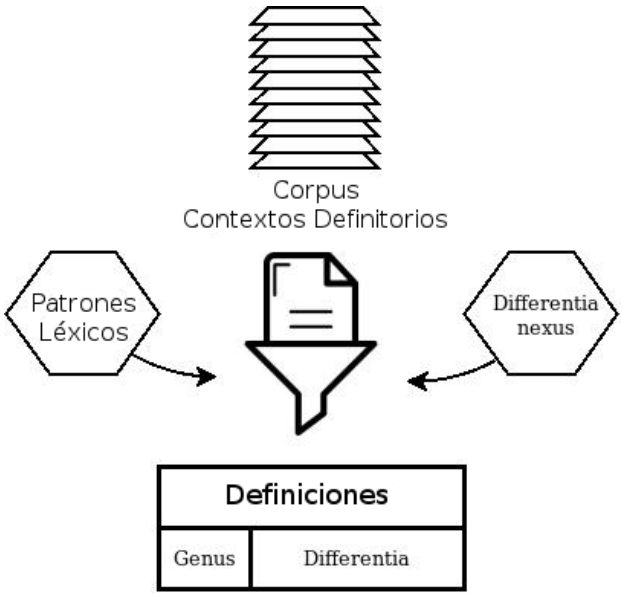


Figura 3: Refinamiento del corpus.

El Cuadro 1 puede servir para ilustrar qué es un contexto definitorio, la estructura de los patrones. Además, ejemplifica una definición analítica que sigue la formación del patrón.

### Extracción de hiperónimos

El sistema de extracción de hiperónimos trabaja en dos etapas: en la primera se extraen todos los genus y en la segunda se agrupan por frecuencia, ofreciéndonos así los candidatos a hiperónimos con su respectiva frecuencia de aparición.

A continuación se clarifican ambas etapas: en la primera se extraen todos aquellos elementos que se encuentran circunscritos tanto por los DVs como por el ND; mientras que en la segunda se agrupan todos los hiperónimos que son completamente iguales con la intención de mostrar qué hiperónimo podría ser el más pertinente para cada término. Tal como se puede ver en la Figura 4.

Al final de cada uno de los procesos es posible obtener definiciones precisas y candidatos a hiperónimos. Dichos resultados son posibles gracias

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lexicografía | Macroestructura | | | | | | Microestructura | | | | | | |
| | Entrada, lema, definido, definendum o | | Verbo definitorio | | | | Definición, definiens, expresión explicativa, descripción, explicación, exposición, explanación o declaración | | | | | | |
| | entidad léxica | | | | GENUS | | DIFFERENTIA | | | | | | |
| Artículo lexicográfico | Cajero automático | | | | Máquina | | que permite retirar y depositar dinero en efectivo a los clientes de un banco en casi cualquier parte del mundo a cualquier hora del día | | | | | | |
| Contexto definitorio | El | cajero automático | es | la | máquina | que | permite | retirar y depositar dinero en efectivo | a los clientes de un banco | en case cualquier parte del mundo | a cualquier hora del dia | | |
| Sintaxis | Art. 1 | Sintagma nominal | Verbo 1 | Art. 2 | Sintagma nominal | Pron. Rel. Nexus differentia | Verbo 2 | Sintagma nominal | Sintagma prep. | Sintagma prep. | Sintagma prep. | | |
| | M.D. | Núcleo Sujeto | Núcleo Pred | | Atributo | M.D. | Núcleo Pred. | CD | CI | CCL | CCT | | |
| | Oración principal | | | | | Oración subordinada relativa especificativa | | | | | | | |
| | Oración compuesta | | | | | | | | | | | | |

Cuadro 1: Estructura de una definición y del contexto definitorio de donde se extrae.

a la suma del *nexus differentia* que nos permite elevar la precisión en la tarea de la extracción de definiciones.

La extracción de hiperónimos, al igual que la de definiciones analíticas, es automática tomando en cuenta los resultados arrojados en el primer proceso (la extracción de definiciones). La tarea es sencilla, pues lo único que se hace es asociar el género del término definido (Figura 4).

