Extraños-misteriosos-insondables-inescrutables son los caminos del Señor: extracción de relaciones paradigmáticas mediante análisis estadístico de textos

Rogelio Nazar

Seminario presentado en el IULA,
Universidad Pompeu Fabra, 18 de abril 2013

Resumen

Este trabajo presenta un método para la clasificación no supervisada de unidades léxicas en clases morfológicas y semánticas, basado en una medida de similitud paradigmática como único recurso, es decir que no incorpora conocimiento externo al corpus analizado. La medida se puede entender mejor con la idea del "comodín" o "asterisco", que se refiere a una posición de palabra dentro de un enegrama extraído del corpus. Se trata de reemplazar una de las palabras del enegrama por un asterisco para luego ver qué palabras aparecen en el corpus ocupando esa posición. Este registro permite entonces el análisis por clustering distribucional de las unidades léxicas según la cantidad de enegramas que tienen en común. Una de las utilidades prácticas del método sería utilizarlo dentro de un etiquetador morfosintáctico o bien como un mecanismo para engrosar una ontología léxica existente.

1. Introducción

Este seminario presenta los resultados preliminares de un trabajo[1] que está actualmente en curso y que consiste en el desarrollo de una metodología para la clasificación de unidades léxicas según criterios morfológicos y semánticos mediante el análisis cuantitativo de un corpus de grandes dimensiones. El modelo de trabajo es conceptual y metodológicamente simple: se contrastan unidades léxicas no en presencia unas de otras (eje sintagmático) sino en ausencia (eje paradigmático). La idea es caracterizar paradigmas como posiciones en determinados contextos y utilizar esta información como medida de asociación de forma tal que sea posible agrupar palabras en clases (análisis cluster). Según esta medida de asociación, dos o más palabras se agrupan en una misma clase si muestran tendencia a aparecer en los mismos contextos, definidos como secuencias de palabras (enegramas).

La clasificación se realiza de manera no supervisada, es decir que no existe una fase previa de "entrenamiento" ni hay tampoco procesamiento lingüístico o información externa al corpus. La aplicación de la técnica de clustering resulta en un número indeterminado de conjuntos de palabras que de manera natural exhiben características en común. De esta manera, un cluster puede

[1]Esta investigación contó con la colaboración de Irene Renau.

reflejar el aspecto morfológico, como el caso adjetivo-masculino-plural, o el semántico, cuando agrupa nombres de personas o palabras pertenecientes a una misma clase semántica, tal como en el ejemplo que lleva el título de este trabajo. Si introducimos la expresión "los caminos del Señor" en un motor de búsqueda, obtenemos resultados en los que se encuentran palabras como *misteriosos*, *insondables*, *inescrutables*, etc., que son intercambiables en el mismo contexto y que, como hablantes del castellano, reconocemos intuitivamente que poseen en común características morfosintácticas y semánticas.

Llamaremos *coeficiente de similitud paradigmática* a la medida utilizada para hacer este agrupamiento, que sería una estimación de la probabilidad que tienen distintas palabras de aparecer en los mismos contextos, definidos de manera restringida como secuencias de entre tres y cinco palabras. Estas agrupaciones, si corresponden aproximadamente con clases semánticas o sintácticas, tienen un potencial interesante en campos como la lexicografía y la lingüística tanto computacional como general. Naturalmente, la idea del "asterisco" no es nueva en lingüística de corpus. Ya era posible examinar de manera manual o con la ayuda de cualquier extractor de concordancias casos como los de los ejemplos *Jake* y *Janine* (cuadro 1), tomados del British National Corpus (BNC). Alguien interesado por ejemplo en los *phrasal verbs* o en el complemento de régimen preposicional, podía extraer concordancias solicitando una secuencia como [nombre propio] seguido de [verbo] seguido de palabras como [*up*] o [*at*], y así obtener contextos de aparición y distintas instancias de los verbos buscados.

Cuadro 1: Ejemplos de enegramas con asteriscos

Jake	*	up
	brought	
	wakes	
	lifted	
	picked	
	...	

Janine	*	at
	looked	
	glanced	
	sat	
	is	
	...	

Lo que este trabajo propone es en esencia algo muy similar. Sin embargo, en contraste con esta técnica semi-manual, la aplicación de una técnica de clustering, y sobre todo el análisis masivo de grandes corpus, nos permite relacionar palabras no solamente porque aparecen juntas en estos enegramas sino en muchos otros, y así las clases morfológico-semánticas que resultan interesan porque los grandes números proporcionan solidez y objetividad al análisis. En este caso no se trata solamente de ahorrar trabajo manual sino que el análisis estadístico hace posible llegar a resultados que no obtendríamos de otra manera, dado que la cantidad de datos que analizamos computacionalmente no está en la escala de lo que puede hacer una persona.

En cuanto a los límites de la propuesta, debe aclararse que el trabajo actual solo pretende conseguir la generación de las clases semánticas, y no incluye por tanto el etiquetado semántico de las palabras en contexto, ya que esto es una tarea distinta y tiene otras dificultades. Determinar que una palabra en contexto pertenece a una clase u otra es un desafío entre otra razones porque

uno ya debe contar con una taxonomía bien desarrollada y también con algún sistema de resolución de los problemas de homonimia y polisemia. En ese caso, el contexto puede aportar pistas para determinar, por ejemplo, si un nombre propio refiere a una persona o un lugar o si un museo es un organismo o un edificio.

Existe un amplio abanico de aplicaciones para estas agrupaciones de palabras. Los resultados pueden servir para la expansión de vocabularios y ontologías léxicas, pero también pueden interesar a quienes estudian la combinatoria léxica, a quienes desarrollan etiquetadores morfosintácticos o a los interesados en la detección de colocaciones (donde podría servir como complemento a otras técnicas tales como las que se describen en Ferraro et al. [FNW11]). A continuación explicaré brevemente cuál era la motivación para obtener específicamente estos resultados. Luego comentaré brevemente algunos trabajos relacionados y pasaré entonces a exponer la metodología en más detalle, siguiendo con un análisis de los resultados para terminar con una reflexión sobre el trabajo realizado y algunas líneas de trabajo futuro.

