Pontificia Universidad Católica de Valparaíso

Facultad de Filosofía y Educación

Instituto de Literatura y Ciencias del Lenguaje

Asignación de hiperónimos para sustantivos polisémicos en una taxonomía de generación automática: propuesta metodológica a partir de las similitudes de coocurrencia verbal de hipónimos de segundo orden

Tesis para optar al grado académico de Licenciado en Lengua y Literatura Hispánica

Profesores guía:

Rogelio Nazar

Irene Renau

Estudiante:

Javier Obreque Zamora

Viña del Mar, enero de 2019

Agradecimientos

En primer lugar, a Javiera Andrea, mi esposa, por su cariñoso compañerismo y apoyo en la cotidianidad de los días de estudio y trabajo que comprometió la realización de esta tesis.

En segundo lugar, a Irene Renau y Rogelio Nazar, mis profesores guía, por su entusiasmo y dedicación en la construcción de este estudio, y por su acompañamiento jovial y fraternal durante este período formativo inicial.

Y en tercer lugar, a los compañeros de camino, particularmente aquellos con quienes he compartido inquietudes, conversaciones y, en ocasiones, acaloradas discusiones en torno al apasionante fenómeno del lenguaje.

Índice

1. Introducción

El presente trabajo de tesis aborda el fenómeno problemático de la polisemia en los sustantivos, que podría definirse como la situación en la que un mismo sustantivo puede tener, potencialmente, más de un hiperónimo o sustantivo definidor. La asignación de un hiperónimo para un sustantivo es una operación común y natural para el ser humano cuando, por ejemplo, generaliza nombres de personas, objetos, eventos, etc., y en la mayoría de las ocasiones es el contexto sintagmático o pragmático-discursivo el que permite desambiguar entre todos los candidatos posibles a hiperónimo. Sin embargo, esta tarea se vuelve problemática en el marco de la conformación de una taxonomía generada por un sistema automático. Uno de estos casos dificultosos bien podría ser el de una tríada de relación hiperonímica como la siguiente: *tarántula* ⇒ *araña* ⇒ *animal/artefacto*. En un caso como este, el hiperónimo de *tarántula* es *araña* sin que exista posibilidad de ambigüedad; ahora bien, el problema se presenta luego, porque *araña* es un término polisémico, y bien podría entenderse a *araña*, y, por consiguiente, su hipónimo y nuestra palabra diana *tarántula,* como un tipo de *animal* o como un *artefacto*, en cuanto a que también existen el tipo de 'lámparas arañas'. En estos casos, el error de la automatización de una taxonomía semasiológica (ver **subsección 2.4**) se puede producir en la serie de los nodos de hiperonimia, porque si el sustantivo diana es *tarántula,* entonces el hiperónimo de segundo orden o grado no puede ser *artefacto*, sino *animal*. Un sistema automático podría, pues, realizar erróneamente el vínculo *tarántula* ⇒ *araña* ⇒ *artefacto.*

En estos términos, la dificultad explicitada es comprendida como un problema de clasificación, uno de los tipos de dificultades más comunes manifestadas en la construcción de taxonomías automáticas de sustantivos (Bordea et al., 2015). Por taxonomías comprendemos las estructuras que emergen de la combinación de todas las relaciones de hiperonimia entre las unidades léxicas de una lengua, especialmente entre los sustantivos (Lyons, 1977). Estas construcciones son, según Cruse (1986), una de las más elementales descripciones semánticas de la lengua. De esta manera, por ejemplo, se puede indicar que una relación taxonómica entre dos términos se manifiesta cuando el significado de un hipónimo incluye el de su hiperónimo, como en el caso en que *perro* incluye el significado de *animal*, porque este actúa como hiperónimo del primero. De ahí que las relaciones de

herencia semántica sean un verdadero problema que solucionar en cuanto a la generación automática de taxonomías léxicas.

Ahora bien, nuestro problema se centra en las taxonomías pobladas de manera automática. Hacer esta distinción es necesaria porque da cuenta de que también existe su contraparte, es decir, las estructuras pobladas de forma manual, cuyos antecedentes desarrollaremos con mayor detención en el marco teórico de este trabajo (**subsección 2.4**). *A priori*, diremos, por un lado, que estas últimas se vuelven rápidamente obsoletas debido al dinamismo de la lengua en el eje temporal y a la demora en el tiempo que plantea su actualización, y, por otro lado, que las primeras deben lidiar con problemas de estructura y polisemia que hoy mantienen –en cuanto a los resultados– un rango de precisión que fluctúa entre el 60% y el 80% (Nazar y Renau, 2016). Esta medida está por debajo de lo esperado y con el presente trabajo creemos posible aumentarla.

En función de la anterior cifra porcentual y, por cierto, del problema que hemos planteado, nos proponemos resolver la pregunta sobre cómo logramos que a partir de un proceso automático *tarántula*, hipónimo de *araña*, que es, a su vez, hipónimo de *animal*, se enlace con *animal* en vez de con *artefacto*, para luego reestablecer la relación hiperonímica entre los tres sustantivos, desambiguando el sentido de *araña*. En efecto, nuestro objetivo general es construir un método que, utilizando la información semántica proporcionada por un hipónimo de un sustantivo polisémico, logre seleccionar automáticamente para él un sustantivo hiperónimo entre dos o más posibles sentidos. Esto, a partir de una comparación de similitud estadística extraída de una amplia muestra de coocurrencias de verbos asociados cotextualmente al hipónimo de segundo grado (ej. *tarántula*) y a los posibles hiperónimos de segundo grado (ej. *animal* y *artefacto*). Por eso, el hipónimo del sustantivo polisémico, es decir, *tarántula*, es tan importante para el proceso –de ahí la formación de tríadas–, porque son sus coocurrencias léxicas, particularmente de tipo verbal, las que inciden directamente en la selección adecuada del hiperónimo (*animal*/*artefacto*) de su propio hiperónimo (*araña*). Este es el caso que intentamos clarificar en la **figura 1**.

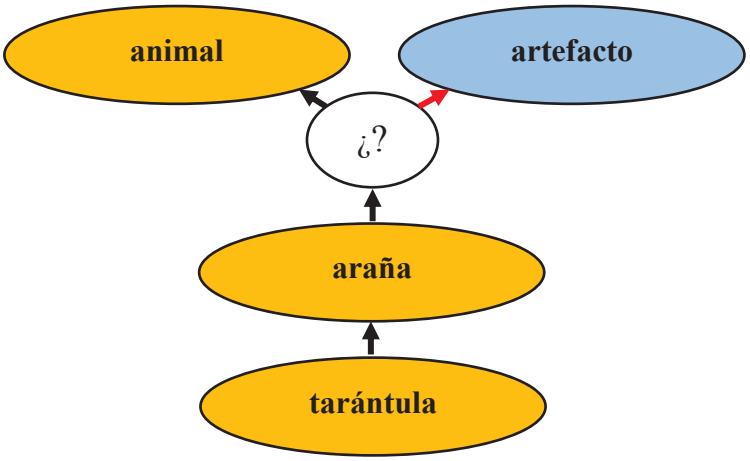

Figura 1: Estructura taxonómica (elipses amarillos) que presenta un evento polisémico a partir de su segundo elipse (araña), cuya selección correcta debiera ser animal. No obstante, ocurre que en estas estructuras aún existen selecciones inconsistentes, como la presentada en el elipse celeste.

En función del objetivo general propuesto, se ejecutará una metodología que considera, por un lado, los principios de la semántica léxica y, por otro lado, los principios de la lingüística distribucional aplicados en la estadística de corpus por medio de las herramientas que nos ofrece la lingüística computacional. En concreto, se elige como principal elemento clasificador la coocurrencia verbal en el eje sintagmático de distintos sustantivos. Estos últimos son, por una parte, nuestras unidades de análisis, es decir, las que debemos clasificar. Los verbos, por otra parte, son las variables que utilizamos para la discriminación en dicha clasificación. Aunque su utilización como variables se relaciona con la necesidad de acotar el estudio, también responde a la capacidad predictiva que eventualmente podrían tener en razón de los sustantivos con los que se vinculan con mayor frecuencia. De este modo, y como ejemplo, por un lado, el sustantivo *tarántula* se vincula en una frecuencia considerable con verbos como *tener*, *estar*, *volver* e *ir* (relacionados con seres vivos). Por otro lado, el vector clase (ver **apartado 3.6.2**) generado para *artefacto* a partir de los verbos que coocurren con diez hipónimos de segundo grado de *artefacto*, párametro elegido de manera arbitraria para la construcción, se vincula con verbos como *traer, costar, regalar* o *requerir*, *a priori* distintos que los verbos que se encontrarán con más frecuencia en el eje sintagmático de los sustantivos de seres vivos, en que

encontraremos (en el caso de *araña*) *tejer, ir, cazar, habitar* o *alimentarse*. Así pues, se crea una matriz en la que cada variable es un verbo y las unidades de muestra vectoriales son los casos diana (hipónimos del hiperónimo a desambiguar) y los vectores clases ya creados (tipos de hiperónimos). A partir de la información contenida en esta matriz, y haciendo uso de un algoritmo creado para este problema (ver **apartado 3.6.5**), logramos entonces hacer uso de la información ordenada en ella (coocurrencia verbo-sustantivo), para extraer el grado de similitud de un hipónimo (diana) como *tarántula* con un tipo de vector clase del carácter de hiperónimo de segundo grado como *animal o artefacto*.

Los resultados que se pretenden alcanzar para el desarrollo de esta tesis de final de grado son dos, consecutivos el uno del otro. Por una parte, se procura comprobar que la relación sustantivo-verbo, desde la perspectiva de la estadística de corpus, es capaz de clasificar y desambiguar un hipónimo polisémico en su hiperónimo correcto siguiendo la lógica relacional de las taxonomías. Por otra parte, y como consecuencia de lo anterior, se pretende producir una tabla de las clasificaciones efectuadas por el algoritmo creado a partir de la matriz que manifiesten el grado de similitud de la palabra diana (hipónimos) con los vectores clases (hiperónimos de segundo grado), y, por lo tanto, mostrar cómo estas logran vincularse correctamente con ellos.

Este trabajo de tesis tiene una alta relevancia en el ámbito del procesamiento del lenguaje natural y la lingüística computacional. Por medio de la aplicación de las herramientas de esta última se está proponiendo un método que contribuye a solucionar el problema de la desambiguación semántica de manera automática. Y, por cierto, su impacto repercutirá de manera considerable en el mejoramiento de la construcción de taxonomías léxicas automáticas, estructuras de gran uso en diversos ámbitos científicos y tecnológicos.

El desarrollo de este trabajo se compone de 3 capítulos. En primer lugar, un marco teórico que aborda las unidades que están en el centro de nuestra propuesta, a saber: el significado léxico, el fenómeno de la polisemia, los presupuestos de la lingüística distribucional y la construcción de taxonomías de sustantivos (**apartado 2**). En segundo lugar, un marco metodológico que explica la forma de ejecución de la propuesta (**apartado 3**). Y en tercer lugar, la presentación analítica de los resultados del método ejecutado y sus posibles alcances (**apartado 4**). Por último, se explicitan las conclusiones del trabajo (**apartado 5**).