A continuación se muestran los resultados y la evaluación de nuestro método con la intención de esclarecer la mejora propuesta.

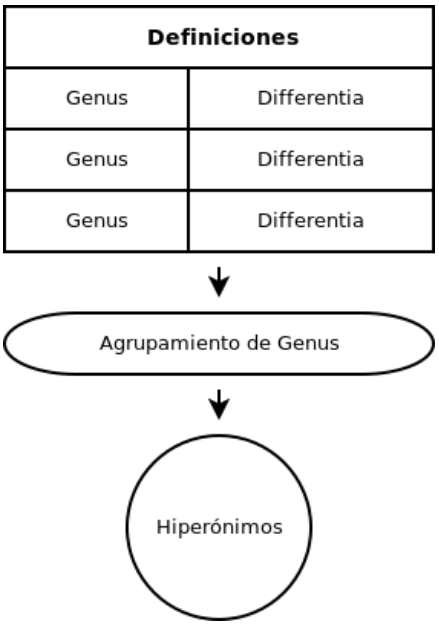


Figura 4: Extracción de hiperónimos.

## 5 Resultados

Los resultados de la extracción de las definiciones analíticas e hiperónimos se evalúan cuantitativamente y cualitativamente, respecto a otros métodos ya descritos anteriormente.

### Resultados cuantitativos de la extracción de definiciones

Tras correr las pruebas de extracción de definiciones y compararlas con un conjunto de definiciones analíticas obtenidas del CORCODE, ésta herramienta nos ofreció 985 definiciones analíticas dentro de un total de 1426 contextos definitorios recopilados.

El Cuadro 2 muestra la matriz de confusión que se presenta en el sistema; por un lado, la columna de datos de la izquierda muestra cantidad de resultados que el sistema extrajo como definiciones; por otra parte, en la derecha de la tabla

se muestra la cantidad de contextos descartados por el sistema al buscar definiciones. Al final, la primera fila de datos corresponde a la cantidad de contextos que efectivamente tienen definiciones, mientras que la segunda muestra la cantidad de los contextos que no presentan una definición.

| Real / Sistema | Positivo | Negativo |
|---|---|---|
| Verdadero | 147 | 838 |
| Falso | 12 | 429 |

Cuadro 2: Matriz de confusión que se presenta en el sistema.

Como se puede ver, se extrajeron 159 definiciones, de las cuales 147 fueron correctamente catalogadas, es decir, hay una gran cantidad de contextos que se clasifican correctamente como definiciones dentro del total de contextos catalogados por el sistema; sin embargo, la relación entre las definiciones obtenidas, del total de definiciones que se encontraban presentes, es muy baja. Por lo que se logra una precisión del 92.5 % y una exactitud del 40.5 %. Dicho resultado se obtiene de dividir las definiciones analíticas correctas entre las definiciones analíticas obtenidas (frecuencia relativa).

En este caso los datos podrían parecer desalentadores si lo que se busca no fuera la precisión, pero como creemos que este es un elemento importante para la generación de definiciones, este dato es el que consideramos que tiene mayor peso para desarrollos futuros.

Se hace pertinente contrastarlos con los ya existentes en el método que se pretende mejorar. De esta manera, podemos observar que en ECODE la precisión es de 58 %, mientras que la de nuestro desarrollo es de 92.5 %. Lo anterior nos permite decir que nuestro desarrollo mejora en un 32.5 % la precisión de ECODE.

En general, estos datos son buenos con miras a la generación automática de definiciones, pues ofrecen resultados confiables y más precisos.

### Resultados cualitativos de la extracción de definiciones

Con la intención de mostrar también los resultados cualitativos, es decir, algunas de las definiciones analíticas ofrecidas por el desarrollo, a continuación se presenta el Cuadro 3 donde se pueden ver cinco términos con las definiciones que ofrece el sistema propuesto.

