Artículo publicado en las actas del V Congreso Internacional de Lexicografía Hispánica, Madrid, 25-27 de junio de 2012.

Agrupación semántica de sustantivos basada en similitud distribucional. Implicaciones lexicográficas

ROGELIO NAZAR, IRENE RENAU
Instituto Universitario de Lingüística Aplicada
Universitat Pompeu Fabra
C/Roc Boronat 138, 08018 Barcelona

0. INTRODUCCIÓN Y OBJETIVOS

Este artículo[1] describe un experimento de agrupación semántica de sustantivos basada en la hipótesis distribucional de Harris (1954), según la cual la similitud semántica entre unidades léxicas puede detectarse a través de la búsqueda de coincidencias en el contexto lingüístico. Nuestro estudio busca demostrar que, dado un conjunto de sustantivos con distintos hiperónimos (en este caso, bebidas, quesos, sombreros, vehículos y animales), es posible su agrupación bajo el hiperónimo correcto por procedimientos estadísticos que prácticamente no requieren conocimiento explícito de la lengua analizada.

Para ilustrar el argumento, la Tabla 1 muestra el ejemplo de tres sustantivos en castellano, *cerveza*, *café* y *té*, con sus respectivos elementos coocurrentes, que son palabras que muestran una tendencia a aparecer con frecuencia en el corpus junto a cada unidad en cuestión. Esto es, es muy frecuente que palabras como *beber*, *tomar*, *servir*, etc., aparezcan con frecuencia en combinación con estas tres unidades léxicas porque son bebidas. Tal como la tabla indica, las unidades que coocurren con frecuencia con cada unidad analizada contienen importante información semántica sobre ellas, la cual nos permite compararlas entre sí y agruparlas en función de la cantidad de atributos compartidos. En verdad se trata de atributos en muchos de los casos, como el color, el sabor o la temperatura. Por un lado, estos atributos nos permiten ver lo que las tres unidades tienen en común, ya que comparten el vocabulario que acompaña típicamente a las bebidas y que cabría en la intersección en la Figura 1 entre los tres conjuntos analizados; pero por otro lado también nos permitiría continuar haciendo distinciones dentro del grupo de las bebidas, ya que, por ejemplo, *té* y *café* tienen algunas unidades en común que las caracterizan como bebidas calientes que se beben

1 Este trabajo ha sido posible gracias a la financiación de los proyectos “Agrupacion semántica y relaciones lexicológicas en el diccionario”, dirigido por J. DeCesaris (HUM2009-07588/FILO) y por el proyecto APLE: “Procesos de actualización del léxico del español a partir de la prensa escrita”, dirigido por M.T. Cabre (FFI2009-12188-C05-01/FILO).

en taza (con coocurrentes como *caliente*, *preparar*, *taza*, *cucharada*, etc.) que como cabe esperar no aparecen con *cerveza*. En menor medida existen también atributos compartidos entre *cerveza* y *café*, como el amargor, o entre la *cerveza* y el *té*, como una temperatura inadecuada para el consumo (*tibia*/*o*).

cerveza	**café**	**té**
beber tomar pedir querer servir	beber tomar pedir querer servir	beber tomar pedir querer servir
amarga	amargo	
tibia		tibio
	caliente negro preparar taza plantación cucharada	caliente negro preparar taza plantación cucharada
fría fresca local embotellar botella lata jarra vaso	triste irlandés humeante gustar máquina poso	verde aromático bolsa hoja infusión

Tabla 1: Comparación de palabras que coocurren con tres sustantivos

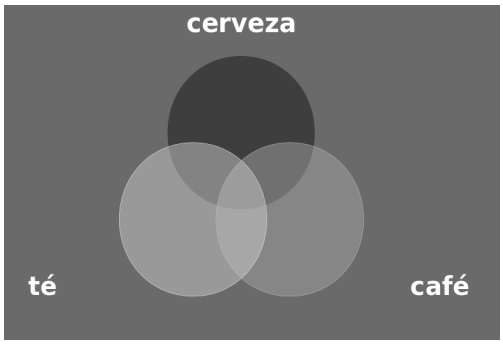

Figura 1: representación de la intersección de unidades léxicas como conjuntos de unidades coocurrentes

Sin lugar a dudas el estudio de la similitud distribucional de las unidades léxicas puede dar pie a una serie de investigaciones distintas. Sin embargo, en el presente artículo nos limitamos a una primera exploración de la posibilidad de agrupar unidades mediante la técnica estadística de clustering, tomando como función de similitud la cantidad de palabras coocurrentes que los comparandos tienen en común. Para ello adoptamos un enfoque técnico sencillo y más bien

conservador, dejando para trabajo futuro una mayor elaboración de los algoritmos utilizados y una mayor experimentación con la modificación de distintas variables, tal como se explica en mayor detalle en la sección de trabajo futuro.