2. Marco teórico

En este apartado se realiza un recorrido por los conceptos y concepciones teóricas que sostienen la propuesta de clasificación que desarrollaremos en la metodología (**apartado 3**). Este despliegue teórico se presenta del siguiente modo:

1) Se expone desde qué perspectiva abordaremos el significado y el léxico, definiendo qué entendemos por palabra, los tipos de relaciones de sentido entre las unidades que se categorizan como tal y el rol del contexto donde aparecen (**subsección 2.1**).
2) Se explica qué es y cómo se manifiesta el fenómeno de la polisemia en el caudal léxico (**subsección 2.2**).
3) Se hacen explícitos los principios de la lingüística distribucional que están en la base de nuestra propuesta metodológica para tratar el significado léxico (**subsección 2.3**).
4) Se desarrolla qué son las taxonomías léxicas, su función, antecedentes y estado actual de sus resultados (**subsección 2.4**).

2.1 El significado léxico

Desde la formación de las primeras comunidades, los seres humanos reconocieron en el lenguaje un carácter creativo y poderoso, divino para algunos. Sin este instrumento con el que el hombre da forma a su pensamiento, sentimientos, emociones, aspiraciones y voliciones (Hjelmslev, 1971), y que es, asimismo, la manifestación plena de una actividad humana, no habría cultura. Todo lo que hace el hombre nos viene vehiculizado por el lenguaje (Gadamer, 2015). No obstante, lo que hace que a algo lo llamemos lenguaje, es decir, que comunique intencionadamente, es que tenga la capacidad de significar. No hay lenguaje sin significado (Valdés, 1991), y no hay lenguaje que signifique sin palabras que lo conformen, aquellas que resultan de lo que Martinet (1974) distingue como primera articulación en el fenómeno de lo que llamó doble articulación del lenguaje.

En la Antigüedad clásica las palabras fueron descritas como meras unidades básicas que se combinan (Ullman, 1972; Piera, 2009). No obstante, en el presente trabajo entenderemos por palabra a una forma libre mínima capaz de sostenerse a sí misma y actuar como expresión completa (Bloomfield, 1964; Ullman 1972; 1979), es decir, dotada de significado (Lyons, 1997). Parcialmente distinta es la unidad básica lexema, que no comparte la característica de independencia y movilidad que sí tiene la palabra, y cuya mayoría de casos

no puede aparecer aisladamente como expresión completa (Escandell, 2007). El léxico de una lengua está formado por el conjunto de estas unidades.

El léxico es el componente estructurado, dinámico y creativo que constituye parte esencial del conocimiento que un hablante tiene de la estructura de su lengua (Felíu, 2009). Su dinamismo, manifestado en la creación de nuevas palabras y muerte o inutilización de otras (Hanks, 2013) es la causa de que se transforme en el lugar donde con mayor rapidez se muestran los cambios de una lengua (Álvarez de Miranda, 2009). Ejemplos paradigmáticos de ello son los fenómenos de la neología léxica –o la aparición de nuevas palabras– y la neología semántica –es decir, la aparición un nuevo significado sobre una base léxica ya existente– (Cabré, 2006; Álvarez de Miranda, 2009).

Ahora bien, las palabras que componen la estructura léxica de una lengua establecen entre sí relaciones formales y semánticas (Felíu, 2009) que García y Pascual (2009) proponen agrupar en siete casos. Primero, la relación de inclusión, que se manifiesta de dos maneras. Por una parte, como de hiperonimia-hiponimia, es decir, en la que un término incluye el significado de otro (Leech, 1977). De esta manera se puede decir que *mujer* es hipónimo de *humano* porque incluye su significado. Por otra parte, de tipo holonimia-meronimia, o de relación del todo a las partes (Escandell, 2007), como cuando distinguimos *cuerpo* (holónimo) y *mano* (parte de un cuerpo, merónimo). En la holonimia, a diferencia de en la hiperonimia, el significado de *cuerpo* no se corresponde con el de *mano*. Segundo, la relación de exclusión, que referencia a palabras que se contraponen semánticamente en varias de sus posibles dimensiones –por lo que no es correcto decir que se corresponde con términos antónimos– (Leech, 1977). Un ejemplo de exclusión es el de *ornitorrinco* (tipo de animal) y *atornillador* (tipo de artefacto), sustantivos que bajo ninguna de sus dimensiones podrían tener algún tipo de relación semántica. Tercero, la relación de identidad, comprendida como el caso en el que dos palabras comparten parcialmente el mismo significado (Leech, 1977; García y Pascual, 2009). Un ejemplo podría ser el de la relación existente entre los verbos *demoler* y *derribar*. Este tipo de relación es usualmente llamada como sinonimia parcial porque se corresponde con la relación de términos que poseen el mismo contenido descriptivo, pero no pueden utilizarse ambos en todos los contextos posibles (Escandell, 2007). Cuarto, la relación de oposición semántica sí se corresponde

con la antonimia, fenómeno que se entiende como la oposición de palabras que aunque comparten su categoría tienen un sentido opuesto (Leech, 1977; Espinal y Mateu, 2014). Tal es el caso de las siguientes composiciones binarias: *material*/*inmaterial*, *alto*/*bajo*, *viejo*/*joven*, etc. Quinto, las relaciones de campo léxico, referidas a los casos de palabras que, siendo distintas, comparten un campo de significación o área conceptual (Ullman, 1979; Espinal y Mateu, 2014). Es el fenómeno que sucede, por ejemplo, entre *casado, soltero, divorciado* y *viudo*, que sin significar lo mismo, se asocian a un campo léxico común, en su caso, el estado civil de las relaciones formales de una persona. Sexto, la relación de combinatoria léxica, que hace referencia al fenómeno lingüístico en que ciertas palabras se asocian comúnmente en el eje sintagmático, lo que a menudo restringe el significado de entre los posibles en una palabra (García y Pascual, 2009). Este es el caso, por ejemplo, en que un verbo como *abordar* toma un significado distinto, y en consecuencia, restringe el acceso de otro, si está acompañado por un sustantivo como *nave*, a diferencia de estar acompañado por otro sustantivo como *problema*: *abordar una nave* / *abordar un problema*.. Y séptimo, las relaciones con el mundo, referidas a los grados de implicancias de factores extralingüísticos que afectan el significado de las palabras y como, asimismo, las palabras categorizan realidades en el mundo (García y Pascual, 2009). Este tipo de relaciones son habituales hoy con las metáforas de las tecnologías de la información, donde el enunciado *abrir una ventana* verá afectado su significado habitual, y podrá tanto responder a la *ventana* como objeto físico o a la *ventana* como objeto virtual inmaterial si este se manifiesta en el contexto de la construcción de un edificio o de reunión de trabajo en el ámbito informática. Estos siete tipos de relaciones son una descripción muy útil que permite, a su vez, describir –entre otros fenómenos lingüísticos– los significados de toda la estructura léxica o vocabulario de una lengua.

Retomando el primero de los tipos de relación de inclusión semántica previamente descritos, un hiperónimo es un término cuyo significado está en un nivel de abstracción más alto que el de su hipónimo (Gómez Macker y Peronard, 2005). Este podría ser el caso de la palabra *árbol,* que es el hiperónimo de tipos de árboles como *alerce*, *mirto*, *peumo* o cualquier otro. Estos últimos, en razón de *árbol* son hipónimos, pues heredan el significado de su hiperónimo, pero lo aplican a especies específicas (Cruse, 1986; Espinal y Mateu, 2014). Dada esta explicitación, y considerando que Lyons (1983) señala que la estructura

léxica de una lengua puede ser descrita como una red de relaciones de sentido en la que todas sus partes están conectadas, tal y como sucedería con una tela de araña, la distinción de relaciones de hiperonimia es importante. Esta importancia radica en la posibilidad de que la distinción de todas las relaciones de hiperonimia, que implica analíticamente una relación hiponímica (Espinal y Mateu, 2014), permite la descripción semántica de toda la estructura léxica de una lengua (Lyons, 1977).

No obstante, evidenciar el significado de una palabra es una acción problemática, fundamentalmente porque las unidades léxicas no son portadoras de un significado unívoco. En este sentido, importante es la tradición lingüística y filosófica que ha sido muy clara en enfatizar que el significado de una palabra no puede ser descubierto en razón de la palabra aislada, sino en relación con su contexto de aparición, es decir, de uso, con el que está intrínsecamente ligado (Malinowski, 1923; Harris, 1954; Bloomfield, 1964; Guiraud, 1976; Trujillo, 1976; Ullman, 1979; Wittgenstein, 2009; Gadamer, 2015; entre otros). A partir de la distinción empírica de este fenómeno, Hanks (2013) concluye que una unidad léxica puede tener más de un significado potencial, activando uno de estos sentidos de acuerdo con el contexto (sintagmático) en el que aparece. A modo de ejemplo de este último fenómeno vinculado a la coocurrencia léxica en el eje sintagmático, presentamos a continuación, a través de datos proporcionados por el corpus esTenTen (Kilgarriff y Renau, 2013), el caso de la unidad léxica *gato* para ejemplificar un fenómeno común de desambiguación del significado léxico.

> **Ejemplo 1**: *Las uñas del **gato** tienen varias capas. Cuando un **gato** trepa a un árbol o utiliza un rascador, se le cae la capa externa de la uña y aparece [...]*
>
> **Ejemplo 2**: *Esto es aprovechado por sus dueños para cambiar las ruedas de su vehículo y abusar del **gato**. En realidad, cambiar las ruedas es una excusa [...]*

En estos dos ejemplos (**1** y **2**), se comprueba que la forma léxica *gato* es polisémica, es decir, posee más de un significado. Ello, porque observando ambos contextos sintagmáticos (cotextos), es posible advertir que el **ejemplo 1** corresponde al significado del hiperónimo *animal*, mientras que el sustantivo del **ejemplo 2** al significado del hiperónimo *artefacto*. En esta línea evaluativa del cotexto, los ejemplos **1** y **2** entregan evidencias sobre su distinción. Por un parte, en el **ejemplo 1** *gato* actúa como complemento del nombre del

sustantivo plural *uñas* (un merónimo típico de un felino), así como también es sujeto del verbo *trepar*. Por otra parte, en el **ejemplo 2** el sustantivo *gato* se manifiesta en posición de objeto indirecto de la acción humana que transporta el verbo *abusar*. Aunque la utilización de este verbo (*abusar*) puede ser ambigüa para diferenciar entre el mamífero y la máquina, cotextualmente este abuso de los dueños se vincula directamente con el verbo *cambiar* y cercano a sustantivos como *ruedas* y *vehículo*. Es decir, y según el cotexto, si es hay abuso, se debe al evento de cambiar ruedas al vehículo. Por medio de estas vinculaciones cotextuales el *gato* aquí es un objeto a utilizar, un *artefacto* directamente relacionado con la acción de cambiar las ruedas. Ambos significados corresponden a las acepciones 1 y 5 de *gato* en el *Diccionario de la lengua española* (DLE, 2014, en línea). La primera acepción es el "mamífero carnívoro de la familia de los félidos", mientras que la quinta acepción es la "máquina que sirve para levantar grandes pesos a poca altura", utilizada para levantar vehículos también.

2.2 El fenómeno de la polisemia

Como ya se ha adelantado, entendemos por polisemia el fenómeno por el cual una palabra o entrada léxica tiene asociado más de un significado (Espinal y Mateu, 2014). Casos como el que acaba de explicarse en el apartado anterior, la palabra *gato*, no son en absoluto un defecto del lenguaje, sino de un rasgo esencial de este (Ullmann, 1972; 1979). Esta multifuncionalidad de los signos lingüísticos manifestada en la polisemia, señala Battaner (2008), es un fenómeno que le permite a los sistemas de lengua ser muy potentes y económicos.