En el Cuadro 3 se puede apreciar la pertinencia de las definiciones analíticas propuestas. Lo cual se debe a que todas cuentan con el ND del cual se ha hablado anteriormente.

| **Término** | **Definiciones ofrecidas por nuestro desarrollo** |
|---|---|
| Factor de activación de la transcripción | Proteína celular que en principio fue identificada como un factor estimulador de la transcripción de la unidad de transcripción e4 de adenovirus, la cual se activa en la fase inicial de la infección. |
| Servidor | Programa que se ejecuta en un terminal remoto y trabaja conjuntamente con el cliente. |
| Turbocompresor | Dispositivo de sobrealimentación que produce el ingreso del aire a una presión por encima de la atmosférica. |
| Relevador | Dispositivo que provoca un cambio brusco en uno o más circuitos eléctricos de control, cuando la cantidad o cantidades medidas a las cuales responde cambian de una manera predeterminada. |
| Entorno del Servidor de Internet de SGI (ISE) | Solución muy efectiva que incluye herramientas avanzadas de administración, monitoreo y seguridad además de programas integrados de instalación y una interfaz basada en la Web. |

Cuadro 3: Resultados cualitativos de nuestro sistema.

### Resultados cualitativos de la extracción de hiperónimos

Para esta sección, se han buscado ocho términos y han sido divididos en dos grupos según si los términos pertenecían a las ciencias de la salud (Biología, Cerebro, Hepatitis y Oncología) o no (Física, Matemáticas, Ontología, Sintaxis). La elección de estos términos ha venido determinada por los recursos con los que se contaba para evaluar. Así, se han elegido campos en los que existían ontologías, y en último caso se ha recurrido a wordnet, versión 3.0.

Para el primer grupo se ha utilizado la base de datos de Descriptores en Ciencias de la Salud (DeCS), cuyos conceptos se organizan con una estructura jerárquica para permitirle usarlo como una taxonomía (Cuadro 4). Mientras que para la otra mitad de los términos se ha utilizado el tesauro de la UNESCO,[2] que es una lista controlada y estructurada de términos utilizados para el análisis de temas y la búsqueda de documentos en los campos de educación, cultura, ciencias naturales, ciencias sociales y comunicación (Cuadro 5).

[2] http://vocabularies.unesco.org/browser/thesaurus/es/

| **Término** | **Nuestro sistema** | **WordNet** | **DeCS** | **Humano** |
|---|---|---|---|---|
| Biología | Ciencia<br>Rama de las ciencias naturales | Ciencia de la vida<br>Bio-ciencia<br>Vida<br>Colección<br>Agregación<br>Acumulación<br>Montaje | Ciencia<br>Disciplina de ciencia<br>Biológica | Ciencia<br>Ciencia exacta |
| Cerebro | Órgano | Estructura neuronal<br>Inteligencia<br>Cognición<br>Conocimiento<br>Noesis<br>Intelecto<br>Variedad de carne<br>Órgano | Telencéfalo | Órgano<br>Cuerpo<br>Anatomía<br>Aparato nervioso central<br>Cabeza |
| Hepatitis | Inflamación de hígado<br>Enfermedad inflamatoria<br>Enfermedad | Enfermedad infecciosa<br>Enfermedad del hígado | Hepatopatía<br>Virosis | Enfermedad<br>ETS |
| Oncología | Especialidad médica<br>Especialidad<br>Rama de la medicina<br>Ciencia | Medicina<br>Especialidad médica | Medicina interna | Medicina<br>Cáncer<br>Disciplina<br>Estudio |

Cuadro 4: Para los términos de ciencias de la salud, se muestran los hiperónimos de la base de datos de DeCS.

| **Término** | **Nuestro Sistema** | **WordNet** | **UNESCO** | **Humano** |
|---|---|---|---|---|
| Física | Ciencia | Ciencia Natural | Ciencia<br>Ciencias Físicas | Ciencia<br>Ciencia exacta |
| Matemáticas | Ciencia | Ciencia<br>Disciplina científica | Ciencia<br>Matemáticas y estadística | Ciencia<br>Ciencia exacta<br>Números |
| Ontología | Rama de la metafísica de la filosofía<br>Ciencia de las esencias<br>Base de datos | Disposición<br>Organización<br>Sistema<br>Metafísica | Metafísica | Filosofia<br>Estudio<br>Metafísica<br>Rama de la filosofía |
| Sintaxis | Parte de la gramática<br>Conjunto de reglas | Estructura<br>Sistema<br>Esquema<br>Gramática | Gramática | Lingüística<br>Nivel de lengua<br>Orden<br>Lengua |

Cuadro 5: Tablas comparativas para los resultados dela evaluación del sistema. Ambas tablas incluyen los hiperónimos obtenidos de WordNet y de la anotación humana.