2. Contexto de la investigación

La motivación de este trabajo se encuentra en un proyecto más amplio en colaboración con Irene Renau [RN11], en el que hacemos un análisis de las unidades lingüísticas en un corpus en castellano para detectar los distintos patrones de uso de estas unidades, en la misma línea de lo que se está haciendo en inglés con el proyecto Corpus Pattern Analysis [Han04]. Nuestra idea es crear un recurso lexicográfico en el que una entrada como la del verbo "volcar", por ejemplo, detallaría distintos patrones de uso tales como los siguientes:

- 1: [[Vehicle]] volcar
- 2: [[Human]] volcar [[Liquid]]
- 3: [[Animate - Process]] volcar [[Artifact]]
- 4: [[Human]] volcarse (en - con [[Activity]])
- 5: [[Human 1]] volcarse [[con Human 2]]

El primer caso corresponde al patrón de uso del verbo *volcar* en el que toma como sujetos a sustantivos que pertenecen a la clase semántica de los vehículos. Así, en un accidente, por ejemplo, puede volcar un *autobús*, un *coche*, un *camión*, etc. En el segundo caso, una persona puede por ejemplo volcar un líquido sobre una superficie, en el sentido de derramar. En el tercero, el verbo puede tener como sujeto a entidades animadas o procesos y como objeto un vehículo o algún otro tipo de artefacto, como cuando los manifestantes vuelcan contenedores, vehículos o cubos de basura. El cuarto y el quinto patrón corresponden a usos como en el que decimos que una persona se vuelca con entusiasmo a una actividad o con otra persona. Y es posible continuar encontrando patrones del

verbo ya que el número no está determinado: podemos pensar en otros usos como cuando decimos que vamos a volcar datos de un formulario en una base de datos.

Para poder llevar a cabo la extracción de estos patrones de manera automática necesitamos disponer de una taxonomía del castellano que sea lo suficientemente grande como para tener una amplia cobertura en cualquier corpus analizado. La versión de castellano de WordNet no tiene todavía el tamaño suficiente para conseguir el objetivo de proveer las relaciones hiperonímicas entre los argumentos del verbo y alguno de los tipos semánticos, y por esta razón hemos decidido emprender nosotros mismos la creación de una gran taxonomía del castellano a partir de diversas fuentes. Además de WordNet, entre otros recursos estamos utilizando la Wikipedia en castellano, ya que por su estructura interna (los vínculos entre categorías) se puede interpretar y utilizar como una ontología. La limitación de este recurso es que sirve más bien para poblar una ontología existente con instancias de nombres propios de personas, lugares, organizaciones, terminología científico-técnica, etc. Evidentemente, nos interesa tener la capacidad de reconocer estas entidades en los textos, pero necesitamos además un mecanismo que sea útil para el vocabulario general.

En el caso concreto del léxico general hemos presentado cuatro metodologías distintas para la generación automática de taxonomías. Nuestra idea es implementar estos cuatro algoritmos de manera tal que interactúen entre ellos, es decir que se refuercen mutuamente en sus decisiones. Así, si la mayoría de ellos considera que, por ejemplo, un autobús es un tipo de vehículo, entonces esta afirmación será más segura que en el caso de que cada uno de ellos tenga una "opinión" diferente.

Para más detalles sobre las características de estos algoritmos generadores de taxonomías referiré a otras publicaciones. Para ofrecer aquí una explicación en pocas palabras, se puede decir que el primer algoritmo está basado en estadísticas sobre un corpus lexicográfico, y registra la cantidad de veces en que coocurren las palabras en el definiens y en el definiendum [NJ10, RN12]. El segundo es un análisis de los bigramas (secuencias de dos palabras) del corpus de Google Books [MSA+11], y se trata de un experimento en el que clasificamos un conjunto limitado de palabras en función de la palabra inmediatamente anterior y posterior que tienen en común [NR12a], un método que tiene una alta precisión pero con un elevado coste computacional debido a su complejidad cuadrática y al gran tamaño del corpus analizado. El tercer algoritmo está basado en el que se presenta en [Naz10], pero se aplica al léxico general en lugar de la terminología especializada, y ya ha sido implementado y evaluado [NR12b]. Este método no tiene tan buenos resultados cuando es aplicado al vocabulario general, lo cual no es realmente sorprendente ya que los mecanismos de generación de taxonomías en la terminología especializada presentan unas características diferentes a las del vocabulario general.

El cuarto algoritmo viene a ser el que presento hoy aquí, y tiene por objeto agrupar unidades léxicas en clases para luego distribuirlas en la mencionada taxonomía del castellano, una vez que ha sido generada al menos en su estructura elemental. En esta línea de investigación hay también un trabajo pendiente

que será la unificación de todas estas técnicas en una taxonomía del castellano (revisada y corregida de forma manual), además de un mecanismo de generación de estas taxonomías que, en la medida de lo posible, sea independiente de lengua.

3. Trabajo relacionado

La bibliografía relacionada con este trabajo es sumamente extensa y cualquier intento de síntesis resulta irremediablemente arbitrario. Las lecturas pueden ser más generales o más específicas. Aquí dejaré de lado las nociones más generales sobre las técnicas de análisis de la combinatoria léxica en corpus, asumiendo que es conocimiento compartido. Buenas primeras lecturas se pueden encontrar, por ejemplo, en el trabajo de John Sinclair [Sin91] o el de Church y Hanks [CH90]. Esta sección no incluye, tampoco, referencias a la extensa cantidad de trabajo que se ha llevado a cabo en la extracción de relaciones de hiperonimia a partir de diccionarios como fuente o bien corpus textuales utilizando los llamados patrones gramaticales del tipo *X es un tipo de Y*, debido a que en la presente propuesta nos vamos a interesar exclusivamente por los enfoques basados en estadística. Los interesados en esta vertiente podrán encontrar una revisión bibliográfica en [NVW12a].