La causa de origen de la polisemia es la diversificación del significado de una base léxica. Según Escandell (2007), ello responde a variados fenómenos manifestados en el uso. Primero, la designación de nuevos objetos cuando el ya conocido comparte ciertas notas características con el objeto nuevo a designar (ej. *pluma* de ave y *pluma* de artefacto de escritura). Segundo, la utilización de una palabra en un ámbito especializado, como podría ser entendida la palabra *anillo* en áreas como la química o la astronomía. Tercero, el resultado de procesos de metaforización (usos figurados). Y cuarto, la adopción de calcos semánticos, según el cual en una lengua una palabra adopta el significado que tiene una palabra semejante en otra lengua.

Ahora bien, cabe precisar que en el caso de la polisemia léxica los significados de la palabra en cuestión han de relacionarse entre sí, porque en caso contrario estaríamos frente al fenómeno de la homonimia. Este último, distinto a la polisemia, se considera como la relación entre dos palabras o bases léxicas distintas[1].

2.3 Los principios de la semántica distribucional aplicada al análisis del significado lingüístico

El fenómeno del significado lingüístico ha preocupado al ser humano desde el origen de la reflexión teorética (Yallop, 2004). Sin embargo, desde el comienzo de los estudios lingüísticos ha sido considerado como una zona de desarrollo frágil y débil (Bloomfield, 1964). Por ello, se han realizado muchos esfuerzos para descubrir cómo actúa y se articula el significado en el lenguaje humano. Respecto a esta cuestión, desde los estudios lingüísticos ha emanado más de un tipo de aproximación al estudio del significado. De esta manera, indica Martí (2018), en las décadas de 1970 y 1980 fueron populares los modelos de carácter simbólico basado en reglas (basadas en la lógica), posteriormente se evolucionó a modelos híbridos, es decir, que combinaban el carácter simbólico y referencial de las reglas con datos empíricos, y por último, ya entrada la década de 1990, y debido al progresivo avance de la tecnología que permitió el procesamiento automático de grandes cantidades de datos (textos) –y que, por cierto, repercutió en los estudios sobre el lenguaje– se inauguraron modelos radicalmente empíricos basados en corpus –es decir, en muestras amplias y representativas de la variabilidad de una lengua (Gelbukh y Sidorov, 2010)– tratados de manera estadística; este último abordaje es el adoptado por este trabajo de tesis.

Ahora bien, los modelos radicalmente empíricos basados en estadística de corpus tienen su origen en los presupuestos de la escuela de Yale, cuyo más insigne representante es Leonard Bloomfield (Villar, 2009). Pero es Zellig Harris, a mediados del siglo XX, quien desarrolla más plenamente la perspectiva del estudio distribucional del lenguaje (Mounin, 1977; Villar, 2009). La propuesta analítica de Harris (1954; 1963), adelantada para su tiempo, y probablemente influida por el positivismo lógico y el empirismo de cuna anglosajona, representa todo un giro, una mirada completamente distinta de observar y

[1] Cabe precisar que la distinción entre homonimia y polisemia no es siempre, y en todos los casos, una diferencia clara, ya que la semejanza de significados es, en último extremo, una cuestión de grado ye, efecto, hay significados relacionados más estrechamente que otros. A veces, para poder determinar si estamos ante un caso de homonimia o polisemia es necesario conocer la etimología y el origen de la palabra (Escandell, 2007).

trabajar con los fenómenos del lenguaje (Mounin, 1977). El lenguaje ya no se describirá con las abstracciones de lo que los lingüistas piensan de manera introspectiva, sino con los datos reales que se manifiestan en el uso del lenguaje. El marco de estudio de esta propuesta científica es el principio de distribución en la lengua, principio que Harris (1963) define como la suma de todos los contextos en los cuales un elemento lingüístico se manifiesta[2], o, en otras palabras, la suma de todas las posiciones diferentes de un elemento lingüístico con relación a otros.

Los presupuestos teóricos y metodológicos de la lingüística distribucional, que adoptamos también en este trabajo de tesis, se resumen en cuatro puntos que exponemos a continuación. En primer lugar, las unidades que conforman el lenguaje no se manifiestan de manera arbitraria en situaciones comunicativas reales, sino que tienden a la regularización de coocurrencias entre elementos[3]. De este modo, no es una casualidad que elementos lingüísticos de distinta clase coocurran habitualmente juntos, sino que ello es una manifestación de la característica distributiva del lenguaje (Harris, 1954). En segundo lugar, es plausible agrupar en clases todos los elementos que conforman una lengua cuando su ocurrencia de manifestaciones se pueda establecer con exactitud en una muestra dada. Ahora bien, cuando la coocurrencia entre elementos se manifiesta entre unidades de distinta clasificación, entonces será necesario hablar en términos de probabilidad de acuerdo con la medición de la frecuencia en que ese fenómeno se presente (Harris, 1954). En tercer lugar, existe la posibilidad de establecer con exactitud lógica y matemática el grado de encuentro (coocurrente) y relación semántica entre unidades lingüísticas sin la necesidad de recurrir a cualquier otro tipo de información extralingüística (Harris, 1954). Y en cuarto lugar, cuando la ocurrencia de un fenómeno lingüístico es relativa, es decir, no es del todo clara y concluyente, declara Harris (1954), la descripción es más simple y, a su vez, estricta, utilizando una red de descripciones interrelacionada (por conjuntos de ocurrencias) en vez de una medición por separado de los mismos fenómenos. En otras palabras, si la ocurrencia de un fenómeno lingüístico no es concluyente, la mejor descripción resultará de la medición

[2] En 1926, el lingüista francés Antoine Meillet ya había hecho mención a este principio en *Linguistique historique et linguistique générale* (1952).

[3] La noción de coocurrencia pone especial énfasis en las relaciones sintagmáticas. Para los presupuestos distribucionalistas esta observación de la proximidad o el entorno es una relación combinatoria fundamental (Baylon y Fabre, 1994).

y comparación de grandes cantidades de datos de los fenómenos a estudiar –y de los adyacentes, relacionados con él– en conjunto.

En función de los presupuestos teóricos y metodológicos ya declarados, Harris (1954; 1963) elabora un marco relacional entre el lenguaje y el significado. De esta forma nacen los estudios de semántica distribucional bajo la idea de que si existe una correlación entre la similitud distributiva y la similitud del significado, entonces la primera (a cuyos datos tenemos acceso hoy) nos permite estimar la segunda (Sahlgren, 2008; Baroni, 2013; Martí, 2018). En esta línea, y bajo la aplicación de estos principios, será correcto indicar, por lo tanto, que si dos morfemas tienen significados diferentes, entonces también diferirán en alguna parte o grado de su distribución, entendiendo esta como la suma de todos los contextos en los que aparece (Hernández, 1991).

Ahora bien, en el marco del tratamiento del significado a través de las herramientas proporcionadas por la lingüística computacional, como lo es el caso de este trabajo de tesis, el modelo distribucional desarrollado por Harris (1954; 1963) hace medio siglo es aún del todo productivo. Primero, porque nos permite alcanzar el contenido semántico de una palabra basándonos en su distribución, obteniendo datos cuantitativos y graduales, prescindiendo de los rasgos simbólicos y categoriales propios de los modelos de referencia (Martí, 2018). Segundo, porque las representaciones del contenido semántico de una palabra son relacionales y no referenciales. Estas se expresan en vectores, entendidos como una lista ordenada de valores numéricos que manifiestan la cantidad de veces que una palabra a coocurrido en una muestra de corpus y su localización cotextual (en referencia con otras palabras). Son estos valores numéricos los que permiten que las representaciones de contenido semántico sean tratadas en un lenguaje computacional con grandes cantidades de datos y plausibles de ser aplicadas a cualquier idioma. Y tercero, y muy importante, es un método de análisis autónomo de la lengua que plantea una aproximación radicalmente empírica (Martí, 2018).

2.4 La descripción semántica de la lengua a partir de taxonomías

Clasificar es un verbo utilizado para expresar la acción de ordenar o disponer por clases distintos tipos de entidades (DLE, 2014, acep. 1). Esta acción de ordenar cosas ha sido una labor constante del ser humano desde el inicio de la historia documentada, cuánto más

también en el ámbito del conocimiento teorético y científico. Por ello, Aristóteles (s. IV a. C) ya pensó en distinguir entidades universales por sobre otras particulares en su *Metafísica* (trad. en 1994), que es, en el fondo, una propuesta de ontología, y presentó luego las distinciones entre tipos de nombres de entidades en sus *Categorías* (trad. en 1982), que podría considerarse un primer tratado sobre la organización conceptual de las entidades conocidas por el ser humano.

La clasificación de entidades ha permitido una infinidad de avances en cualquier ámbito que nos pudiéramos imaginar, como la distinción de fenómenos físicos, la diferenciación entre tipos de especies biológicas, el ordenamiento de teorías psicosociales y educativas, la aparición y funciones de nuevos proyectos tecnológicos y empresariales, etc. En la tarea de describir semánticamente la estructura léxica de una lengua, donde los esfuerzos han sido muchos, la clasificación de las unidades que la componen, es decir, las palabras, ocupa un lugar prioritario. Esta prioridad obedece a que son las relaciones de hiperonimia las que permiten la construcción de una descripción semántica de todas las palabras que componen el léxico de una determinada lengua. Por eso Lyons pudo indicar que la estructura léxica de una lengua, o, en otras palabras, su vocabulario podía ser descrito como una red de relaciones de sentido similar a "una enorme tela de araña multidimensional en la que cada tramo establece una relación" (1983, p. 81). Este fenómeno es evidente en la construcción de diccionarios, donde la definición de un sustantivo inicia –generalmente– con el hiperónimo, es decir, la explicitación del significado de un hipónimo incluye el significado de su hiperónimo, y así sucesivamente (Marín, 2009). Debido a ello es que podemos definir a un *rinoceronte* como un tipo de *animal* o a un *lápiz* como un tipo de *artefacto*.

En el marco de la lingüística contemporánea, los esfuerzos por la realización de una descripción semántica del vocabulario de una lengua se han incrementado considerablemente, sobre todo por las posibilidades que han ofrecido el desarrollo de la lingüística de corpus desde el último tercio del s. XX (Berber, 2000; Rojo, 2002; 2008), la inteligencia artificial y, por consiguiente, la lingüística computacional. En esta línea, sin la conformación de grandes conjuntos de textos auténticos en formato electrónico que garanticen altos parámetros de representatividad de la variedad de la lengua (Villayandre, 2008), la base empírica para un estudio del significado no sería posible. Desde este

contexto surge la idea de la construcción de taxonomías léxicas que consideren lo que Baldinger (1980) expresa como el desarrollo de una perspectiva semasiológica, vale decir, de la palabra hacia la manifestación de su significado, en oposición a una perspectiva onomasiológica, es decir, desde el concepto o significado hacia las posibles formas de manifestarlo[4].

Una taxonomía es la estructura que emerge de la combinación de todas las relaciones de hiperonimia entre los sustantivos de una lengua (Lyons, 1977). Esta estructura es, según Cruse (1986), la más elemental descripción semántica de las unidades léxicas que conforman la lengua, desde la más básica hasta la más compleja. En ella se explicitan las relaciones de jerarquía entre palabras en el plano semasiológico (Baldinger, 1980). De esta manera, por ejemplo, una relación taxonómica entre dos términos se manifiesta cuando el significado de un hipónimo hereda el de su hiperónimo, como en el caso en que *martillo* hereda las propiedades y sentido de *artefacto*, porque este último actúa como su hiperónimo.

Como lo hemos esbozado previamente, esta lógica de las relaciones taxonómicas no es en absoluto moderna. Es una idea clásica que ya estaba presente en el s. IV a. C. por medio de Aristóteles, fundamentalmente en el libro de las *Categorías*, pero amparado, por cierto, en todo su sistema filosófico: en su ontología metafísica (conceptual, no léxica) y en las reglas de lo que conocemos como lógica formal clásica, de ahí que sus explicaciones estriben en términos como 'substancia primera', 'substancia segunda' o 'especie'.