La estructura del artículo es la siguiente. Después de esta introducción, el artículo se divide en las restantes cinco secciones que abordan respectivamente los siguientes aspectos: en la sección 1 ofrecemos una descripción del contexto de esta investigación y explicamos por qué es relevante para la lexicografía; en la sección 2 comentamos el trabajo relacionado con el proyecto; en la sección 3 explicamos nuestra metodología cuyos resultados se detallan en la sección 4 y, finalmente, en la sección 5 exponemos nuestras conclusiones y posibles líneas de trabajo futuro.

1. CONTEXTO E IMPLICACIONES LEXICOGRÁFICAS

1.1. Motivación

La motivación de este trabajo está en el desarrollo de una metodología para la generación de taxonomías como una pieza más general de un proyecto de estudio de las relaciones predicado-argumento. En un trabajo anterior (Renau y Nazar 2011) describimos una metodología de análisis de corpus basado en el Corpus Pattern Analysis (CPA) de Hanks (2004) con el cual pretendemos determinar las clases de sustantivos que en castellano pueden operar como argumentos de cada verbo (para una presentación y adaptación de CPA al castellano con fines lexicográficos, *cf.* Renau y Alonso, en prensa, y Renau 2012).

El análisis de textos reales se ve favorecido con la detección automática de hiperónimos ya que esto permite aumentar la capacidad de generalización a partir del corpus. Para mencionar un ejemplo, considérese el verbo *abrir* en dos de sus acepciones: una, ‘quitar lo que cubre o tapa un recipiente u objeto similar’ (abrir una caja, lata, botella, etc.), y otra, ‘inaugurar un negocio’ (abrir una tienda, empresa, cadena, etc.) (Renau, 2012). La estructura de ambos usos es la misma, se trata de verbos transitivos, pero el significado cambia pues está ligado a la combinatoria. Nuestra idea entonces es sustentar este tipo de distinción en los usos de los verbos por medio del análisis estadístico de grandes cantidades de texto. Para poder llevar a cabo tal estudio, se necesita un sistema capaz de relacionar las entidades que aparecen en el corpus analizado con la o las categorías a las que estas entidades pertenecen. En el caso de nombres propios de persona, estos deberían ser reconocidos con la etiqueta semántica

PERSONA, y de la misma forma, el resto de las unidades que pueden funcionar como argumentos de un verbo tienen que ser asociadas a su o sus categorías generales más próximas (sus hiperónimos), como la palabra *tienda* con respecto a un tipo semántico ESTABLECIMIENTO o la palabra *caja* con respecto a OBJETO. Es evidente que se encuentra aquí una diversidad de problemas, entre ellos el de la polisemia porque *tienda*, *establecimiento* o *caja* pueden pueden tener significados distintos. Disponemos, sin embargo, de algoritmos para la inducción de los sentidos de unidades polisémicas con una fiabilidad superior al 90% (Nazar, 2010). Más complejo resulta plantearse qué nivel de riqueza y granularidad debe alcanzar la taxonomía y si debe contener unidades poliléxicas, porque debe encontrarse un equilibrio entre la arbitrariedad de la clasificación hecha por quienes desarrollan la taxonomía y las clases que se generan automáticamente con los datos del corpus. Por ejemplo, uno debe decidir si incluir o no un nodo como ESTABLECIMIENTO COMERCIAL para distinguirlo de ESTABLECIMIENTO u OBJETO QUE SE ABRE para ser más específico respecto a OBJETO, decisiones que a nuestro juicio deben tomarse en el ámbito de la lexicografía o donde sea que se aplique tal taxonomía.

Para un proyecto de este tipo se necesita una taxonomía del castellano muy completa o, mejor aún, un sistema capaz de asignar el hiperónimo más probable para a una determinada unidad léxica potencialmente desconocida. Esto es entonces lo que motiva nuestros trabajos en relación con la detección de similitud semántica y extracción de taxonomías.

1.2. Aplicaciones

El experimento de agrupación de unidades semánticamente similares interesa a la lexicografía en varios sentidos. En primer lugar, contribuye al análisis semántico de las unidades léxicas partiendo de datos de corpus. La observación de datos empíricos se considera una condición *sine qua non* para la realización de diccionarios (Sinclair 1987, 1991), y son muchos los investigadores que han reclamado una lexicografía sistemática (Apresjan, 2008). No obstante, el análisis de dichos datos es costoso en tiempo y esfuerzo, lo que conlleva que los diccionarios no siempre se hayan confeccionado basándose en dicho análisis o que este se haya hecho de forma poco metódica. Se están comenzando a desarrollar, con más empuje desde la década de 1990, herramientas que faciliten esta tarea, entre ellas la detección de similitud semántica en una línea inspirada en el trabajo de Zellig Harris (Grefenstette, 1994; Lin, 1998).