Baroni y Lenci [Len08, BL08] ofrecen una adecuada presentación del campo que llamamos análisis distribucional o semántica distribucional, y también hay una selección bibliográfica sobre el tema en [Naz10]. A menudo se atribuye a Zellig Harris [Har54] el papel de iniciador de este campo. También sería justo destacar el trabajo pionero de Martin Phillips [Phi85]. Aquí ya se encuentran los fundamentos o nociones elementales de lo que es el clustering distribucional y cómo se calculan las asociaciones sintagmáticas y paradigmáticas y otras ideas fundamentales como que las relaciones de coocurrencia pueden ser tanto simétricas como asimétricas, es decir que una palabra puede mostrar una tendencia a coocurrir con otra pero es posible que esta relación no sea recíproca. Su trabajo es continuado por Gregory Grefenstette [Gre94], quien incorpora el análisis gramatical en el proceso (relaciones de dependencia) en lugar de analizar solamente coocurrencia simple como hacía Phillips. Grefenstette demuestra que es capaz de recuperar "pseudo-sinónimos", que son palabras que están ortográficamente alteradas, y las relaciona con su versión correctamente escrita gracias a su similitud distribucional.

En una línea muy similar aparece citado con frecuencia el trabajo de Dekang Lin [Lin98] sobre análisis distribucional. La explicación que el autor nos ofrece resulta muy intuitiva: si uno encuentra secuencias como *una botella de ** o *se emborrachó con **, entonces es de suponer que * representa a una palabra como *vino*, *vodka*, *cerveza*, etc. Lin también utiliza el análisis de dependencias para extraer tripletas del corpus, que consisten en dos palabras unidas por una determinada relación gramatical. Por ejemplo, en un corpus de genética, una palabra en inglés como *cell* (célula) puede ser el sujeto de verbos como *to absorb*, *to adapt* o *to behave* y el objeto de verbos como *to attack*, *to bludgeon*, *to call*, etc. Procesando un corpus de 64 millones de palabras, Lin agrupó las unidades

léxicas en función de su similitud en cuanto a estas relaciones de dependencia y descubrió que podía generar un tesauro que resulta más similar a WordNet de lo que es el tesaurus Roget's. El tesauro que trae incorporado el Sketch Engine [KRST04] también funciona de una manera esencialmente similar.

También se ha planteado trabajar en la misma línea pero sin incorporar el análisis de dependencias [SP97], que es algo que encarece el análisis. Una referencia muy frecuente en esta línea es la del *Latent Semantic Analysis* (LSA) [LD97], un modelo computacional capaz de elaborar representaciones semánticas únicamente a partir de las coocurrencias en corpus, sin información adicional. Está basado en una matriz término x documento, como es habitual en el campo de la recuperación de información, y utiliza solamente 300 dimensiones que son definidas como palabras coocurrentes sin importar el orden o la gramática. Las palabras, a su vez, son definidas simplemente como secuencias de letras entre espacios en blanco o signos de puntuación. Usando una enciclopedia como corpus, estos autores han sido capaces de emular el resultado de un humano frente a una prueba tipo *multiple choice* de sinónimos, el llamado TOEFL (Test of English as a Foreign Language). En esta prueba, el estudiante recibe una palabra como input y tiene que elegir un sinónimo entre cuatro opciones que se le ofrecen. El algoritmo de LSA fue capaz de obtener resultados similares a la media de los humanos, que no son hablantes nativos de inglés pero se supone llegan a la prueba con un nivel universitario.

La primera conclusión a la que llegan los autores es que la idea puede tener una aplicación práctica, pero la segunda es un poco más radical, ya que presentan al LSA como una teoría del conocimiento, es decir, un modelo que explica la forma en que el ser humano lleva a cabo la adquisición del conocimiento. Ciertamente, los niños aprenden el significado de la mayoría de las palabras no porque alguien se las explique o porque las busquen en el diccionario, sino que infieren su significado gracias a los contextos en que las palabras son usadas. Más allá del aspecto práctico, entonces, estos hallazgos dan pie a una reflexión más profunda desde un punto de vista psicológico, ya que esta propiedad del lenguaje sugiere pistas sobre el funcionamiento de nuestra mente.

Una idea parecida aparece en el trabajo de John Bullinaria [Bul08], que hace un clustering distribucional de las palabras del BNC utilizando un contexto muy reducido (una palabra a la izquierda y otra a la derecha). Se trata de una aproximación bastante relacionada con la de la presente propuesta ya que consiste en una clasificación no supervisada (clustering). La diferencia es que utiliza una medida de asociación (la información mutua o MI) y una herramienta externa para el clustering (CLUTO). En un experimento con 44 sustantivos concretos, por ejemplo, obtiene como resultado seis categorías semánticas que se corresponden con lo que cabía esperar, es decir que acaban en un mismo cluster palabras como *hammer*, *chisel* o *screwdriver*, que luego se funden en el dendrograma resultante con otro cluster que contiene las palabras *pencil* y *pen*, y luego con otro que contiene *scissors*, *knife* y *spoon*. El autor comenta algo que también hemos visto castellano, y es que palabras como "chiken" suele ser clasificadas junto con los alimentos y no con animales, lo cual no es sorprendente pero nos recuerda la complejidad del problema de la taxonomía, ya que una

misma palabra se puede clasificar de varias maneras.

Hay otros trabajos específicamente relacionados con la metodología propuesta pero que se interesan más por los aspectos sintácticos, como la tesis doctoral de Alexander Clark [Cla01], en el contexto de los métodos no supervisados para la inducción de sintaxis. En el caso concreto de este autor (en el capítulo 5 de la tesis), nos presenta un método para la inducción de categorías gramaticales. También trabaja con el BNC y consigue extraer 77 clusters de palabras que se corresponden aproximadamente con las categorías gramaticales.