En razón de la definición de estructura taxonómica propuesta por Lyons (1977), se puede distinguir que estas poseen al menos tres características fundamentales en función de las relaciones que forman. En primer lugar, la herencia, concepto central para una estructura jerárquica. Por medio de esta propiedad se dice que un nodo inferior (hipónimo) hereda las propiedades de su nodo superior (hiperónimo), o, en otras palabras, el significado del inferior incluye el de su significado superior, de ahí que nazca su relación. En segundo lugar, la asimetría, según la cual una unidad léxica no puede ser superior (hiperónima) e inferior (hipónima) de otra unidad léxica al mismo tiempo. Y en tercer lugar, la

[4] Sobre esta cuestión, Baldinger (1977) especifica que se trata de puntos de vista complementarios, dado que este doble aspecto se corresponde con "la doble naturaleza del signo lingüístico como forma y contenido" (p. 120).

transitividad, según la cual los hipónimos de un determinado caso siempre pueden ligarse a su hiperónimo directo, y este, a su vez, a su hiperónimo directo, y así sucesivamente hasta alcanzar los tipos semánticos más generales o abstractos, teniendo como ejemplo que si (a → b) y (b → c), entonces también se puede decir que (a → c) (Nazar y Renau, 2016).

Los esfuerzos por construir taxonomías del componente léxico de una lengua han ido en progresivo aumento durante el último cuarto del s. XX y los albores de este s. XXI, principalmente por las posibilidades que nos ofrecen en el ámbito del procesamiento del lenguaje natural. La lingüística computacional, al aparecer en el panorama del desarrollo de la lingüística moderna, por un lado utilizó términos como *red semántica* u *ontología* para referirse a estructuras de datos relacionados y, por otro, aunque siempre en la búsqueda del procesamiento automático de una gran cantidad de datos, se interesó por el desarrollo de ontologías que pudieran ser leídas por un computador (Sowa, 2000). Los primeros intentos para construir taxonomías se desarrollaron, por un lado, de forma manual, en casos como los de CyC (Lenat, 1995), WordNet (Fellbaum, 1998), EuroWordNet (Vossen, 1998), Snomed (Stearns et al., 2001), entre otros; y, por otro lado, paralelamente se desplegaron esfuerzos para su construcción de manera automática a partir de diccionarios (Chodorow et al. 1985; Guthrie et al., 1990) y, luego, a partir de corpus (Hearst, 1992; Lin, 1998 ; Snow et al., 2006).

Ahora bien, cabe mencionar que tanto las estructuras de ontologías léxicas construidas de forma manual como las de generación automática son, aunque de un valor incalculable, actualmente problemáticas. Por un lado, las primeras se vuelven rápidamente obsoletas debido al dinamismo de la lengua en el eje temporal, y, por otro, las segundas deben lidiar con problemas de estructura y polisemia que hoy mantienen –en cuanto a los resultados– un rango de precisión que fluctúa entre el 60% y el 80% (Nazar y Renau, 2016), muy por debajo de lo adecuado. Sin embargo, las taxonomías generadas a través de procedimientos de automatización, aun con sus limitaciones, pueden contribuir a complementar e incluso sustituir las taxonomías generadas de forma manual, además de brindar la posibilidad de adaptar este tipo de recursos a diferentes idiomas y propósitos (Nazar y Renau, 2016), manteniéndolo permanentemente actualizado.

3. Marco metodológico

3.1 Planteamiento del problema

Las taxonomías generadas de manera automática tiene una precisión que todavía está por debajo de los parámetros adecuados (ver **subsección 2.4**). Solucionar el problema de la polisemia léxica que en ellas se manifiesta, y que provoca esos errores, podría aumentar su rango de exactitud. Como ya se expuso en el **apartado 1,** un caso problemático de polisemia léxica puede presentarse como la estructura taxonómica en tríada como el de la **figura 2**. *Laucha* es un tipo *ratón*, pero *ratón* (su hiperónimo polisémico) podría ser un tipo de *animal* o un tipo de *artefacto* ('periférico del computador'), por lo que *laucha* quedaría conectado potencialmente tanto a *animal* como a *artefacto*.

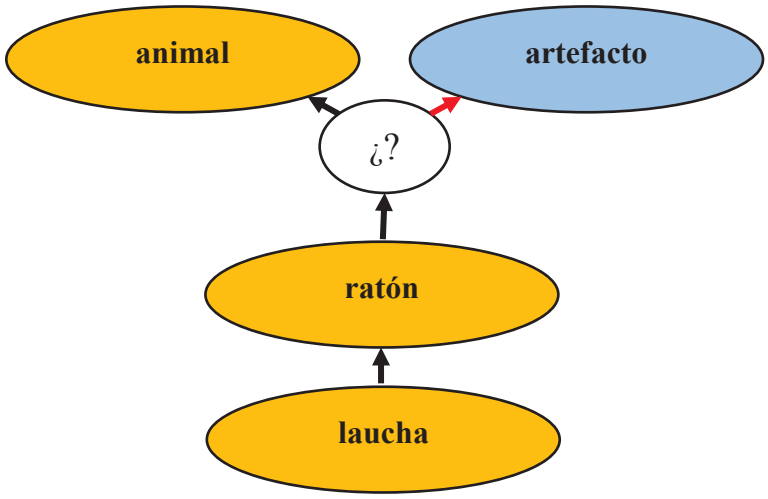

Figura 2: Planteamiento del problema en base a la tríada de relación taxonómica laucha-ratón-animal

Ahora bien, considerando que la operación por medio de la cual una máquina asigna un determinado significado a una palabra a partir de un inventario de significados disponibles por el sistema (Ide y Véronis, 1998; Nazar, 2013) no ha sido en lo absoluto fácil de resolver. En caso contrario, los resultados de precisión de las taxonomías de sustantivos no hubiesen quedado estancados como lo han demostrado las cifras aportadas por Nazar y Renau (2016), que distan de ser satisfactorias. Hoy resulta necesario e importante continuar proponiendo métodos de clasificación de significados de palabras para aumentar grado de perfección del procesamiento del lenguaje natural y, por cierto, y tal como lo hemos expuesto a través de nuestro caso, mejorar los sistemas generación automática de taxonomías que puedan describir la lengua. De este cometido nace la propuesta de este

trabajo de grado por crear un clasificador que seleccione automáticamente entre dos o más sentidos una palabra polisémica en calidad de hipónimo a través de la elección de su hiperónimo correcto, del que hereda su significado. Utilizando, para ello, métodos distribucionales basados en corpus que evalúen la similitud de coocurrencias verbales entre conjuntos de hipónimos de cada hiperónimo problemático. Esta es la propuesta que se ejemplifica en la **figura 3**.

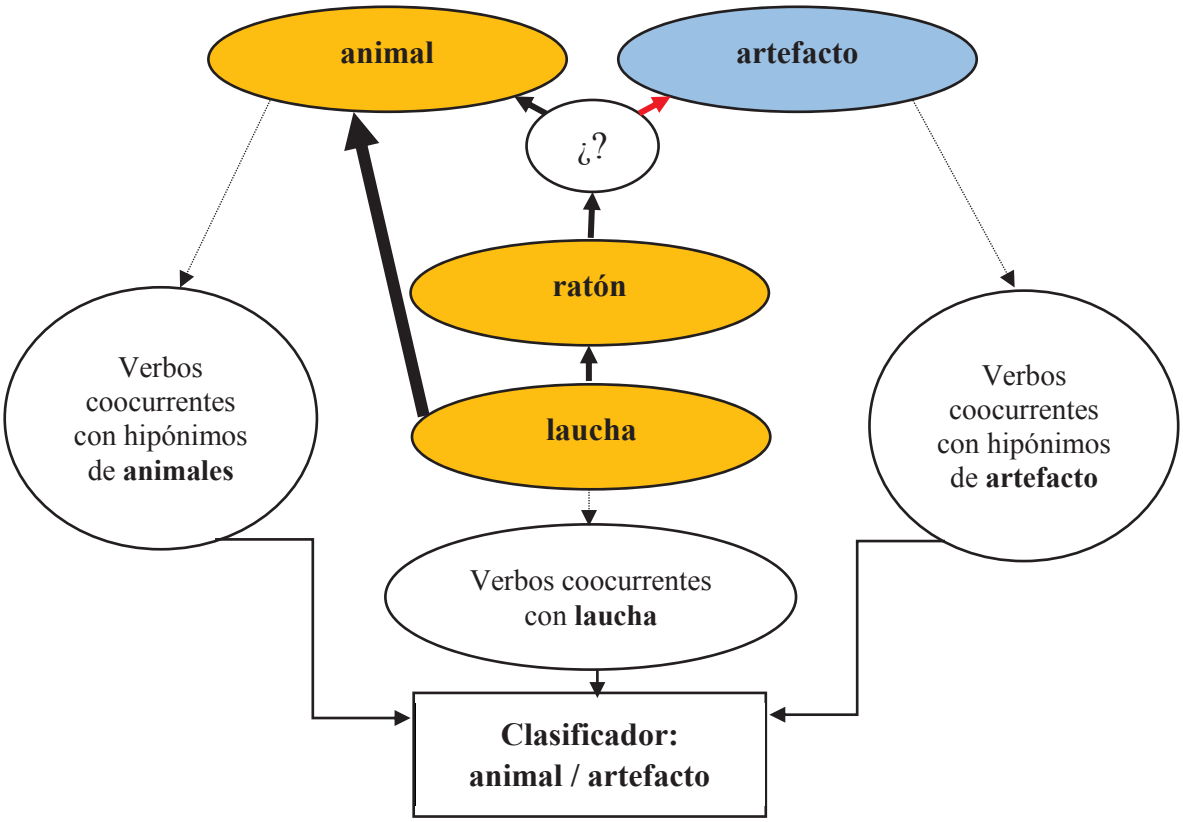

Figura 3. Se presenta el fenómeno problemático con la propuesta de resolución a través de la comparación de similitud de las coocurrencias verbales del hipónimo de segundo orden (laucha) con las coocurrencias verbales de los hipónimos de cada eventual hiperónimo (animal y artefacto para este caso). El algoritmo clasificador es el que elige a qué hiperónimo corresponde.

3.2 Pregunta de investigación

La pregunta de investigación que se propone responder por medio de este trabajo final de grado es la siguiente: ¿Cómo seleccionar, en el marco de una taxonomía poblada automáticamente, un hiperónimo de segundo orden cuando el hiperónimo de primer orden de un sustantivo hipónimo es polisémico?

3.3 Objetivo general

Construir un método que automatice la correcta elección de un hiperónimo de segundo orden cuando el hiperónimo de primer orden de un sustantivo hipónimo es polisémico.

3.4 Objetivos particulares

a) Determinar en qué grado es posible seleccionar el significado correcto de una palabra por medio de la aplicación del principio de distribución y la estadística de corpus.
b) Determinar si la medida de similitud de coocurrencia sustantivo-verbo comparada a través de conjuntos funciona como elemento de selección automática de hiperónimos de segundo orden, cuyo hiperónimo de primer orden de un sustantivo hipónimo es polisémico.

3.5 Hipótesis

La medida de similitud entre el vector de coocurrencia verbal de un hipónimo de segundo orden, cuyo hiperónimo de primer orden es polisémico, con el mismo tipo de vectores de sus posibles hiperónimos de segundo orden propuestos por una taxonomía automática, son capaces de seleccionar adecuadamente el nodo hiperonímico correcto.