Así pues, las utilidades que se pueden derivar de este estudio están relacionadas con el uso que se le dé para agrupar los lemas y definirlos acorde con sus rasgos semánticos, o con poder ofrecer esta agrupación al usuario en un diccionario electrónico. La relevancia del experimento específicamente para la lexicografía viene dada por posibles aplicaciones como:

1) facilitar y mejorar el análisis semántico basado en corpus, pues la agrupación de concordancias por similitud semántica de sus argumentos facilita y acelera el lento y arduo proceso manual de tratamiento de los datos, uno de los principales escollos de tipo práctico con los que se ha topado la lexicografía de corpus
2) la creación, ampliación o actualización del lemario, que puede ser agrupado con criterio semántico y, de este modo, ser tratado de manera más homogénea, en correspondencia con el lemario ya disponible si es el caso
3) la agrupación de palabras para una redacción más coherente, pues se pueden crear grupos de objetos similares y definirlos con el mismo patrón de definición (p. ej. definir los quesos o los sombreros de la misma manera), lo que hasta ahora se venía haciendo con cierta ayuda de procedimientos informáticos pero insuficientes
4) permitir una consulta por rasgos semánticos al usuario en el caso de un diccionario web.

1.3. El contexto de la investigación

Querríamos referirnos brevemente a otros trabajos sobre extracción de taxonomías en los que estamos trabajando en paralelo y que esperamos poder integrar en un sistema final, tal como anunciábamos en el apartado 1.1. En general se trata de estrategias sencillas e independientes de lengua que si bien aún no ofrecen resultados de calidad suficiente para un usuario final cuando se aplican de forma aislada, su combinación, que dejamos para trabajo futuro, sí parece prometedora ya que esto permitiría aumentar la cantidad y certeza de la información contenida.

Nuestro primer enfoque fue la extracción de relaciones de hiperonimia a partir de múltiples diccionarios electrónicos utilizando estadísticas de coocurrencia entre las palabras del *definiens* y las del *definiendum*, siguiendo la intuición de que una las palabras más frecuentes en un conjunto de definiciones de una misma unidad léxica suele ser su hiperónimo

(Renau y Nazar, 2012). También investigamos la posibilidad de extraer las relaciones de hiperonimia a través de la asociación de coocurrencia asimétrica de las palabras en corpus (Nazar y Renau, 2012). De esta forma encontramos por ejemplo que palabras como *ciclomotor* o *bicicleta* tienen tendencia a aparecer con frecuencia en oraciones con la palabra *vehículo* y que esta relación no es recíproca. Las palabras tienen tendencia a aparecer en las mismas oraciones pero la asociación es más fuerte en un sentido, es decir que tiene una dirección, y esto permite construir un grafo dirigido que se asemeja a una taxonomía "natural", ya que surge directamente del texto y no a través de la introspección.

Finalmente, estamos llevando a cabo también un experimento de agrupación de unidades léxicas (no sólo sustantivos) de acuerdo con su similitud paradigmática en un sentido estricto, que es calculada como la probabilidad de dos palabras de aparecer exactamente en los mismos fragmentos de texto tales como secuencias de entre tres y cuatro palabras (Nazar y Renau, en preparación). Así, una secuencia en inglés como "are the ways of the Lord" puede ser precedida por una serie limitada de palabras, tales como *inscrutable*, *strange*, *mysterious*, *infinite*, *unsearchable*, etc. Este mecanismo sirve para agrupar palabras en clases semánticas y morfológicas, pero en cualquier caso se puede decir que se trata de unidades de comportamiento sintáctico similar. Por eso, esta última línea de investigación promete resultados no solo en relación a la agrupación de unidades semánticamente similares sino también como una nueva vía para la inducción automática de sintaxis y su aplicación al etiquetado morfosintáctico.

2. TRABAJO RELACIONADO

El campo de la extracción automática de relaciones semánticas (con particular énfasis en la extracción de relaciones de hiperonimia) comenzó a desarrollarse a partir de la publicación de los primeros diccionarios en formato electrónico en la década del setenta. Este nuevo recurso favoreció la aparición de una serie de publicaciones sobre distintos métodos para convertir el diccionario pensado para el usuario humano en una base de datos con información semántica accesible para el ordenador (Chodorow et al., 1985; Alshawi, 1989; Fox et al., 1988; Nakamura y Nagao, 1988; Wilks et al, 1989; Guthrie et al, 1990; Boguraev, 1991; entre muchos otros). Estos autores comparten la idea de tomar un diccionario electrónico y estudiar las regularidades en las definiciones que permitan la extracción de relaciones semánticas mediante sistemas de reglas. Dependiendo del diccionario, una de estas reglas podría ser que

el primer sustantivo en la definición de un sustantivo es su hiperónimo. El problema está, naturalmente, en que no siempre es el caso, y ello requiere entonces de nuevas reglas que salven las excepciones.