Otro trabajo que interesa también desde el punto de vista sintáctico es el proyecto StringNet [WT10], en el que se trabaja con el concepto de "enegramas híbridos". A diferencia de los enegramas tradicionales, que son simplemente secuencias de palabras, en los híbridos se reemplazan algunas de las palabras por categorías gramaticales (algo similar a nuestros asteriscos). Por ejemplo, un enegrama como *It's the thought that counts*, se convierte en *it's the [noun] that [verb]*, lo cual permite generalizar hacia muchas otras expresiones que cumplen el mismo patrón. Otros ejemplos serían: *It is safe to [verb] that...* (por ejemplo, en el caso de *It is safe to [assume/say/predict] that...*, o bien *There is a tendency for [noun] to [verb]*, *from [Possessive determiner] point of view*, etc. Su propuesta, por tanto, explora la zona fronteriza entre los enegramas simples y las construcciones sintácticas. El trabajo presentado hoy es diferente porque pone el énfasis en el aspecto semántico más que en el sintáctico.

Entre los trabajos que se interesan específicamente por el aspecto semántico podemos mencionar una línea de investigación muy activa que consiste en la adquisición de información léxica con el objeto de poblar recursos lexicográficos ya existentes. En esta línea se encuentra el trabajo de Massimiliano Ciaramita [Cia02], que hace la prueba de expandir un "mini-wordnet". Su enfoque es similar al de la clasificación de documentos, solo que aquí lo que se clasifican los palabras y las clases se corresponden con las categorías semánticas de una taxonomía (el mencionado mini-wordnet). En este esquema, el autor asimila a la idea de un "documento" lo que en realidad es el contexto de aparición de una palabra, es decir un fragmento del texto en el corpus en donde aparece la palabra que se desea clasificar. Los rasgos que se utilizan para la clasificación son las palabras que se encuentran en el contexto, en la vecindad de la palabra analizada, a los que se refiere como "colocativos". El suyo es el enfoque del aprendizaje supervisado, que complementa con gran cantidad de trabajo manual en la generación de diversas reglas. La idea es similar sin embargo a lo que ya hemos visto en otros autores. Por ejemplo, si en el contexto encontramos que la palabra *comer* aparece junto a un sustantivo, entonces ese sustantivo se clasificará dentro la clase semántica COMIDA y no en clases como PLANTAS o EMPRESAS. Otra medida que utiliza, y que resulta particularmente interesante, es la explotación de los rasgos morfológicos de las palabras para clasificarlas semánticamente, ya que hasta cierto punto existe una correlación entre los afijos de una palabra y su categoría semántica. Pensemos por ejemplo en el sufijo *-isis* u *-osis* y la clase semántica de las enfermedades. La información morfológica incluye incluye número (plural, singular), la alternancia mayúscula / minúscula, o la presencia de unidades morfológicamente complejas que comparten el mismo núcleo (por

ejemplo en el caso de *drinking age* y *age* o *chairman* y *man*). Le podríamos criticar sin embargo el haber compilado a mano los listados de afijos, cuando en realidad la asociación entre información morfológica y la pertenencia a clases semánticas se puede aprender de manera automática (como hacemos, por ejemplo, en [NVW12b]).

El trabajo de Ciaramita se puede agrupar junto con el de otros autores que se dedican a la clasificación de entidades nombradas, como el caso de Alfonseca y Manandhar [AM02], que también presentan una manera de expandir recursos léxicos como Wordnet. De nuevo, aquí se considera que las palabras poseen marcas distintivas en el plano semántico conformada por aquellas con las que suele coocurrir. De esta forma, un determinado sustantivo muestra tendencia a aparecer como el sujeto o el objeto de un número limitado de verbos, de la misma manera que aparece también combinado con un número limitado de adjetivos. La clasificación se lleva a cabo entonces mediante una combinación de estas marcas distintivas, y la dificultad y el mérito del trabajo es que se intenta clasificar las palabras en su contexto, lo cual implica la necesidad de sortear los posibles problemas de ambigüedad. Una ciudad o un país, por ejemplo, pueden ser una organización (un gobierno) o bien un lugar, dependiendo del contexto.

El campo de la la clasificación de entidades nombradas (Named Entity Categorization o NEC), en el que puede incluir el trabajo de Alfonseca y Manandhar, es muy amplio y muy complejo, pero se puede decir que consiste básicamente en clasificar los nombres propios que aparecen en un texto en las clases de *Person*, *Location* y *Organization*. Para un panorama general del campo se puede consultar el trabajo de Nadeau y Sekine [NS07]. El campo de investigación en terminología especializada o terminografía computacional está íntimamente relacionado con este tema, pero ambos campos parecen evolucionar de manera paralela (para referencias bibliográficas, véase [NVW12a]). Son, sin embargo, campos distintos ya que en el caso de la terminología la taxonomía sería mucho más amplia que estas tres clases. En la categorización de entidades nombradas se suele utilizar "palabras detonantes" (*trigger words*, [GW98]) que forman parte del nombre de una entidad que se desea clasificar y que ofrecen pistas sobre su categoría semántica. Por ejemplo, una entidad como "Wing and Prayer Airlines" es casi seguramente una empresa debido a la presencia de la palabra "Airlines", y por tanto podríamos añadirle la etiqueta *Organization*; de la misma manera la expresión "Bay of Pigs" será probablemente un lugar (*Location*), porque contiene la palabra "Bay". Esta es sin embargo una línea de investigación que ya no se relaciona directamente con el objeto de interés del presente trabajo.