3.6 Metodología

Se ha expresado que para la lingüística distribucional el significado lingüístico es diferencial y no referencial. Si el significado se correspondiera con este último caso (referencial), siempre se requeriría analizar y evaluar información extralingüística para extraerlo, y, por el contrario, en este trabajo hemos adherido al presupuesto de que las diferencias de significado entre dos términos están mediadas por diferencias de distribución. La aplicación de una metodología distributiva, como la que proponemos, permite descubrir el nivel de diferencia y similitud semántica entre unidades léxicas mediante una medición cuantificable y comprobable empíricamente (Sahlgren, 2008).

A continuación, se explica sistemáticamente cada etapa del proceso de construcción de un método que evalúa automáticamente la similitud entre vectores de coocurrencia verbal de hipónimos de segundo grado e hiperónimos de segundo grado, buscando medir el nivel de diferencia y similitud semántica.

3.6.1 Selección de casos y palabras diana

La selección de los casos que operan como muestra representativa en la creación de este método automático de clasificación de sustantivos hiperónimos de una estructura taxonómica fue arbitraria, pero con base en el conocimiento empírico. Es decir, se puso especial atención en que el caso fuera recurrente en una taxonomía de poblamiento automático. Para ello se observó el funcionamiento de estos casos en las herramientas de formación de estructuras taxonómicas Kind (Nazar y Renau, 2016) y CPA Ontology (Hanks, en línea). A partir de lo anteriormente descrito, se distinguieron relaciones taxonómicas en forma de tríada, es decir, de tres nodos, donde existían un hipónimo al que llamamos palabra *diana* (de entrada, ej. *laucha*), su hiperónimo de primer orden, que es polisémico (ej. *ratón*), y dos posibles hiperónimos de segundo orden entre los cuales el sistema debe elegir (ej. *animal / artefacto*). Como ya se ha ido exponiendo, esta tríada tiene una estructura jerárquica, vale decir, comienza por el sustantivo de carácter más particular (palabra *diana*) hasta ascender al tipo semántico más general y, por ende, abstracto (hiperónimo de segundo orden).

Los casos escogidos para esta investigación son los que se ordenan en la **tabla 1**. Esta utiliza tres columnas: la primera para explicitar la palabra *diana* (hipónima), la segunda para mostrar el hiperónimo de primer orden (*hiper1*), y polisémico en relación con la palabra *diana*, y la tercera para exponer los eventuales tipos semánticos en los que puede terminar la clasificación. A estos últimos les llamamos hiperónimos de segundo orden (*hiper 2*), y solo uno de ellos es el correcto.

	diana	**hiper 1**	**hiper 2**
1	guanaco	llama	animal | evento

2	caniche	perro	animal | artefacto | humano
3	tarántula	araña	animal | artefacto
4	siamés	gato	animal | artefacto
5	laucha	ratón	animal | artefacto
6	chimpancé	mono	animal | prenda
7	timbal	tambor	instrumento musical | artefacto
8	cencerro	campana	instrumento musical | máquina
9	hígado	órgano	parte del cuerpo humano | instrumento musical
10	verano	estación	período de tiempo | lugar

Tabla 1. Casos de eventos polisémicos en tríadas de estructuras taxonómicas.

3.6.2 Selección de tipos semánticos y construcción de vectores-clase

Los hiperónimos de primer y segundo orden (hiper 1 y 2) están tomados de la CPA Ontology de Hanks (en línea), una ontología que está basada en datos empíricos obtenidos de corpus y, por tanto, es coherente con la aproximación de esta tesis. En este sentido, los tipos semánticos, correspondientes a los hiperónimos de segundo orden, para nuestro experimento son diez: *animal*, *evento*, *artefacto*, *instrumento musical*, *prenda*, *máquina*, *lugar*, *parte del cuerpo humano*, *período de tiempo* y *humano*. Ahora bien, estos tipos semánticos (hiper2) no siempre aparecen lexicalizados en un corpus; es decir, puede ser frecuente encontrar *llama* asociada a *mamífero* o *rumiante* y menos frecuente a *animal.* Por ello, retomamos aquí una propuesta realizada por Schütze y Pedersen (1997) para construir un vector-clase (o vector-suma) que sea el resultado de la suma de los vectores o puntos de coocurrencia entre sustantivos hipónimos del hiperónimo en cuestión (hiper 2) y los verbos que se le asocian en los cotextos inmediatos. En otras palabras, el vector clase es una

construcción metodológica útil que pretende representar un tipo semántico abstracto a partir de los verbos con los que coocurre y, para el caso de nuestra investigación, serán diez vectores.

En este trabajo, la construcción de cada vector clase (hiper 2) se realiza a partir de la selección de 10 casos de hipónimos de cada tipo semántico. Lo que se pretende hacer con ellos en un paso posterior es la construcción de una matriz de coocurrencia entre el sustantivo y los verbos con los que aparece, para así generar el vector clase. Sin embargo, en esta parte del proceso cabe indicar que los 10 hipónimos de cada tipo semántico fueron seleccionados de manera parcialmente arbitraria. Esta parcialidad obedece a que parte ellos, los que estuvieran disponibles, fueron extraídos de los ejemplos de casos de tipos semánticos propuestos en la CPA Ontology (Hanks, en línea), mientras que otros, aquellos en que la CPA Ontology no proponía ejemplos de casos de corpus, fueron seleccionados arbitrariamente, colocando en ellos hipónimos del tipo semántico de manera arbitraria. Las **tablas 2** y **3** muestran los datos proporcionados por cada tipo semántico en la que hay 10 tipos semánticos (hiper 2) y, por ende, 100 casos de hipónimos que conforman, de diez en diez, cada vector clase.

	animal	**artefacto**	**máquina**	**período de tiempo**	**instrumento musical**
1	canario	pizarra	cocina	mañana	guitarra
2	gorrión	abridor	refrigerador	anoche	piano
3	elefante	secador	juguera	primavera	violín
4	león	bolso	motocicleta	otoño	trompeta
5	caballo	plancha	automóvil	invierno	flauta
6	delfín	candado	estufa	minuto	charango
7	lobo	cepillo	soldadora	siglo	saxofón

8	jirafa	ventana	aspiradora	diciembre	oboe
9	canguro	tubo	cortadora	semana	bombo
10	ornitorrinco	manguera	horno	marzo	clarinete

Tabla 2. Hipónimos seleccionados para la construcción de los vectores clase animal, artefacto, máquina período de tiempo e instrumento musical.

	prenda	**lugar**	**humano**	**evento**	**parte del cuerpo humano**
1	pantalón	estadio	estudiante	lluvia	pulmón
2	calcetín	hospital	soldado	precipitación	estómago
3	polera	colegio	electricista	erupción	riñón
4	chaleco	supermercado	profesor	tormenta	intestino
5	calzón	cárcel	mecánico	ciclón	corazón
6	cinturón	puerto	mamá	sismo	pies
7	blusa	plaza	mucama	huracán	pelo
8	sudadera	universidad	periodista	vendaval	cabeza
9	camisa	baño	cocinero	explosión	lengua
10	corbata	tienda	gobernador	incendio	brazo

Tabla 3. Hipónimos seleccionados para la construcción de los vectores clase prenda, lugar, humano, evento y parte del cuerpo humano.

3.6.3 Tratamiento de los datos proporcionados por corpus

En razón de las **tablas 1**, **2** y **3** anteriormente descritas, es decir, de las palabras diana (hipónimos de cada tríada polisémica) y de los hipónimos seleccionados para la

construcción de los vectores clase que corresponden a los tipos semánticos, es menester ahora extraer los materiales de corpus a utilizar. El corpus utilizado para este trabajo fue el esTenTen (Kilgarriff y Renau, 2013), y su tratamiento se realizó a través de Jaguar (Nazar, Robledo y Acosta, en línea), una herramienta virtual proporcionada para el análisis cuantitativo de corpus. A través de Jaguar se tuvo acceso al corpus previamente mencionado, además, etiquetado gramaticalmente y seleccionando la cantidad de cotexto requerido para el estudio de la forma léxica a estudiar. En nuestro caso, el cotexto inmediato lo forman las 20 unidades léxicas que rodean a la unidad de análisis (el sustantivo), es decir, 10 unidades por el cotexto izquierdo de la unidad y 10 por el derecho. Esta decisión obedece a la posibilidad de extraer las variables verbos que, en ocasiones, aparecen con cierta distancia en el eje sintagmático de los sustantivos. Cabe precisar que aunque las concordancias seleccionadas de manera aleatoria para cada forma léxica fueron 5.000 casos de corpus, en algunos casos la muestra era menor porque el corpus no proporcionaba más concordancias respecto de la unidad de estudio (el sustantivo).

Bajo esta descripción, las concordancias correspondientes a cada una de las palabras diana eran 5.000, mismo número de concordancias utilizadas para cada hipónimo que formaría el vector clase. De esta manera, y en razón de tener 10 hipónimos para la construcción de cada uno de los vectores clase (**tablas 2** y **3**), las concordancias utilizadas para la formación de cada vector suma son aproximadamente 50.000, cada una, por cierto, con un cotexto –como se ha dicho– de 20 unidades léxicas. Estas concordancias tienen la misma estructura que la **figura 3**, perteneciente a una de las concordancias de la forma léxica y palabra diana *guanaco* (hipónimo de *llama* y *animal*) extraída por medio de Jaguar desde el corpus esTenTen.

5	general/NC/general argentino/ADJ/argentino Juan/NP/Juan Manuel/NP/Manuel de/PREP/de Rosas/NP/Rosas tenía/ULfin/tener tan/ADU/tan mala/ADJ/malo hostia/NC/hostia porque/CSUBX/porque se/SE/se crió/ULfin/criar mamando/ULadj/mamando leche/ULfin/lechar de/PREP/de guanaca/NC/guanaca (/LP/(o/CC/o de/PREP/de	guanaco/NC/guanaco	,/CM/, las/ART/el fuentes/NC/fuente lo/PPC/él discuten/ULfin/discutir)/RP/) Este/DM/esto empleo/NC/empleo indiscriminado/ADJ/indiscriminado como/CSUBX/como fuerzas/NC/fuerza de/PREP/de choque/NC/choque debilitó/ULfin/debilitar considerablemente/ADU/considerablemente la/ART/el población/NC/población y/CC/y obligó/ULfin/obligar

Figura 4. Ejemplo de concordancia perteneciente a la palabra diana *guanaco*.

3.6.4 Selección de variables: los verbos

Considerando que se ha postulado que la coocurrencia verbal puede funcionar como elemento predictivo del significado de un sustantivo, haciendo que un término hipónimo se relacione con su hiperónimo correcto, será justamente este tipo de relación, vale decir, la de sustantivo-verbo, la encargada de ejecutar dicha elección predictiva. Por ello, de cada concordancia extraída por unidad léxica de análisis (ya sea palabra diana o hipónimo que forma un vector clase), pretendemos extraer las variables verbos que con ellas coocurren. En función de lo anterior, se construyó un *script* que leyera todos los archivos correspondientes a los casos a estudiar, y extrajera, de cada uno de los ficheros, los verbos que coocurren más de cinco veces con el sustantivo en cuestión. Estos verbos seleccionados por cada caso serán nuestras variables (V) para estudiar. La frecuencia de mayor a 5 coocurrencias verbales para la selección es arbitraria, pero fundada. Dada la muestra que utilizamos (5.000 concordancias seleccionadas aleatoriamente), se considera que es una alta frecuencia, cuya coocurrencia no es accidental o fruto de algún tipo de error.