Con el boom de la lingüística de corpus en la década de los noventa, el interés de los investigadores se trasladó de la extracción de relaciones semánticas a partir de diccionarios hacia la extracción a partir de corpus, aunque con una filosofía similar ya que lo que se busca en corpus en este caso son patrones léxico-sintácticos tales como "X es un tipo de Y", donde la relación de hiperonimia entre las unidades X e Y encontradas en corpus se expresa de una manera más o menos explícita. Esta línea tiene muchos partidarios particularmente en el campo de la terminología especializada (Hearst, 1992; Barrière y Popowich, 1996; Pearson, 1998; Meyer, 2001; Rydin, 2002; Potrich y Pianta, 2008; Auger y Barrière, 2008; Aussenac-Gilles y Jacques, 2008, entre diversos otros). Algunos de los inconvenientes de este enfoque es que no siempre los patrones recolectados expresan las relaciones esperadas, que las relaciones pueden aparecer en patrones que quienes desarrollan el sistema de extracción no son capaces de prever y, sobre todo, que todavía no se ha encontrado una manera de combinar eficientemente la información que se encuentra en el corpus. Un ejemplo de ello es que en general no se tiene en cuenta como indicador de veracidad de la supuesta relación de hiperonimia la cantidad de veces en que un mismo par de entidades X e Y aparece en corpus instanciando estos patrones léxico-sintácticos, una información que con seguridad contribuiría a aumentar la calidad de los resultados.

Un enfoque diferente y cuantitativo por naturaleza es el de los autores que, también a partir de corpus, buscan encontrar unidades léxicas que son semánticamente similares, ya sea por sinonimia o equivalencia como por pertenencia a una misma clase semántica y, en consecuencia, gobernada por un mismo hiperónimo (Grefenstette, 1994; Schütze y Pedersen, 1997; Lin, 1998; Alfonseca y Manandhar, 2002; Kilgarriff et al, 2004; Curran, 2004; Pekar et al. 2004; Bullinaria, 2008; entre muchos otros). Estos trabajos parten también de la línea iniciada por Harris (1954) y en general hablan de extracción de tesauros más que de taxonomías. Una mención aparte merece la aplicación de la misma estrategia para la extracción de unidades léxicas equivalentes en distintas lenguas, lo cual añade un grado más de complejidad al problema (Fung y McKeown, 1997; Rapp, 1999; Nazar, 2010).

Nuestro enfoque en el presente artículo se acerca más a este tercer grupo de autores mencionado, pero probablemente sea la primera vez que se intenta hacer algo así en

castellano, ya que no hemos sido capaces de encontrar trabajos similares en esta lengua. Naturalmente, no hay razones para esperar que con respecto a los estudios de semántica distribucional el castellano se comporte de una manera muy distinta al inglés.

3. METODOLOGÍA

3.1. Materiales

En este experimento intentamos agrupar palabras en categorías semánticas utilizando una medida de similitud paradigmática, es decir, una comparación de los contextos de ocurrencia de estas palabras. Para ello, tomamos como contexto una sola palabra a la derecha y a la izquierda de la unidad analizada en el corpus de enegramas de Google Books (Michel et al. 2011), cuyo tamaño es de alrededor de 45.000 millones de palabras, la más grande colección de castellano escrito en formato digital existente en la actualidad.

Para este experimento se seleccionaron arbitrariamente 145 sustantivos que, en su significado más frecuente, tuvieran como hiperónimo principal las palabras *bebida*, *queso*, *sombrero*, *vehículo* y *animal*. Tales son unidades como *ajenjo*, *brandy*, *horchata*, *limonada* o *ponche*, como bebidas, *gorgonzola*, *gruyer*, *mozzarella o roquefort,* como quesos, etc. Naturalmente, esto no implica que el algoritmo exija un número reducido de unidades para analizar ni que estas pertenezcan a un número reducido de clases semánticas. El motivo de este diseño experimental es facilitar la evaluación, pero damos por supuesto que la aplicación del método es general a cualquier muestra.

3.2. Extracción de datos

Se buscaron en el corpus los bigramas (secuencias de dos palabras) en los que aparecen los sustantivos de la muestra. En el caso de la palabra *brandy*, por ejemplo, encontramos en el corpus bigramas como *mucho brandy*, *buen brandy*, *brandy español*, *brandy francés*, *bebiendo brandy*, *tomar brandy*, *brandy barato*, etc.