Como se puede ver, entre los términos ofrecidos por nuestro sistema y los ofrecidos por otros recursos, existe una semejanza y, en algunos casos como el de Hepatitis, una aportación que podría resultar pertinente a la hora de tratar de entender a la Hepatitis no sólo como enfermedad, sino también como una inflamación del hígado. Cabe destacar que ninguno de los conceptos cae fuera del área de especialidad del que proviene el término y la mayoría de ellos es el mismo o es similar al resultado consultado o al dado por (DeCS), WordNet, UNESCO o humanos. Adicionalmente, se ha preguntando a colegas universitarios por los hiperónimos relevantes de los términos dados, así como una comparación manual entre diferentes recursos para estimar la precisión de nuestro sistema.

La evaluación elaborada mediante una metodología cualitativa arroja los resultados de el Cuadro 6.

| Precisión | Exactitud | Exhaustividad | F-Score |
|---|---|---|---|
| 92.5 % | 40.4 % | 14.9 % | 25.7 % |

Cuadro 6: Resultados de la evaluación cualitativa.

La precisión nos indica un factor de pertinencia. Consiste en encontrar la proporción que existe entre los resultados que se extrajeron como definiciones y los que realmente son definiciones.

Por su parte, la exhaustividad se refiere a la relación de la cantidad de definiciones que fueron extraídas contra todas las definiciones que pudieron haber sido obtenidas del corpus.

La exactitud es una medida que obtiene la proporción de oraciones clasificadas correctamente, es decir la unión de verdaderos/positivos y falso/negativo en proporción con el total de elementos en la evaluación.

Por último, el F-score es un valor único ponderado de la precisión y la exhaustividad. Se trata de una media armónica que combina ambos valores.

### Resultados cuantitativos de la extracción de definiciones

Como ya se dijo, se aplicó una encuesta a un grupo de 170 voluntarios para recuperar la aceptación que el estándar humano podría dar a cada uno de los candidatos a hiperónimos dados por el sistema. Ellos pertenecen a los últimos semestres de los estudios de Lengua y Literatura Hispánicas de la UNAM. Se recurrió a ellos tomando en cuenta todo su conocimiento lingüístico adquirido.

En la encuesta se presentó a los voluntarios el par de candidatos a término con la instrucción de calificar el candidato como un hiperónimo correcto o incorrecto para el término. Una tercera opción fue dada para el caso donde el voluntario no podría decir la categoría por sus propios medios.

Como se puede observar en el Cuadro 7, la aprobación de los humanos es muy satisfactoria: pasan el radio del 75 % llegando hasta el 96 %. Esto permite afirmar que, si bien aun se pueden mejorar los resultados, se va por buen camino hacia una mejora sustancial en esta tarea.

| Término | Radio |
|---|---|
| Biologia | 82.1 % |
| Cerebro | 88.3 % |
| Física | 96.3 % |
| Matemáticas | 84.0 % |
| Sintaxis | 75.1 % |
| Oncologia | 84.3 % |
| Ontologia | 63.6 % |
| Hepatitis | 95.7 % |
| **Evaluación del sistema: 83.67 %** | |

Cuadro 7: Aprobación desde la perspectiva humana de un set de candidatos a hiperónimos dados por el sistema.

## 6 Conclusiones y trabajo futuro

El método propuesto ha mejorado la precisión en la extracción de definiciones y ha mostrado candidatos a hiperónimos que, aunque no han sido evaluados, parecen pertinentes para los términos de entrada. A diferencia de los patrones utilizados en contextos definitorios, lo que fuerza la salida tanto de una definición como de un hiperónimo es el *nexus differentia* que nos permite recuperar definiciones como las ya mostradas anteriormente.