Finalmente, el trabajo de Chris Biemann [BBQ03], está algo cerca de la presente propuesta. Este trabajo se centra en el análisis y clasificación no supervisada de las unidades de un corpus en clases que se corresponden con categorías morfológicas y semánticas, y hablan específicamente de relaciones paradigmáticas. Como otros autores, hablan también de colocaciones en un sentido laxo, es decir, para referirse a combinaciones de unidades que no forman necesariamente colocaciones pero sí muestran una tendencia a coocurrir en los mismos contextos. Concretamente, hablan de cuatro tipos de coocurrencia: la adjetivo-sustantivo, sustantivo-adjetivo, sustantivo-verbo y, finalmente, una cooccurrencia entre sus-

tantivos que se da en el contexto de una oración (unidades no adyacentes), sin tener en cuenta el orden de aparición ni la distancia entre ellas. Posteriormente, las palabras se comparan entre sí y se asocian aquellas que comparten colocativos. Así es posible relacionar, por ejemplo, palabras como *caliente*, *fresco* y *frío* porque suelen aplicarse a sustantivos para los que la temperatura es un atributo relevante.

En las últimas décadas se ha publicado una gran cantidad de trabajos relacionados con este tema. Los hay de mayor o menor complejidad y presentan técnicas muy variadas, aunque en general tienen en común la idea de que las palabras que son semánticamente similares aparecen en los mismos contextos. Una de las conclusiones que se desprende de la lectura de todo este material es que no existe todavía una manera de comparar rigurosamente el resultado de estas distintas estrategias. Los números de precisión y cobertura que mencionan los autores no son comparables ya que se obtienen de experimentos muy distintos. No se ha determinado aún que un método sea mejor que otro, en términos de efectividad o eficiencia. Tampoco parece que la presente propuesta sea directamente comparable con alguno de estos métodos, si bien es evidente que existen múltiples similitudes con muchos de ellos, al menos con los que se enmarcan en el conjunto de los algoritmos no supervisados.

4. Materiales y métodos

En comparación con los algoritmos normalmente utilizados en los trabajos relacionados, la metodología de la presente propuesta es verdaderamente simple. El hecho de no tener ningún tipo de procesamiento lingüístico de los textos reduce considerablemente la complejidad técnica, y los pasos realizados se pueden explicar con muy pocas palabras.

Por el momento hay dos experimentos realizados, uno en inglés con el BNC y otro en castellano con una muestra de artículos del diario El País, ambos con un tamaño cercano a los cien millones de palabras. Está actualmente en curso un nuevo experimento con un corpus varias veces más grande: la nueva versión del corpus de Google Books [LML+12], de un tamaño que supera los setenta mil millones de palabras. Los resultados, sin embargo, se harán esperar algunos días más ya que, debido a las limitaciones de memoria de los equipos informáticos actualmente disponibles, el procesamiento de un corpus de estas dimensiones exigirá cambios en el diseño metodológico para optimizar el procesamiento.

En las secciones siguientes se explican paso a paso los procesos de indexación del corpus (4.1), la extracción de parejas de palabras que comparten la posición del asterisco (4.2) y el filtrado de unidades poco informativas (4.3) para terminar en el proceso de clustering de las parejas de palabras (4.4), cuyo resultado se explica luego en la sección 5. Cabe aclarar que es posible que en el clustering final una misma forma partenezca a más de una categoría, medida con la que se pretende atajar el problema de la polisemia.

4.1. Indexación del corpus

El primer paso en el procesamiento del corpus es su indexación. La indexación implica en general el convertir el texto en tablas que registran la frecuencia de aparición de las palabras. En el experimento que nos ocupa registramos más bien las secuencias de palabras (entre tres y cinco), pero más allá de esto, el procedimiento es estándar. La particularidad es que, una vez generadas estas tablas de enegramas, las alteramos de manera automática reemplazando uno de los componentes interiores de cada enegrama por un asterisco. De esta manera, un enegrama como, por ejemplo, *just down the road* se convierte en *just * the road*, y, en un ciclo posterior, también en *just down * road* (cuadro 2), y así sucesivamente con todos los enegramas. Como ya se sabe, el propósito es registrar en cada uno de los reemplazos las palabras que ocupan la posición del asterisco, información que se obtiene de las mismas tablas del índice. Las tablas 3, 4 y 5 podrán servir para ilustrar la forma en que se almacena esta información.

Cuadro 2: Otros ejemplos de palabras en posición del asterisco

just	*	the road
	down	
	across	
	up	
	along	
	over	
	...	

just	*	if they
	as	
	wondered	
	wondering	
	asking	
	...	

Cuadro 3: Palabras en la posición del asterisco en: *are * sound*

*are * sound*
absolutely 2, structurally 2, financially 2, theoretically 1, scientifically 1, nutritionally 1, ideologically 1, geographically 1, basically 1, commercially 1, electrically 1
the 3, given 1, they 1, of 1, therefore 1, punching 1, not 1, no 1, still 1, after 1, proving 1, also 1, very 1, for 1, minor 1, as 1, poor 1, considered 1, some 1, effective 1

Cuadro 4: Referencias a aeropuertos capturadas en la posición del asterisco

*at * airport* [645]
the 225, Heathrow 56, Manchester 23, Gatwick 19, London 15, Frankfurt 15, Teesside 10, Edinburgh 8, an 7, Glasgow 7, Stansted 7, Dublin 6, Birmingham 5, Coventry 5, Aberdeen 5, Athens 5, Shannon 4, Manila 4, Faro 4, Vienna 4, Amsterdam 3, Milan 3, Liverpool 3, Sarajevo 3, Aldergrove 3, Christchurch 3, Teheran 3, Malaga 3, Geneva 3, Chicago 3, Eglinton 3, Orly 3, ...

Cuadro 5: Nombres de *colleges* en la posición del asterisco

*at * college [1092]*
the 189, university 121, Trinity 52, new 39, Imperial 32,
Darlington 22, a 20, Birkbeck 18, Magee 17, art 13,
Jesus 12, Pembroke 12, Newnham 12, Balliol 12, Goldsmiths 11,
Bedford 10, Somerville 10, Malvern 9, Magdalen 9, Longlands 9,
Telford 8, Merton 8, Clifton 7, Ruskin 7, Dulwich 7, Exeter 7,
Napier 7, her 7, Wellington 7, Swindon 6, Kings 6,
Cheltenham 6, Eton 6, Queens 6, his 6, Wadham 6, Kirby 6,
National 5, Brighton 5, Oriel 5, Chelsea 4, Westminster 4,
Girton 4, Marlborough 4, Worcester 4, Methodist 4, this 4,
Winchester 4, Gresham 3, Coloma 3, Theological 3, Southlands 3,
Glasgow 3, Chester 3, Mason 3, another 3, Edinburgh 3, your 3,
Newcastle 3, Owens 3, Morley 3, Aquinas 3, Jordanhill 3,
Wolfson 3, Emmanuel 3, Wellesley 3, Keble 3, Ushaw 3, Westfield 3, ...