3.3. Filtrado de los datos

Las unidades más comunes y que por tanto resultan escasamente informativas (palabras como *mucho, buen, como,* etc) se eliminan del análisis por medio del coeficiente *w* definido en la ecuación 1, donde f_o es la frecuencia observada y f_e la frecuencia esperada. Esta última representa la probabilidad de una palabra de aparecer en un contexto cualquiera, y se calcula

registrando la frecuencia de tal unidad en un corpus de referencia que se supone representa una muestra de lengua general. El proceso de filtrado se lleva a cabo eliminado aquellas unidades que obtienen una puntuación *w* por debajo de un umbral arbitrario.

$$w(i) = \frac{f_o(i)}{(f_e(i) + 1)} \qquad 1$$

3.4. Construcción de los vectores

Con las unidades restantes, pasamos a construir una estructura de datos en la que cada sustantivo queda asociado a una lista de palabras (lematizadas) con las que comparte bigramas, la cual representamos como un conjunto o vector:

brandy = { *beber*, *francés*, *tomar*, *barato*, *presidente*… }

3.5. Similitud entre vectores

Utilizamos estos vectores para comparar los sustantivos entre sí y agrupar aquellos que resultan más similares, calculando esta similitud como la cantidad de palabras que tienen en común. Como dijimos en la introducción en el caso de las bebidas (Tabla 1), algunos quesos suelen cooccurrir con palabras como *rallar*, algunos animales con *feroz*, etc. La comparación entre vectores puede hacerse utilizando alguna de las diversas medidas de similitud que existen, como por ejemplo los coeficientes de Jaccard, Dice y overlap (o solapamiento), definidas en las ecuaciones 2, 3 y 4, respectivamente.

$$\text{Jaccard(X, Y)} = \frac{|X \cap Y|}{|X \cup Y|} \qquad 2$$

$$\text{Dice(X, Y)} = \frac{2|X \cap Y|}{|X| + |Y|} \qquad 3$$

$$\text{Overlap(X, Y)} = \frac{|X \cap Y|}{min(|X|, |Y|)} \qquad 4$$

Comprobamos empíricamente que los mejores resultados se obtienen con el coeficiente de Jaccard, aunque una confirmación concluyente se alcanzaría con una investigación sistemática con distintos tipos de coeficientes que dejamos para trabajo futuro. Es importante advertir que, debido a la forma en que se ha implementado este coeficiente, la comparación admite sólo valores binarios (es decir presencia o ausencia de los elementos coocurrentes) sin tener en cuenta la frecuencia de aparición de estos elementos. Esto se hizo bajo el supuesto de que las unidades que aparecen con una frecuencia estadísticamente no significativa ya han sido

eliminadas por el filtro de la sección 3.3., sin embargo no se puede obviar el hecho de que la frecuencia todavía tiene un papel importante a la hora de ponderar los atributos cuando se comparan los vectores. Sin embargo, también en este caso tenemos que dejar para trabajo futuro la implementación de medidas más eficientes que las basadas en valores binarios.

3.6. Clustering

El proceso de clustering requiere la comparación de todas las unidades entre sí utilizando el coeficiente de similitud elegido, para lo cual necesitamos una tabla de distancias tal como la que se ejemplifica en la Tabla 2. Para facilitar la explicación, supongamos que son solo cuatro las unidades que se someten al clustering, a las que llamaremos *a*, *b*, *c* y *d*. Dispuestas en las filas y en las columnas de la tabla, cada celda especifica el valor obtenido en la comparación de las dos unidades. Como la tabla es simétrica (es decir, obtenemos el mismo valor comparando *a* con *b* que *b* con *a*), sólo usamos la mitad superior a la diagonal principal de la tabla. Al ser 145 o cualquier otra cantidad de unidades, la forma de la tabla no cambia, simplemente contiene más celdas y columnas. El límite en la cantidad de unidades para analizar está en la capacidad tecnológica y viene dado por la complejidad cuadrática de la tabla (a un aumento lineal de unidades corresponde un aumento exponencial del coste computacional).

	b	**c**	**d**
a	...	...	...
b		...	...
c			...

Tabla 2: Tabla de distancias entre las unidades *a*, *b*, *c* y *d*

El proceso de clustering comienza con la elaboración de la tabla de distancias y de ella se obtiene el par de unidades que muestran mayor similitud. A continuación, los miembros de este par se funden en lo que sería un primer cluster, que pasa a ocupar el lugar de las dos unidades seleccionadas y a contener la suma de los atributos de ambas. El proceso es iterativo, es decir que de nuevo se vuelve a producir una nueva tabla de distancias aunque cada vez con un elemento menos. Este proceso puede detenerse cuando ya no hay más unidades que agrupar o bien según se especifique mediante distintos parámetros tales como un número máximo de clusters o un umbral de similitud que represente la cantidad mínima de atributos en común que tienen que tener dos unidades para formar parte de un mismo cluster.