Los resultados arrojados por el sistema apuntan hacia que los *nexus differentia* son un factor decisivo para el mejoramiento de la precisión en la extracción de definiciones analíticas e hiperónimos, por ello, como trabajo futuro se plantea explorar el resto de *nexus differentia* del español, con la finalidad de aumentar el recall de nuestro sistema.

En cuanto a los hiperónimos, asumimos que los resultados son aceptables. Aunque los términos resultantes no siempre pertenecen al mismo nivel semántico, están relacionados. Además, los cambios se pueden atribuir a diferentes niveles de especialización entre las fuentes, es decir, cuanto más especializada sea una fuente, más "cercano" será el hiperónimo dado para una palabra. Por ejemplo, una fuente no especializada puede dar como hiperónimo de perro: "animal", una fuente más especializada puede recuperar "canino", y una fuente aún más especializada podría devolver "canis lupus", todos los cuales son hiperónimos correctos.

Dicho todo lo anterior, creemos que en futuros trabajos será posible comenzar a generar definiciones que sean adecuadas para cada término solicitado y que al mismo tiempo nos permitan estructurar el conocimiento.

## Agradecimientos

Este artículo ha sido elaborado gracias a los proyectos PAPIIT IA400117 y Fronteras de la Ciencia 2016-01-2225.

## Referencias


Acosta, Olga, César Aguilar & Gerardo Sierra. 2010. A method for extracting hyponymy-hypernymy relations from specialized corpora using genus terms. En *Workshop in Natural Language Processing and Web-based Technologies*, 1–10.

Aguilar, Cesar. 2009. *Análisis lingüístico de definiciones en contextos definitorios*: Universidad Nacional Autónoma de México. Tesis Doctoral.

Alarcón, Rodrigo. 2006. Extracción automática de contextos definitorios. propuesta para el desarrollo de un eccode (extractor de candidatos a contextos definitorios). Proyecto de tesis de doctorado. Universitat Pompeu Fabra.

Alarcón, Rodrigo. 2009. *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios*: Universitat Pompeu Fabra. Tesis Doctoral.

Alarcón, Rodrigo, Carme Bach & Gerardo Sierra. 2008. Extracción de contextos definitorios en corpus especializados: Hacia la elaboración de una herramienta de ayuda terminográfica. *Revista Española de Lingüística* 37. 247–278.

Alarcón, Rodrigo, Gerardo Sierra & Carme Bach. 2008. ECODE: a pattern based approach for definitional knowledge extraction. En *XIII EURALEX International Congress*, 923–928.

Alshawi, Hiyan. 1987. Processing dictionary definitions with phrasal pattern hierarchies. *Computational Linguistics* 13(3–4). 195–202.

Bunescu, Razvan C. & Raymond J. Mooney. 2007. Learning to extract relations from the web using minimal supervision. En *45th Annual Meeting of the Association for Computational Linguistics (ACL'2007)*, 576–583.

Calzolari, Nicoletta. 1984. Detecting patterns in a lexical data base. En *22nd Annual Meeting of the Association for Computational Linguistics (ACL'1984)*, 170–173.

Caraballo, Sharon. 1999. Automatic constraction of a hypernym-labeled noun hierarchy from text. En *37th Annual Meeting of the Association for Computational Linguistics (ACL'1999)*, 120–126.

Cederberg, Scott & Dominic Widdows. 2003. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. En *The SIGNLL Conference on Computational Natural Language Learning (CoNLL'2003)*, 111–118.

Cimiano, Philipp, Andreas Hotho & Steffen Staab. 2004. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. En *16th European Conference on Artificial Intelligence (ECAI'2004)*, 435–439.

Dorantes, Miguel A. 2016. La estructura definitoria en lexicografía: sintaxis de la deinición analítica para sustantivos en un diccionario especializado. Tesis de Licenciatura. Universidad Nacional Autónoma de México.