4.2. Extracción de parejas de elementos

Una vez que hemos registrado las tablas que contienen la información de las palabras que aparecen en cada enegrama en la posición del asterisco, el paso siguiente es hacer parejas de palabras que muestren tendencia a aparecer en los mismos enegramas. Debido a la limitación de la memoria disponible, este proceso se tiene que hacer por etapas en las que se van analizando fragmentos del corpus o, más precisamente, de las tablas de indexación. El resultado de este proceso es una nueva serie de tablas en las que ordenamos las parejas por frecuencia, ignorando las que tengan frecuencia menor a tres. Así, en algunas de las tablas que obtenemos en este proceso podemos observar que las palabras *years* y *weeks* tienen una relación paradigmática porque aparecen en la posición del asterisco de distintos enegramas. Lo mismo puede decirse de otras parejas como *minutes - months*, *five - three*, *will - would*, *less - more* (cuadro 6) o, curiosamente, en el caso de nombre de condados ingleses (cuadro 7), como *Oxfordshire - Gloucestershire*, *Oxfordshire - Buckinghamshire*, etc.

4.3. Filtrado de elementos poco informativos

De los listados de parejas solo nos interesan aquellas que sean especialmente informativas y por eso tenemos que filtrarlas con algún criterio estadístico. Un criterio bastante fácil de comprender y de aplicar es limitar el número de parejas con las que se aparea una misma palabra, y una forma de hacerlo es determinar un umbral arbitrario (un parámetro de ejecución definido empíricamente) para el número de unidades diferentes con las que se puede combinar cada miembro de la pareja, valor que, naturalmente, estará correlacionado con la frecuencia. Desaparecerán así de los listados unidades poco interesantes para los fines de este trabajo, tales como las altamente frecuentes *the*, *that*, *this*, etc. Este método resulta más efectivo que simplemente determinar un umbral de frecuencia, ya que es posible que haya unidades que tengan una frecuencia importante y aparezcan a la vez con un número limitado de parejas.

Cuadro 6: Parejas de palabras en posición del asterisco en distintos enegramas

three - more	322
that - because	320
some - this	320
five - three	315
which - this	314
what - which	309
less - more	307
this - each	301
where - that	298
will - would	297
each - that	290
more - even	289
they - that	288
even - less	286
three - other	285
another - this	281
other - some	278

years - weeks	9
lead - years	9
years - minutes	8
years - season	7
years - months	7
hours - minutes	7
when - that	6
with - after	6
period - years	6
lead - minutes	6
minutes - months	6
years - june	6

4.4. Clustering de los elementos

El último paso del análisis es la creación de clusters a partir de las tablas de parejas de palabras generadas en el paso anterior. Es posible advertir ya, por ejemplo en el caso de parejas formadas por las palabras *Oxfordshire*, *Gloucestershire*, *Buckinghamshire*, etc., una tendencia a formar grupos como si se tratara de contactos sociales. La técnica de clustering es ideal para una situación así, ya que nos ayudará a extraer estas agrupaciones de una forma explícita.

El algoritmo de clustering aplicado aquí está basado en uno que fue originalmente utilizado para desambiguación semántica ([Naz10]), pero puede ser utilizado para aplicar de forma general una clasificación no supervisada. El algoritmo aparece explicado a continuación en la forma de pseudocódigo, donde definimos cada $jp \in G(t)$ como las parejas de elementos que leemos de las tablas generadas en el proceso anterior, denominadas conjunto $G(t)$, y que pasan a clasificarse en un conjunto de clusters C. El símbolo k hace referencia al grado de solapamiento entre las parejas y los clusters. Si, por ejemplo, uno de los miembros de la pareja *Oxfordshire* - *Gloucestershire* está incluido en un cluster C_i, entonces decimos que hay un solapamiento de 1. Si las dos palabras están ya en ese cluster (con otras parejas) entonces el solapamiento es de 2.

Algoritmo de clustering

for all $jp \in G(t)$ **do**

1. si no hay clusters C, creamos uno con la primera jp,
2. caso contrario comparamos jp con cada cluster C_i ya creado.
3. si hay el mismo solapamiento con más de un cluster, $\rightarrow$ no clasificamos
4. si en cambio hay solapamiento $> k$ con C_i, $\rightarrow jp \in C_i$
5. si en cambio no hay solapamiento con ninguno, $\rightarrow jp \in$ nuevo cluster C_n
6. si dos clusters se disputan frecuentemente las mismas parejas, entonces se funden

end for

Cuadro 7: Nombres de condados de Inglaterra capturados en las parejas

Oxfordshire - Gloucestershire	91
reporting - Gloucestershire	61
reporting - Oxfordshire	57
Oxfordshire - Buckinghamshire	56
Oxfordshire - Oxford	52
Oxfordshire - Swindon	50
Gloucestershire - Buckinghamshire	49
Oxford - Gloucestershire	49
Gloucestershire - Swindon	44
Gloucestershire - Wiltshire	41
Wiltshire - Oxfordshire	40
Oxford - reporting	39
Oxfordshire - Gloucester	38
reporting - Buckinghamshire	36
Gloucestershire - Gloucester	35
Herefordshire - Gloucestershire	34
mother - Gloucestershire	34
Buckinghamshire - Oxford	32
Oxfordshire - mother	31
Swindon - Oxford	30
Buckinghamshire - Wiltshire	29
Wiltshire - reporting	29
Worcestershire - Oxfordshire	28

Solo se asigna una pareja a un cluster si no se produce el mismo grado de solapamiento con otro. A modo de ilustración, en el ejemplo A, se muestra un ejemplo en el que una pareja como *referred - linked* no es clasificada porque cada uno de sus miembros aparece en clusters distintos. En el segundo caso, en el ejemplo B, la pareja *bacon - nath* es asignada al cluster C579 porque ambos miembros aparecen en él, solapamiento que no se da con ningún otro cluster.