4. RESULTADOS

Hacemos una descripción de los resultados en dos momentos del proceso de clustering. En una primera medición, interrumpimos el proceso por la mitad y tomamos nota de las agrupaciones que el sistema llevaba hechas hasta el momento, y luego dejamos que el proceso continuara para analizar el resultado final, cuando el proceso acaba por no poder hacer más agrupaciones.

Cuando pausamos el proceso para la primera medición, el algoritmo había creado 28 grupos, clasificando correctamente la mayoría de los sustantivos. De los 145 iniciales, en esta primera agrupación se clasificaron correctamente 75 sustantivos y solo 3 erróneamente, lo cual representa una precisión de 96% y una cobertura de 51%. En la segunda medición, cuando el proceso había concluido, los 28 grupos iniciales quedaron reducidos a 5 que, tal como se muestra en la Tabla 3, coinciden justamente con los grupos iniciales: *vehículos*, *quesos*, *bebidas*, *sombreros* y *animales*.

Cluster	**Miembros del Cluster**
1	*carro, automóvil, coche, autobús, tranvía, carroza, carruaje, camión, jeep, camioneta*
2	*brie, parmesano, camembert, mozzarella, gorgonzola, roquefort, gruyer*
3	*chocolate, licor, chicha, cerveza, aguardiente*
4	*pavero, tricornio, bicornio, guarapón, canotier, calañés*
5	*venado, ciervo, tigre, elefante, perro, gato, puerco, cerdo, carnero, conejo, ratón, rata*

Tabla 3: Resultado final con seis agrupaciones de unidades.

En el resultado final no hay elementos mal clasificados (100% de precisión), pero en contrapartida la cobertura ha caído significativamente ya que son solo 40 los elementos que se consiguen clasificar (27,5% de cobertura).

5. CONCLUSIONES Y TRABAJO FUTURO

Con los datos obtenidos en este artículo hemos demostrado que la similitud semántica entre las unidades léxicas se ve reflejada en las similitudes de sus contextos de uso. Los porcentajes de cobertura obtenidos todavía no son lo suficientemente elevados para cumplir objetivos prácticos como el desarrollo de una taxonomía. Sin embargo, y tal como hemos explicado en la sección 1.3., esta taxonomía final deberá ser el resultado de una combinación de estrategias distintas que queda para trabajo futuro.

Además de la integración de estrategias, en un próximo artículo afinaremos también la expuesta en este artículo. Estamos explorando distintas alternativas para elevar el porcentaje de cobertura buscando un equilibrio óptimo con la precisión. Este equilibrio se encontrará por medio de la comprobación empírica ajustando los parámetros de distinta manera y evaluando con distintas muestras de unidades léxicas. Una forma de atacar el problema de la baja cobertura es ajustar parámetros para que el algoritmo sea más "permisivo" a la hora de juzgar que dos unidades son similares (es decir, relajando el umbral de similitud) aunque esto implica que incurrirá en errores con mayor frecuencia.

Más allá del dato de la cobertura, consideramos que se trata de un resultado prometedor sobre todo si se tiene en cuenta que el algoritmo no sabe cuántas ni cuáles son las clases semánticas. Aquí no tenemos un típico escenario de clasificación en el que se dispone previamente de una serie de clases semánticas a las que asignar los hipónimos (como en Alfonseca y Manandhar, 2002 o Ciaramita, 2002) sino que las clases son también el producto de la clasificación. Es por este motivo que creemos que en esta etapa del proceso es más importante la precisión que la cobertura, porque es conveniente tener primero un número reducido de clusters en los que se tenga un grado mínimo de confianza, como nuestros cinco grupos en este caso, para comenzar entonces a distribuir en estas clases los elementos que quedaron sin clasificar, procedimiento complejo que queda pendiente explorar porque, naturalmente, siempre debe reservarse la posibilidad de crear un nuevo grupo si una unidad no se ajustara a ninguno de los que ya se han creado.

Como trabajo futuro tiene importancia también la tarea de asignar un nombre a los clusters generados. Esto es un aspecto que en este artículo no hemos evaluado sistemáticamente, sin embargo se observa que en los clusters que se han generado, el atributo que aparece con más frecuencia suele ser el nombre correcto del cluster. Es decir, por ejemplo, que en el cluster que corresponde a los quesos, la palabra que coocurre con más miembros de ese cluster es justamente la palabra *queso*, etc. Por lo tanto, creemos que es conveniente separar las distintas etapas del proceso. Ahora nos hemos concentrado en la agrupación de unidades, pero en una etapa posterior nos concentraremos en desarrollar un algoritmo apropiado para la selección del mejor nombre para los clusters generados basado en la búsqueda del atributo más recurrente dentro de cada cluster. Conseguir esta segunda etapa implicaría haber pasado de simplemente agrupar palabras semánticamente similares a construir una taxonomía propiamente dicha, es decir con unidades gobernadas por sus

correspondientes hiperónimos (*horchata* como un tipo de *bebida*, *bicicleta* como un tipo de *vehículo*, etc.).