Espinosa-Anke, Luis, Roberto Carlini, Horacio Saggion & Francesco Ronzano. 2016. DEFEXT: a semi supervised definition extraction tool. En *GLOBALEX Workshop, co-located with LREC'2016*, 24–28.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. En *14th Conference on Computational Linguistics*, 539–545.

Klavans, Judith L. & Smaranda Muresan. 2001. Evaluation of the DEFINDER system for fully automatic glossary construction. En *American Medical Informatics Association Symposium (AMIA'2001)*, 324–328.

Lara, Luis Fernando. 1997. *Teoría del diccionario monolingüe*. COLMEX.

Meyer, Ingrid. 2001. Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. En *Recent advances in Computational Terminology*, 279–302. John Benjamins.

Mititelu Barbu, Virginica. 2006. Automatic extraction of patterns displaying hyponym-hypernym co-occurrence from corpora. En *First Central European Student Conference in Linguistics*, s.pp.

Oakes, Michael. 2005. Using Hearst's rules for the automatic acquisition of hyponyms for mining a pharmaceutical corpus. En *Recent Advances in Natural Language Processing (RANLP'2005)*, 63–67.

Ortega, Rosa M. 2007. *Descubrimiento automático de hipónimos a partir de texto no estructurado*: Instituto Nacional de Astrofísica, Óptica y Electrónica. Trabajo de Fin de Máster.

Pantel, Patrick & Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. En *Conference on Computational Linguistics (COLING'2006)*, 113–120.

Pantel, Patrick & Deepak Ravichandran. 2004. Automatically labeling semantic classes. En *North American Chapter of the Association for Computational Linguistics Conference (NAACL'2004)*, 321–328.

Pasca, Marius. 2004. Acquisition of categorized named entities for web search. En *13th ACM international conference on Information and knowledge management*, 137–145.

Pearson, Jennifer. 1998. *Terms in context*. John Benjamins.

Pereira, Fernando, Naftali Tishby & Lillian Lee. 1993. Distributional clustering of English words. En *31st Annual Meeting of the Association for Computational Lingistics (ACL'1993)*, 183–190.

Ravichandran, Deepak, Patrick Pantel & Eduard Hovy. 2004. The terascale challenge. En *KDD Workshop on Mining for and from the Semantic Web*, 1–11.

Richardson, Stephen D., Lucy Vanderwende & William Dolan. 1993. Automatically deriving structured knowledge bases from on-line dictionaries. En *The Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, 69–79.

Riloff, Ellen & Jessica Shepherd. 1997. A corpus-based approach for building semantic lexicons. En *2nd Conference on Empirical Methods in Natural Language Processing (EMNLP'1997)*, 117–124.

Rodríguez, C. 1999. Operaciones metalingüísticas explícitas en textos de especialidad. Trabajo de investigación. IULA, Universitat Pompeu Fabra.

Sager, Juan C. 1993. *Curso práctico sobre el procesamiento de la terminología*. Pirámide.

Sierra, Gerardo, Rodrigo Alarcón, Cesar Aguilar & Carme Bach. 2008. Definitional verbal patterns for semantic relation extraction. *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication* 14(1). 74–98.

Sierra, Gerardo, Rodrigo Alarcón, César Aguilar, Alberto Barrón, Valeria Benítez & Itzia Baca. 2006. Corpus de contextos definitorios: una herramienta para la lexicografía y la terminología. En *X Simposio Iberoamericano de Terminología*, .

Smith, Robin. 2007. Aristotle's logic. Consultado el 3 de julio de 2016, de http://plato.stanford.edu/archives/win2007/entries/aristotle-logic/.

Snow, Rion, Daniel Jurafsky & Andrew Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. En *17th International Conference on Neural Information Processing Systems (NIPS'2004)*, 1297–1304.

Velardi, Paola, Roberto Navigli & Pierluigi D'Amadio. 2008. Mining the web to create specialized glossaries. *IEEE Intelligent Systems* 23(5). 18–25.

Widdows, Dominic. 2003. Unsupervised methods for developing taxonomies by combining syntactic and statistical information. En *North American Chapter of the Association for Computational Linguistics Conference (NAACL'2003)*, 276–283.