Ejemplo A: Caso de empate

```
"referred-linked" 5
 Match: (C235 = 1 // linked - 18 )
 Match: (C181 = 1 // referred - 4 )
 Indeterminado (1 [C181] == 1 [C235])
```

5. Resultados

Los resultados en los dos casos de los experimentos realizados, tanto en el corpus en inglés como el corpus en castellano, son de casi un millar de clusters que hay que examinar manualmente. Son, efectivamente, grupos que presentan características morfológicas o semánticas en común. Algunos de los clusters son muy pequeños e incluyen solo una pareja de miembros, como uno formado por las palabras *biggest* y *largest* u otro con *older* y *younger*, mientras que otros

Ejemplo B: Caso en que hay un ganador

```
"bacon-nath" 3
 Match: (C579 = 1 // bacon - 11 )
 Match: (C579 = 2 // nath - 2 )
Elegido: cluster C579 (2)
 Se agrega "bacon-nath" al cluster C579
```

llegan a estar formadas por centenares de palabras. En el caso de los clusters pequeños, además de antónimos o sinónimos, a menudo representan también derivación flexiva o variantes ortográficas, como en *application - applications* o en *flicked - flickered*.

Además del tamaño, varía también la naturaleza de los elementos que contienen. Por un lado se puede ver que existe claramente una clasificación según la categoría gramatical, persona y número. Por otro lado, a la clasificación morfológica se superpone otra de naturaleza semántica, que se advierte inmediatamente aunque no sea fácil determinar cuál sería el nombre adecuado para estas clases. En otros casos se ve con claridad que se trata de nombres propios de persona o de lugar, meses del año, números, colores, deportes, etc. Los cuadros 8 y 9 pueden servir a modo de ilustración. En el caso del corpus en castellano el resultado es muy similar (cuadro 10). Existen pequeños clusters que agrupan sinónimos (*presuntas - supuestas* , *absurdas - falsas*, *concluidas - finalizadas - terminadas*) o variantes gráficas *chií - shií*. A medida que el tamaño de los clusters crece (por encima de los centenares de miembros), la pertenencia a una misma clase semántica se va diluyendo, mientras que tiende a perdurar la clase gramatical.

6. Conclusiones y trabajo futuro

Este trabajo ha presentado una técnica para la extracción de unidades de alta similitud paradigmática en el contexto más general de un análisis lexicológico. Si bien se da en el contexto de un proyecto de estudio de la lengua castellana, se puede decir que es una técnica aplicable a cualquier lengua, ya que solo se presupone que el corpus estará compuesto por algo tan universal como una secuencia de palabras.

Siendo un trabajo en curso, los resultados son aun preliminares, y es necesario seguir experimentando y analizando los resultados. La cantidad de datos que se genera (los clusters) para un corpus de cien millones de palabras pueden alcanzar el millar, dependiendo de los parámetros. Estos grupos se podrían volver a agrupar, pero ello requerirá el diseño de la metodología apropiada. La evaluación cuantitativa de los datos está ahora en curso, y se da por vía del muestreo aleatorio y análisis de los clusters, examinando uno a uno sus unidades léxicas. A simple vista, sin embargo, resulta evidente que los resultados pueden ser útiles incluso en este estado temprano de desarrollo.

La aplicación general de este procedimiento puede ser de distinta naturaleza.

Cuadro 8: Ejemplos de los clusters formados (1)

Sustantivos	
effort	2
emphasis	3
energy	1
reliance	1
siege	1
stress	2

Adverbios	
entirely	3
exclusively	3
mainly	4
primarily	3
principally	1

Adjetivos	
costeffective	1
efficient	4
elaborate	1
professional	1
subtle	1

Adjetivos	
eastern	3
heineken	1
lefthand	1
northeast	3
northern	7
northwest	4
provincial	1
southern	4
southernmost	2

Adjetivos	
considerable	2
damaging	1
enormous	3
explicit	1
huge	1
intense	1
massive	1
potent	1
rigorous	1
serious	7
sinister	1

Cuadro 9: Ejemplos de clusters formados (2)

Verbos/Participios	
admits	1
asked	4
growled	1
insisted	1
pleaded	1

Nombres de deportes	
chess	1
cricket	1
football	7
golf	3
rugby	3
soccer	1
tennis	2

Nombres de lugares	
Australia	4
Dublin	3
England	8
France	5
India	3
Janeiro	1
Middlesbrough	2
Newcastle	1
Sunderland	2
Yorkshire	1

Una posibilidad sería utilizarlo como complemento a un etiquetador morfosintáctico (POS-tagger). Sería posible, por ejemplo, utilizar estos clusters para ayudar a la desambiguación o para categorizar correctamente unidades que no están en el vocabulario conocido. Una posibilidad sería intentar clasificar la palabra desconocida en alguno de estos clusters a partir de su comportamiento en otros corpus textuales. Otra posibilidad puede ser calcular la similitud morfológica entre la palabra desconocida y las de los miembros de los distintos clusters, etc.

La aplicación más natural es sin dudas la expansión de recursos léxicos, como una taxonomía. Ello requerirá sin embargo el desarrollo de un mecanismo

Cuadro 10: Ejemplos de clusters formados (3)

Sustantivos - explosivos	
cohetes	3
granadas	3
mortero	3
morteros	3

Participio - causa	
causada	3
desencadenada	2
originada	2
provocada	3

Verbos en infinitivo - infringir	
conculcar	1
contravenir	2
infringir	2
violar	3
vulnerar	4

que permita distribuir de manera automática los clusters producidos por este algoritmo en la taxonomía que se está haciendo del castellano.