En la misma línea se puede ir incluso más allá y acometer también el intento de clasificar estos atributos que encontramos coocurriendo frecuentemente con nuestras unidades analizadas. Esto sería un experimento de otra naturaleza ya que requeriría mayor información acerca de la lengua castellana, pero es también una vía de investigación muy interesante. Por un lado, se trataría de intentar que el algoritmo distinga aquellos atributos que no son apropiados como hiperónimos, entre los cuales encontramos típicamente calificativos sobre color, el tamaño, la edad y los distintos aspectos sobre los que los sustantivos estudiados pueden admitir una predicación. Pero por otro lado también podríamos hacer un estudio previo para hacer listas de cuáles son los atributos típicos de ciertos tipos de entidades, por ejemplo los animales pueden ser salvajes, hembra, macho, y de un determinado color, de un tamaño, etc. Lo que queremos decir es que existe un número limitado de predicados para cada tipo de entidad (o al menos un número limitado de predicados típicos o normales, ya que la naturaleza del lenguaje es ilimitada y el corpus abunda siempre en contraejemplos y peculiaridades).

Finalmente, quedan también como trabajo pendiente los siguientes puntos: 1) ampliar el experimento con mayor número de nombres y grupos semánticos; 2) ampliar la ventana de contexto a enegramas de $n > 2$; 3) tener en cuenta la frecuencia de los coocurrentes en lugar de considerarlos como valores binarios, tal como se explica en la sección 3.4.; 4) observar por separado los coocurrentes según su categoría gramatical, ya que en este experimento los hemos tratado a todos de la misma manera, 5) intentar, de la misma forma, la utilización clases semánticas en lugar de los coocurrentes tal como aparecen en el corpus (es decir, cambiar en el corpus coocurrentes como *blanco, negro, azul,* etc., por COLOR o bien *caliente, tibio, helado,* etc., por una etiqueta como TEMPERATURA) ya que esto aumentaría nuevamente nuestra capacidad de generalizar a partir del corpus al permitirnos relacionar elementos formalmente distintos. Finalmente, y como ya dijimos, 6) combinar este método con los otros que estamos desarrollando en paralelo.

7. REFERENCIAS BIBLIOGRÁFICAS

Alfonseca, E. / Manandhar, S. (2002): «Extending a lexical ontology by a combination of distributional semantics signatures». *Proceedings of EKAW'*02, 1-7.

Alshawi, H. (1989): «Analysing the dictionary definitions». En *Computational lexicography for natural language processing*. Longman Publishing Group, White Plains, NY, USA, 153-169.

Apresjan, J. D. (2008): *Systematic Lexicography.* Oxford: Oxford University Press.

Auger, A. / Barrière, C. (eds.) (2008): «Pattern-based Approaches to Semantic Relation Extraction». Número especial de *Terminology* 14(1).

Aussenac-Gilles, N. / Jacques, M. (2008): «Designing and evaluating patterns for relation acquisition from texts with Caméléon». *Terminology* 14(1): 20-44.

Barrière, C. / Popowich, F. (1996): «Building a noun taxonomy from a children's dictionary». *Proceedings of Euralex*'96, Gothenburg, Sweeden, 65-70.

Boguraev, B. (1991): «Building a lexicon: The contribution of computers». *International Journal of Lexicography*, 4(3): 227-260.

Bullinaria, J.A. (2008): «Semantic Categorization Using Simple Word Co-occurrence Statistics». *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics,* Hamburg, Germany.

Chodorow, M. / Byrd, R. / Heidorn, G. (1985): «Extracting semantic hierarchies from a large on-line dictionary». *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, Chicago, Illinois, 299-304.

Ciaramita, M. (2002): «Boosting Automatic Lexical Acquisition with Morphological Information». *Proceedings of the ACL*-02 *Workshop on Unsupervised Lexical Acquisition*, ACL, Stroudsburg, PA, USA, 17-25.

Curran, J. (2004): *From Distributional to Semantic Similarity*. Tesis doctoral, University of Edinburgh.

Fox, E. / Nutter, J. / Ahlswede, T. / Evens, M. / Markowitz, J. (1988): «Building a large thesaurus for information retrieval». *Proceedings of the second conference on Applied natural language processing*, Association for Computational Linguistics, Morristown, NJ, USA, 101-108.

Fung, P. / Mckeown, K. (1997): «Finding Terminology Translations From Non-Parallel Corpora». *Proceedings of the 5th Annual Workshop on Very Large Corpora*, Hong Kong, 192-202.