En adelante, como se ha indicado recientemente, la metodología se implementará a través de scripts. Un script es una instrucción que utiliza una sintaxis comprensible para un computador o, en otras palabras, un lenguaje de programación. En nuestro caso, esta implementación la realizamos a través de Perl (Nazar, 2018).

Ahora bien, y volviendo a la operación de la selección de los verbos asociados a un sustantivo en más de cinco ocasiones de una muestra total de cinco mil concordancias, se ejecutó el siguiente script que muestra la **figura 5**. A este lo llamamos *agregar.pl*, principalmente por cumplir de función de anexar los verbos que buscábamos.

```
1  $filename = $ARGV[0];
2  open( $fh, '<', $filename );
3  while ( $line = <$fh> ) {
4       if ( $line =~ /\/V[^\/ \t]+\/([^\/ \t]+)/ ) {
5            $frec{$1}++;
6       }
7  }
8  close $fh;
9  foreach $w (keys %frec){
10      if ($frec{$w} > 5){
11           print "\n$w\t$frec{$w}";
```

```
12          }
13  }
```

Figura 5. Script agregar.pl

Lo que agregar.pl (**figura 5**) realiza es lo siguiente: primero, abrir y leer automáticamente el o los documentos entregados, en nuestro caso, las 5.000 concordancias por cada sustantivo o unidad de estudio en condición de hipónimo; segundo, se propone la condición de que si en una de las líneas se encuentre un verbo (dispuesto de acuerdo con una expresión regular en la línea 5 de script), pueda contar el número de veces que aparece el mismo elemento en ese documento (línea 7); y tercero, se entrega otra orden, que lo que hace es seleccionar, dentro de los verbos ya cuantificados, aquellos cuya frecuencia de coocurrencia sea superior a 5 veces en el mismo documento. La salida o el *output* de la ejecución de *agregar.pl* es una lista con los verbos que coocurren en más de cinco ocasiones durante todo el documento, sumado al número de veces de esta coocurrencia (línea 13).

Considerando que este y los siguientes scripts son ejecutados por medio de la terminal del computador, desde ahí es posible dirigir su salida por medio de la creación de un archivo o fichero de texto *.txt*, para poder almacenar los resultados de la orden. En el caso de nuestro trabajo, este script fue ejecutado 110 ocasiones, primero en las concordancias de las palabras diana, y luego en las concordancias de los 100 hipónimos que formaban los vectores clase. Por consiguiente, se obtuvo 110 ficheros que contenían los verbos que coocurrían más de 5 veces por cada 5.000 concordancias.

Una vez terminado este procedimiento, se unen los 110 ficheros en un solo gran fichero (en nuestro caso *todojunto.txt*) separados por tipo de sustantivos. Así, los verbos que pertenezcan a las palabras diana irán encabezado con el nombre de sustantivo (ejemplo: ‘tarántula-diana’), mientras que los verbos asociados a sustantivos hipónimos que formarán vectores clases (tipos semánticos) tendrán el encabezado del tipo semántico al que correspondan (ejemplo: ‘animal-canario’), tal y como se muestra en las **tablas 4** y **5** extraídas de ese fichero.

==> tarantula-diana.txt <==	
poner	9
tener	13
dar	8
estar	34
decir	9
poder	18
ser	57
hacer	8
volver	11
saber	7
haber	53
ir	8

Tabla 4. Frecuencia de los verbos asociados al sustantivo diana *tarántula.*

==> animal-canario.txt <==	
considerar	6
deber	8

aprobar	7
querer	6
celebrar	8
estar	20
realizar	12
leer	14
ser	123
decir	13
incluir	6
entender	6
referir	8

Tabla 5. Frecuencia de los verbos asociados al sustantivo hipónimo *canario*.

3.6.5 Formación de vectores clase y conformación de la matriz de estudio

Un vector es una información empírica y cuantificable sobre el comportamiento de un fenómeno; en el caso de la lingüística computacional, y en el particular caso de nuestra investigación, de una palabra. Los vectores que en nuestro trabajo pretendemos estudiar son los de coocurrencia verbal en cotextos sintagmáticos de sustantivos. En estos términos, cada comportamiento total de la coocurrencia sustantivo-verbos es un vector, y nosotros hasta ahora ya hemos extraído ese comportamiento a través del script *agregar.pl*.

Ahora bien, dado que la hipótesis distribucional indica que las palabras que coocurren en cotextos similares tienden a tener significados similares (Torres y Arco, 2016), nos hemos propuesto formar vectores clase que reúnan la información vectorial, es decir, la coocurrencia sustantivo-verbo de cada hipónimo propuesto para su formación. De esta

manera, un vector clase del tipo semántico e hiperónimo *animal* se formará a partir de la suma de la información vectorial (coocurrencia sustantivo-verbo) de todos los hipónimos propuestos para él en el **apartado 3.6.2** (ver **tabla 2**: *canario*, *gorrión*, *elefante*, *león*, etc.). Como ya se ha mencionado, esta construcción metodológica no es una idea nueva, sino solo la aplicación de una propuesta de Schütze y Pedersen (1997). Fueron ellos los que pensaron la creación de un vector suma (o clase) plausible de ser utilizado en trabajos de lingüística distribucional.

En nuestra investigación, los vectores clase correspondientes a cada tipo semántico se conforman por medio de la utilización de un script al que llamamos *matrix_vectorclase.pl*. El nombre obedece a que el vector clase se construye a partir de una matriz que ordena por cada fila de unidades de muestra (sustantivos hipónimos) la ocurrencia de las variables (verbos) en la superficie estudiada o la suma de los ficheros *.txt* particulares en el fichero *todojunto.txt*. De este modo, para asociar las variables con las unidades de muestra, la matriz creada obtiene un resultado sumatorio por cada clase de acuerdo con los rótulos que hemos colocado en el fichero *todojunto.txt*. En efecto, aquellas palabras que tienen el rótulo de uno de los tipos semánticos (‘animal’, ‘humano’, ‘máquina’, etc.) formarán un vector clase que contiene la suma de las ocurrencia en cada uno de los casos que lo conforman, mientras que las palabras que no tienen ese rótulo, y que corresponde a las *dianas*, formarán, cada una, un vector independiente.

La operación sumatoria que realiza el script *matrix_vectorclase.pl* y que da origen a cada vector clase funciona del modo que ejemplifica la **tabla 6**. Si el sustantivo (H) se combina con un verbo (V), el valor asignado será 1, y si no se vincula con un verbo, el valor asignado será 0.

	V_1	V_2	V_3	V_4	V_5	V_n
H_1	0	1	1	0	0	0
H_2	0	0	1	0	1	0
H_3	0	1	1	0	0	1
H_4	0	0	1	0	1	0
H_5	0	1	1	0	0	1
H_6	0	0	0	0	0	1
H_7	0	1	0	0	1	0
H_8	0	1	1	0	1	0
H_9	0	0	1	0	0	0
H_{10}	0	1	1	0	1	0
vector clase	**0**	**1**	**1**	**0**	**1**	**1**

Tabla 6. Ejemplificación de la operación realizada por el script *matrix_vectorclase.pl*

Y el script que realiza dicha operación es el que se presenta en la **figura 6**.

```
1  use strict;
2  my $filename = "todojunto.txt";
3  if (!$filename || ! -e $filename) {
4          die "\nFalta archivo input! \n";
5  }
6  my ($clase, $verb, %matrix, %verbos);
7   open( my $fh, '<', $filename );
8  while ( my $line = <$fh> ) {
9   if ( $line =~ /==> (.+)\-.+\.txt <==/ ) {
10                $clase = $1;
11   } elsif ($line =~ /(.+)\t/ ) {
12              $verb = $1;
```

```
13      if ($clase and $verb =~ /^[a-záéíóúüñ]+$/ and $verb ne "www" and
14  length($verb) < 15 ) {
15          $matrix{$clase}{$verb}++;
16          $verbos{$verb}++;
17       }
18     }
19  }
20  close $fh;
21  foreach $verb (sort keys %verbos) {
22     print "\t". $verb;
23  }
24  foreach $clase (sort keys %matrix){
25   print "\n".$clase;
26  }
27        foreach $verb (sort keys %verbos) {
28               my $resultado = "0";
29                 if ($matrix{$clase}{$verb}) {
30                      $resultado = "1";
31                   }
32                   print "\t".$resultado;
33         }
34  }
```

Figura 6. Script *matrix_vectorclase.pl*

El script precedente se alimenta del fichero *todojunto.txt* previamente descrito en este mismo apartado. En la línea 6 del script de la **figura 6** declaramos las variables que se utilizarán. Luego de leer el archivo, en la línea 7, distinguimos en la línea 9 y 10 cuáles serán las unidades clase (vectores diana y vectores clase) y en la línea 11 y 12 cuáles serán las variables asociadas a cada una. Posteriormente, limpiamos la muestra de aquellas unidades del corpus mal etiquetadas evitando seleccionar unidades que contengan caracteres que no son propios de verbos en su forma infinitiva (líneas 13 y 14), para luego generar la matriz. Una vez creada la matriz, se completan los resultados de coocurrencia entre sustantivos clase y variables (verbos) en valores binarios, es decir, con 0 en caso de que no exista y 1 en caso de que sí haya ocurrido. Por último, en la línea 32 se imprime el resultado de la matriz con las características que mostramos, a modo de ejemplo, en la **tabla 7.**

Vectores diana | clase	abandonar	abarcar	abogar	abordar	abrir
animal	0	0	0	0	0
artefacto	0	0	0	0	1
caniche	0	0	0	0	0
cencerro	0	0	0	0	0
chimpance	0	0	0	0	0
evento	1	1	0	0	1
guanaco	0	0	0	0	0
higado	0	0	0	0	0
humano	1	1	0	0	1
instrumentomusical	1	0	0	0	1
laucha	0	0	0	0	1
lugar	0	0	1	0	1
maquina	0	0	0	0	1
partedelcuerpohumano	0	0	0	0	1
periododetiempo	0	0	0	1	1

prenda	0	0	0	0	1
siames	0	0	0	0	0
tarantula	0	0	0	0	0
timbal	0	0	0	0	0
verano	0	0	0	0	1

Tabla 7. Parte de la matriz generada con el script *matrix_vectorclase.pl*. Contiene los vectores de las palabras diana y los vectores clase de los hiperónimos correspondientes.

3.6.6 Creación del script clasificador que desambigua un hipónimo entre dos o más hiperónimos posibles

Una vez dispuesta la matriz del paso anterior, matriz que para los alcances de esta investigación tiene diez casos (diana) y diez vectores clase (tipos semánticos o hiperónimos), pero que podría ser eventualmente replicable y extensible a todos los tipos semánticos de una taxonomía de poblamiento automático, nos falta ahora desarrollar el script clasificador en razón de la comparación de vectores, por cierto, entre los de tipo diana y los de tipo clase.