Grefenstette, G. (1994): *Explorations in Automatic Thesaurus Construction*. Kluwer, Dordrecht, The Netherlands.

Guthrie, L. / Slator, B. / Wilks, Y. / Bruce, R. (1990): «Is there content in empty heads?» *Proceedings of COLING'*90, Helsinki, Finland, 138-143.

Hanks, P. (2004): «Corpus Pattern Analysis». *Proceedings of EURALEX 2004*, Lorient, France, 87-97.

Harris, Z. (1954): «Distributional structure». *Word* 10(23): 146-162.

Hearst, M. (1992): «Automatic acquisition of hyponyms from large text corpora». *Proceedings of COLING'92*, Nantes, France, 539-545.

Kilgarriff, A. / Rychly, P. / Smrz, P. / Tugwell, D. (2004): «The Sketch Engine». *Proceedings of EURALEX 2004*, Lorient, France, 105-116.

Lin, D. (1998): «Automatic Retrieval and Clustering of Similar Words». *Proceedings of COLING'98*, 768-774.

Meyer, I. (2001): «Extracting Knowledge-rich Contexts for Terminography: A Conceptual and Methodological Framework». En D. Bourigault, C. Jacquemin and M.C. L'

Homme (eds), *Recent Advances in Computational Terminology*, John Benjamins, 279-302.

Michel, J. / Shen, Y. / Aiden, A. / Veres, A. / Gray, M. / The Google Books Team / Pickett, J. / Hoiberg D. / Clancy, D. / Norvig, P. / Orwant, J. / Pinker, S. / Nowak, M. A. / Aiden,

E. (2011): «Quantitative Analysis of Culture Using Millions of Digitized Books». *Science* 331(6014): 176-182.

Nakamura, J. / Nagao, M. (1988): «Extraction of Semantic Information from an Ordinary English Dictionary and its Evaluation». *Proceedings of COLING'*88, Budapest, Hungary, 459-464.

Nazar, R. (2010): *A Quantitative Approach to Concept Analysis*. Tesis doctoral. IULA – Universitat Pompeu Fabra.

Nazar, R. / Renau, I. (2012): «A Co-occurrence Taxonomy from a General Language Corpus». *Proceedings of EURALEX 2012*. Oslo, Norway, 367-375.

Nazar, R. / Renau, I. (en preparación): «Inscrutable-strange-mysterious-infinite-unsearchable are the ways of the Lord: paradigmatic relations and the clustering of semantically similar words».

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

Pekar, V. / Krkoska, M. / Staab, S. (2004): «Feature Weighting for Co-occurrence-based Classification of Words». *Proceedings of COLING'*04.

Potrich, A. / Pianta, E. (2008): «L-hypernymy: Learning Domain Specific hypernymy relations from the Web». *Proceedings of LREC'08*, Marrakech, Morocco. ELRA.

Rapp, R. (1999): «Automatic Identification of Word Translations from Unrelated English and German Corpora». *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, 519-526.

Renau, I. (2012): *Gramática y diccionario: las construcciones con «se» en las entradas verbales del diccionario de español como lengua extranjera*. Tesis doctoral. IULA - Universitat Pompeu Fabra.

Renau, I. / Alonso, A. (En prensa): «Using Corpus Pattern Analysis for the Spanish Learners' Dictionary DAELE (Diccionario de aprendizaje del español como lengua extranjera)». Corpus Linguistics Conference, Birmingham (Reino Unido), 20-22 julio 2011.

Renau, I. / Nazar, R. (2011): «Propuesta metodológica para la creación automática de patrones léxicos usando el Corpus Pattern Analysis». *Proceedings of the 27th Conference of the Spanish Society for Natural Language Processing.* Huelva: University of Huelva.

Renau, I. / Nazar, R. (2012): «Hypernym extraction by definiens-definiendum co-occurrence in multiple dictionaries». *Procesamiento del Lenguaje Natural* (49), 83-90.

Rydin, S. (2002): «Building a hyponymy lexicon with hierarchical structure». *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*, Morristown, NJ, USA. Association for Computational Linguistics, 26-33.

Schütze, H. / Pedersen, J.(1997): «A co-occurrence-based thesaurus and two applications to information retrieval». *Information Processing and Management* 33(3): 307-318.

Sinclair, J. (ed.) (1987). *Looking Up. An Account of the COBUILD Project in Lexical Computing.* London: Collins.

Sinclair, J. (1991). *Corpus. Concordance. Collocation.* Oxford: Oxford University Press.

Wilks, Y. / Fass, D. / Guo, C. / McDonald, J. / Plate, T. / Slator, B. (1989): «A Tractable Machine Dictionary as a Resource for Computational Semantics». En *Computational Lexiography for Natural Language Processing*. B. Boguraev and T. Briscoe (eds): 193-228. Essex, UK: Longman.