Este script, al que hemos llamado *clasificador.pl*, nos proporcionará el resultado final que buscamos conocer, es decir, saber si la comparación de los vectores diana y los vectores clase, que muestran la relación de los sustantivos con los verbos en cotextos inmediatos, es capaz de clasificar correctamente una relación hiperonímica en razón de rangos de similitud. A continuación, y luego de mostrar la construcción en la **figura 7**, explicaremos cómo funciona.

```
1  use strict;
2  my @clases = @ARGV;
3  my $paraeliminar = join ('|', @clases);
4  my $diana = shift @clases;
5  if (!$diana) {
6       die "\n No has entregado los argumentos";
7  }
8  my $filename = "matrixclasesfinal.csv";
```

```
9  my %compar;
10 open( my $fh, '<', $filename );
11 while ( my $line = <$fh> ) {
12       next if ( $line =~ /^\t/);
13       my @campos = split ( /\t/ , $line );
14       my $sust = shift @campos;
15       if ($sust =~ /($paraeliminar)/) {
16             $compar{$sust} = \@campos;
17       }
18 }
19 close $fh;
20
21 foreach my $clase (keys %compar) {
22       if ($clase eq $diana) {
23             next;
24       }
25       print "\nComparando [$diana] con [$clase]:";
26       my $n = scalar @{$compar{$clase}};
27       my $numerador;
28       my $denominador;
29       foreach my $i (0 .. $n) {
30          if ($compar{$clase}->[$i] and $compar{$diana}->[$i] )
31 {
32                $numerador++;
33          }
34          if ($compar{$clase}->[$i] or $compar{$diana}->[$i]) {
35                $denominador++;
36          }
37       }
38       my $res = $numerador / $denominador;
39       print "\nResultado: [$res]";
40 }
```

Figura 7. Script *clasificador.pl*

El script que llamamos *clasificador.pl* realiza la operación final clasificación que ejecuta la elección de un hiperónimo de segundo orden cuando el hiperónimo de primer orden de un sustantivo hipónimo es polisémico. La descripción de su función es la siguiente.

En primer lugar, desde la línea 2 a la 6 generamos un modo de ordenamiento que nos permite leer comparativamente los vectores diana con los vectores clase. En segundo lugar, desde la línea 8 a la 19 entregamos instrucciones de lectura del archivo que construimos en el paso anterior (**subsección 3.6.5**), es decir, la *matrixclasesfinal.csv*. Esta lectura exige guardar los comparandos (vector diana y vectores clase); recorrer la matriz línea por línea y rechazar la fila cabecera que tiene las variables verbos (línea 12); separar, por medio del comando *split*, las líneas por columnas y, por medio del comando *shift*, extraer el primer elemento que coincida, ya sea con un sustantivo diana o con cualquiera de los dos o más tipos semánticos hiperonímicos (clase). Posteriormente, ya en la línea 15, nos quedamos solo con el vector diana y vectores clase seleccionados. En este paso, se cargan, por cierto, los vectores de cada uno de acuerdo con los valores de coocurrencias de 0 y 1. De esta forma, se obtienen los datos requeridos para la comparación. Y en tercer lugar, se realiza la comparación propiamente tal. Para ello se ejecuta un bucle que realiza el proceso de comparación automática del vector correspondiente a la palabra diana con los vectores clase correspondientes a los hiperónimos de segundo orden a seleccionar. En esta operación se utiliza al índice de Jaccard (**fórmula 1**) para comparar vectores binarios. Este índice se ejecuta dividiendo la cantidad de veces en que un vector diana (d) y un vector clase (c) coinciden o intersectan ($\cap$) en la coocurrencia verbal, es decir, cuando la matriz asigna el valor 1, por la suma del número de coincidencias o intersecciones ($\cap$) más el número de ocasiones en que al menos una de estos dos casos (vector diana o vector clase) tiene una coocurrencia sin coincidir el uno con el otro, es decir, cuando solo uno de los dos presenta el valor 1. A esta última situación le llamamos unión entre los comparandos ($\cup$).

$$Jaccard\ (\overrightarrow{d}, \overrightarrow{c}) = \frac{\vec{d}\ \cap\ \vec{c}}{\vec{d}\ \cup\ \vec{c}}$$

Fórmula 1. Índice de Jaccard

Si al comparar el vector de palabra diana con dos o más vectores clase pertenecientes a los hiperónimos de segundo orden, el resultado de uno es superior al otro, entonces dicho

resultado deberá corresponder a aquel que tenga mayor grado de similitud en razón de la relación sustantivo-verbo y, por consiguiente, debería ser el hiperónimo correcto del hipónimo diana en cuestión.

4. Resultados

La aplicación de la propuesta metodológica de clasificación automática de hiperónimos de segundo orden para sustantivos hipónimos cuyo hiperónimo de primer orden era polisémico se empleó en diez casos de corpus que ya han sido detallados en la **tabla 1** del **apartado 3.6.1**.

La ejecución de esta metodología estadística basada en corpus por medio de herramientas proporcionadas por la lingüística computacional logró seleccionar correctamente el hiperónimo de segundo grado en los diez 10 casos propuestos para esta investigación. En este sentido, la precisión de la propuesta metodológica alcanza un 100% de aciertos de acuerdo con la muestra presentada inicialmente (**apartado 3.6.1**). Este es el fenómeno que se grafica en cada uno de los resultados de la **tabla 8**.

	sustantivo diana	**hiper 1**	**hiper 2**	**índice de similitud (Jaccard)**
1	**guanaco**	**llama**	**animal**	**0.0632**
			evento	0.0360
2	**caniche**	**perro**	**animal**	**0.1228**
			artefacto	0.0684
			humano	0.0480
3	**tarántula**	**araña**	**animal**	**0.0760**
			artefacto	0.0446
4	**siamés**	**gato**	**animal**	**0.0409**
			artefacto	0.0240

5	**laucha**	**ratón**	**animal**	**0.4182**
			artefacto	0.2888
6	**chimpancé**	**mono**	**animal**	**0.1942**
			prenda	0.1204
7	**timbal**	**tambor**	**instrumento musical**	**0.0472**
			artefacto	0.0412
8	**cencerro**	**campana**	**instrumento musical**	**0.0275**
			máquina	0.0270
9	**hígado**	**órgano**	**parte del cuerpo humano**	**0.3168**
			instrumento musical	0.3079
10	**verano**	**estación**	**período de tiempo**	**0.2757**
			lugar	0.2409

Tabla 8. *Resultados de la aplicación del índice de Jaccard entre vectores de sustantivos diana y vectores clase correspondientes a hiperónimos de segundo grado. En negrita, los casos en que el índice de Jaccard es superior y, al mismo tiempo, correcto.*

Los casos señalados en negrita muestran el resultado seleccionado por el script de la **figura 7**. Como se puede observar, la elección realizada por *clasificador.pl* es correcta en el 100% de los casos, lo que implica a su vez un precisión del 100%. De acuerdo con estos resultados es posible indicar que la comparación de coocurrencias verbales entre dos o más sustantivos de distinto nivel en una taxonomía es plausible de utilizarse para medir su grado de similitud semántica, y, por consiguiente, clasificar, de acuerdo con la mayor medida de similitud, un hiperónimo de segundo orden para un hipónimo cuyo hiperónimo de primer orden es polisémico.

En este sentido, existe una correspondencia de coocurrencias verbales entre sustantivos de distinto nivel de abstracción que se relacionan taxonómicamente (hipónimos e hiperónimos) que es medible y cuantificable. Esta correlación, bajo la aplicación del principio de distribución (Harris 1954; 1963), permite enlazar taxonómicamente un sustantivo hipónimo con su hiperónimo de segundo grado cuando su hiperónimo de primer grado es polisémico. Asimismo, y como consecuencia de ello, permite también describir su significado. Esta es la operación que se grafica a partir de la **figura 8** que presenta, a modo de ejemplo, un diagrama de Venn que muestra la intersección y similitud de variables de coocurrencias verbal entre tres objetos: uno hipónimo (*caniche*) y dos hiperónimos (*animal* y *artefacto*). Estos datos no pertenecen a una fuente externa, sino al análisis de datos de la matriz de estudio que se construyó con el script *matrix_vectorclase.pl* (**figura 6**).

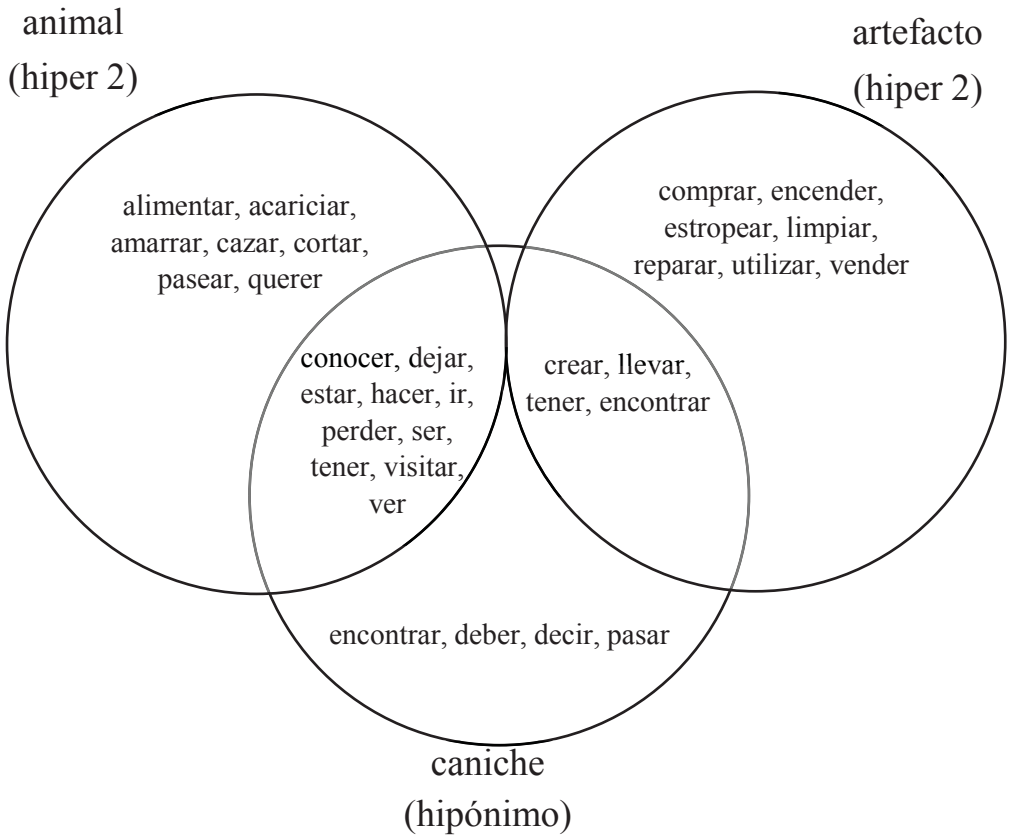

Figura 8. Diagrama de Venn que representa, a modo de ejemplo, las relaciones de intersección de coocurrencia verbal coincidentes entre un hipónimo (sustantivo diana) y sus posibles hiperónimos de segundo grado.

En síntesis, los resultados de esta aplicación metodológica logran superar, al menos en esta muestra, casos de polisemia léxica frecuentes en taxonomías de poblamiento automático.

En el futuro, su aplicación a gran escala, es decir, en todos los tipos semánticos que desarrolle un sistema de generación de taxonomía automática, acompañado de los algoritmos ya existentes, podría estrechar el margen de error de estas estructuras que hasta hoy fluctúa entre el 60% y 80% de precisión (Nazar y Renau, 2016).

5. Conclusiones y trabajo futuro

El presente trabajo de tesis abordó el problema de la asignación de un hiperónimo para una sustantivo polisémico en el marco de la formación de taxonomías automáticas. En razón de ello, se propuso un tratamiento metodológico del problema a partir de los fundamentos de la semántica léxica, los principios de la lingüística distribucional y las herramientas proporcionadas por la lingüística computacional.

En consideración a lo anterior y a los positivos resultados de la propuesta metodológica desarrollada (100% de precisión) que buscaba elegir de manera automática un hiperónimo en segundo grado para un sustantivo hipónimo cuyo hiperónimo de primer grado era un sustantivo polisémico, realizamos las siguientes consideraciones finales.

Primero, el principio de distribución de la lengua es funcional y aplicable desde una investigación con métodos de estadística de corpus. La aplicación de ambos posibilita el tratamiento del significado de unidades léxicas a partir de datos de una muestra amplia de la lengua y relaciones de similitud entre unidades (inclusión e hiperonimia).

Segundo, la similitud semántica explicitada en los lazos de hiperonimia (tipo de relación de inclusión) es plausible de ser medida considerando las unidades que coocurren con ellas habitualmente en los cotextos donde aparecen. Lo anterior, en razón de que existe una tendencia a que ciertas unidades coocurren más habitualmente con un tipo de elementos que con otros en el contexto sintagmático, como resulta que ciertos sustantivos coocurren más con ciertos verbos que con otros.

Tercero, y como profundización de lo anterior, la coocurrencia verbal entre sustantivos funciona como un elemento clasificador y predictor del significado en una relación taxonómica problemática, donde el sustantivo intermedio de una tríada taxonómica (hiperónimo de primer orden) es polisémico. En esta línea, la comparación del grado de similitud de estas coocurrencias verbales entre un sustantivo hipónimo y sus posibles hiperónimos de segundo grado permite la elección del hiperónimo correcto. Por el principio de herencia de las estructuras taxonómicas, se soluciona así el problema de polisemia de provocado por el hiperónimo de primer orden.

Cuarto, en razón de los positivos resultados de la muestra, la medición de la similitud de coocurrencias verbales entre unidades léxicas y su comparación posibilitará el mejoramiento de la producción de taxonomías léxicas automáticas que describan

semánticamente la lengua. Asimismo, el perfeccionamiento de estas, a su vez, podrá abordar, en un futuro próximo, de mejor manera el problema de la desambiguación léxica en el marco del procesamiento del lenguaje natural.

Quinto, como trabajo próximo se estima repetir el mismo método, pero ampliando la muestra, para probar si repite la alta y patente tasa de precisión de este experimento inicial de 10 casos.

Sexto, y para finalizar, el trabajo futuro, por ahora, se vincula a la aplicación de esta propuesta metodológica a una estructura taxonómica automática que presente problemas de polisemia y, seguidamente, evaluar el grado de mejora de esta, tal como el caso del proyecto de inducción taxonómica automática Kind (Nazar y Renau, 2016). En ese caso, formando los vectores clase de los tipos semánticos más abstractos a partir de los datos de corpus ya utilizados por el sistema de generación de la estructura, sobre todo, para automatizar más aún el proceso metodológico.

6. Referencias bibliográficas

Aristóteles. (1982). *Tratados de lógica*. (M. Candel Sanmartín, Trad.). Madrid: Gredos.

Aristóteles. (1994). *Metafísica*. (T. Calvo Martínez, Trad.). Madrid: Gredos.

Álvarez de Miranda, P. (2009). Neología y pérdida léxica. En De Miguel, E. (ed.), *Panorama de lexicología* (pp. 133-156). Barcelona: Ariel.

Baldinger, K. (1977). *Teoría semántica*. Madrid: Alcala.

Baroni, M. (2013). Composition in distributional semantics. *Language and Linguistics Compass*, 7 (10), 511-522.

Battaner, P. (2008). El fenómeno de la polisemia en la lexicografía actual: otra perspectiva. *Revista de Lexicografía*, 14, 7-25.

Baylon, Ch. y Fabre, P. (1994). *La semántica*. Barcelona: Paidós.

Berber, T. (2000). Lingüística de corpus: histórico e problemática. *DELTA*, *16* (2), 323-367.

Bloomfield, L. (1964). *Lenguaje*. (A. Flor Ada, Trad.). Lima: Universidad Nacional Mayor de San Marcos.

Bordea, G., Buitelaar, P., Faralli, S., and Navigli, R. (2015). Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). In *Proc. of the 9th International Workshop on Semantic Evaluation* (SemEval 2015), pp. 902 - 910. Denver, Colorado, June.

Cabré, M. T. (2006). La clasificación de neologismos: Una tarea compleja. *Alfa, 50* (2), 229-250.

Chodorow, M., Byrd, R. y Heidorn, G. (1985). Extracting semantic hierarchies from a large online dictionary. *Proceedings of the 23rd annual meeting on ACL*, pp. 138-143.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Escandell, M. V. (2007). *Apuntes de semántica léxica*. Madrid: UNED.

Espinal, M. T. y Mateu, J. (2014). Palabras y significado. En Espinal, M. T. (coord.), *Semántica* (pp. 59-109). Madrid: Akal.

Felíu, E. (2009). Palabras con estructura interna. En De Miguel, E. (ed.), *Panorama de lexicología* (pp. 51-81). Barcelona: Ariel.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Gadamer, H G. (1975[2015]). *Verdad y Método II*. Salamanca: Sígueme.

García, R, y Pascual, J. (2009). Relaciones de significado entre las palabras. En De Miguel, E. (ed.), *Panorama de lexicología* (pp. 117-131). Barcelona: Ariel.

Gelbukh, A. y Sidorov, G. (2010). *Procesamiento automático del español con enfoque en recursos léxicos grandes*. México, D. F: Instituto Politécnico Nacional.

Gómez, L. y Peronard, M. (2005). *El lenguaje humano*. Valparaíso: Ediciones Universitarias de Valparaíso.

Guiraud, P. (1976). *La semántica*. México D.F: Fondo de Cultura Económica, S. A.

Guthrie, L., Brian S., Yorik W. y Bruce, R. (1990). Is there content in empty heads? *Proceedings of the 13th International Conference on Computational Linguistics*, pp. 20-25.

Hanks, P. (2013). *Lexical Analysis. Norms and exploitations*. Cambridge, Massachusetts: MIT Press.

Hanks, P. (En línea). *CPA Ontology*. URL: http://pdev.org.uk/#onto (última consulta: 20/12/2018).

Harris, Z. (1954). Distributional Structure, *WORD*, *10* (2-3), pp. 146-162.

Harris, Z. (1963). *Structural linguistics*. Chicago: The University of Chicago Press.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th Conference on Computational Linguistics*, pp. 539-545.

Hernández, H. (1991). El análisis semántico-distribucional. Una aportación a la lexicografía. *Revista de Filología. Universidad de La Laguna*, *10*, 221-226.

Hjelmslev, L. (1943[1971]). *Prolegómenos a una teoría del lenguaje*. (J. Díaz de Liaño, Trad.). Madrid: Gredos.

Ide, N. & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics, 24* (1), 1-40.

Kilgarriff, A. y Renau, I. (2013). esTenTen, a vast web corpus of peninsular and american spanish. *Procedia - Social and Behavioral Sciences*, 95: 12-19. 5th International Conference on Corpus Linguistics (CILC2013).

Leech, G. (1974[1977]). *Semántica*. (J. L. Espada, Trad.). Madrid: Alianza Editorial.

Lenat, D. (1995). CYC: a large-scale investment in knowledge infrastructure, *Communications of the ACM*, 38 (11), pp. 33-38.

Lin, D. (1998). Automatic retrieval and clustering of similar words. *Proceedings of COLING'98*, pp. 768-774.

Lyons, J. (1995[1997]). *Semántica lingüística*. (S. Alcoba, Trad.). Barcelona: Paidós.

Malinowski, B. (1923). The problem of meaning in primitive languages. En Odgen, C. K & Richards, I. A (Eds.). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism* (pp. 296-336). Cambridge: Cambridge University Press.

Marín, R. (2009). El tratamiento computacional del léxico y sus aplicaciones. En De Miguel, E. (ed.), *Panorama de lexicología* (pp. 467-486). Barcelona: Ariel.

Martinet, A. (1974). *Elementos de lingüística general*. Madrid: Gredos.

Martí, A. (2018). Modelos de semántica distribucional. Actas del *XIII Congreso de Lingüística Xeral*. Vigo, pp. 16-22

Mateu, J. (2014). Predicación. En Espinal, M. T. (coord.), *Semántica* (pp. 185-221). Madrid: Akal.

Meillet, A. (1952). *Linguistique historique et linguistique générale*. París: Klincksieck.

Mounin, G. (1977). *Lingüística del siglo XX*. Madrid: Gredos.

Nazar, R. (2013). Word Sense Discrimination Using Statistic Analysis of Texts. *BRAC. Barcelona, Research, Art, Creation*, *1* (1), 5-26. doi: 10.4471/brac.2013.01.

Nazar, R. y Renau, I. (mayo, 2016). A taxonomy of Spanish nouns, a statistical algorithm to generate it and its implementation in open source code. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA), pp. 1485-1492.

Nazar, R. (2018). Taller de procesamiento de textos con Perl: una mini-introducción. Pontificia Universidad Católica de Valparaíso, Chile. Recuperado de: http://www.tecling.com/brasil2018/ResumenTallerPerl2018.pdf

Nazar, R, Robledo, H. Acosta, N. (En línea). *Jaguar: a Tool for Quantitative Corpus Analysis.* URL: http://www.tecling.com/cgi-bin/jaguar/jaguar.pl (última consulta: 15/12/2018).

Piera, C. (2009). Una idea de la palabra. En De Miguel, E. (ed.), *Panorama de lexicología* (pp. 25-49). Barcelona: Ariel.

Real Academia Española. (2014). *Diccionario de la lengua española* (23ª ed.). URL: https://dle.rae.es/?w=diccionario (última consulta: 20/12/2018).

Rojo, G. (agosto, 2002). Sobre la Lingüística basada en el análisis de corpus. *Hizkuntza-corpusak.* Oraina eta geroa.

Rojo, G. (2008). Lingüística de corpus y lingüística del español. *XV Congreso de la ALFAL*. Ponencia plenaria XV Congreso de la ALFAL. Montevideo.

Sahlgren, M. (2008). The distributional hypothesis. *Rivisti di Linguistica*, *20* (1), 33-53.

Schütze, H. y Pedersen, J. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, *33* (3), 307 -318.

Snow, R., Jurafsky, D. y Ng. A. (2006). Semantic taxonomy induction from heterogenous evidence. *Proceedings of the 21st International Conference on Computational Linguistics*, pp. 1-8.

Sowa, J. (2000). Knowledge representation logical, philosophical, and computational foundations. Pacific Grove Brooks/Cole cop.

Stearns, M. Q., Price, C. Spackman, K. A., Wang, A. Y. (2001). Snomed clinical terms: overview of the development process and project status. *Proceedings of the AMIA Symposium. American Medical Informatics Association*, pp. 662 - 666.

Torres, C. y Arco, L. (2016). Representación textual en espacios vectoriales semánticos. *Revista Cubana de Ciencias Informáticas*, 10 (2), 148 -180.

Trujillo, R. (1976). *Elementos de semántica lingüística*. Madrid: Cátedra.

Ullman, S. (1962[1972]). *Semántica*. (J. M. Ruiz-Werner, Trad.). Madrid: Aguilar.

Ullman, S. (1973[1979]). *Significado y estilo*. (J. García-Puente, Trad.). Madrid: Aguilar.

Valdés, L (ed). (1991). *La búsqueda del significado*. Madrid: Tecnos.

Villar, M. B. (2009). Modelos estructurales. En De Miguel, E. (ed.), *Panorama de la lexicología* (pp. 219-246). Barcelona: Ariel.

Villayandre, M. (2008). Lingüística con corpus. *E. H. Filología*, 30, pp. 229-249.

Vossen, P. (2004). Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. International Journal of Lexicography, 17 (2). pp. 161-173.

Wittgenstein, L. (1953[2009]). *Tractatus lógico - philosophicus. Investigaciones filosóficas sobre la certeza*. Madrid: Gredos.

Yallop, C. (2004). Words and meaning. En *Lexicology and Corpus Linguistics* (pp. 23-72). London: Continuum.