En algunos días estarán también disponibles los resultados de la repetición de este experimento con los ya mencionados enegramas de Google Books [LML+12], un corpus muchas veces más grande que además incorpora etiquetado morfosintáctico. Una variante en el procedimiento es que ahora se incluye más información lingüística como la detección y categorización de nombre propios, entre otros procedimientos.

Referencias

[AM02] Alfonseca, E. y S. Manandhar: *Extending a lexical ontology by a combination of distributional semantics signatures.* En *Proceedings of EKAW'02*, páginas 1–7, 2002.

[BBQ03] Biemann, C., S. Bordag y U. Quasthoff: *Lernen von paradigmatischen Relationen auf iterierten Kollokationen.* En *Beiträge zum GermaNet-Workshop: Anwendungendes deutschen Wortnetzes in Theorie und Praxis*, Tübingen, Germany, 2003.

[BL08] Baroni, M. y A. Lenci: *Concepts and properties in word spaces.* Italian Journal of Linguistics, páginas 55–88, 2008.

[Bul08] Bullinaria, J.A.: *Semantic Categorization Using Simple Word Co-occurrence statistics.* 2008.

[CH90] Church, K. y P. Hanks: *Word association norms, mutual information and lexicography.* Computational Linguistics, 16(1):22–29, 1990.

[Cia02] Ciaramita, M.: *Boosting Automatic Lexical Acquisition With Morphological Information.* páginas 17–25, 2002.

[Cla01] Clark, A.: *Unsupervised Language Acquisition: Theory and Practice.* Tesis de Doctorado, COGS, University of Sussex, 2001.

[FNW11] Ferraro, G., R. Nazar y L. Wanner: *Collocations: A Challenge in Computer-Assisted Language Learning.* En *Proceedings of the 5th International Conference on Meaning-Text Theory*, September 2011.

[Gre94] Grefenstette, G.: *Explorations in Automatic Thesaurus Discovery.* Kluwer, Dordrecht, The Netherlands., 1994.

[GW98] Gaizauskas, R. y Y. Wilks: *Information Extraction: Beyond Document Retrieval.* Computational Linguistics and Chinese Language Processing, 3(2):17–60, 1998.

[Han04] Hanks, P.: *Corpus Pattern Analysis.* En *Proceedings of EURALEX*, páginas 87–97, Lorient, France, 2004.

[Har54] Harris, Z.: *Distributional structure.* Word, 10(23):146–162, 1954.

[KRST04] Kilgarriff, A., P. Rychly, P. Smrz y D. Tugwell: *The Sketch Engine.* En *Proceedings of EURALEX*, páginas 105–116, Lorient, France, 2004.

[LD97] Landauer, T. K. y S. T. Dutnais: *A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.* Psychological review, páginas 211–240, 1997.

[Len08] Lenci, A.: *Distributional semantics in linguistic and cognitive research.* Italian Journal of Linguistics, 20(1):1–32, 2008.

[Lin98] Lin, D.: *Automatic Retrieval and Clustering of Similar Words.* En *Proceedings of COLING'98*, páginas 768–774, 1998.

[LML+12] Lin, Y., J.B. Michel, E. Lieberman, J. Orwant, W. Brockman y S. Petrov: *Syntactic Annotations for the Google Books Ngram Corpus.* En *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, páginas 169–174, Jeju, Republic of Korea, 2012.

[MSA+11] Michel, J.B., Y. Shen, A. Aiden, A. Veres, M. Gray, The Google Books Team, J. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak y E. Aiden: *Quantitative Analysis of Culture Using Millions of Digitized Books.* Science, 331(6014):176–182, 2011.

[Naz10] Nazar, R.: *A Quantitative Approach to Concept Analysis.* Tesis de Doctorado, Universitat Pompeu Fabra, 2010.

[NJ10] Nazar, R. y M. Janssen: *Combining Resources: Taxonomy Extraction from Multiple Dictionaries.* En *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010.

[NR12a] Nazar, R. y I. Renau: *Agrupación semántica de sustantivos basada en similitud distribucional. Implicaciones lexicográficas.* En *V Congreso Internacional de Lexicografía Hispánica*, Madrid, Spain, 2012.

[NR12b] Nazar, R. y I. Renau: *A Co-occurrence Taxonomy from a General Language Corpus.* En *Proceedings of EURALEX*, páginas 367–375, Oslo, Norway, 2012.

[NS07] Nadeau, D. y S. Sekine: *A survey of named entity recognition and classification.* Lingvisticae Investigationes, 30(1):3–26, 2007.

[NVW12a] Nazar, R., J. Vivaldi y L. Wanner: *Automatic Taxonomy Extraction for Specialized Domains Using Distributional Semantics.* Terminology, 18(2):188–225, 2012.

[NVW12b] Nazar, R., J. Vivaldi y L. Wanner: *Co-occurrence graphs applied to taxonomy extraction in scientific and technical corpora.* Procesamiento del Lenguaje Natural, (49):67–74, 2012.

[Phi85] Phillips, M.: *Aspects of text structure: an investigation of the lexical organisation of text.* North-Holland, 1985.

[RN11] Renau, I. y R. Nazar: *Propuesta metodológica para la creación automática de patrones léxicos usando el Corpus Pattern Analysis.* En *Proceedings of the 27th Conference of the Spanish Society for Natural Language Processing*, Huelva, Spain, 2011.

[RN12] Renau, I. y R. Nazar: *Hypernym extraction by definiens-definiendum co-occurrence in multiple dictionaries.* Procesamiento del Lenguaje Natural, (49):83–90, 2012.

[Sin91] Sinclair, J.: *Corpus, Concordance, Collocation.* Oxford: Oxford University Press, 1991.

[SP97] Schütze, H. y J. Pedersen: *A co-occurrence-based thesaurus and two applications to information retrieval.* Information Processing and Management, 33(3):307–318, 1997.

[WT10] Wible, D. y N. Tsao: *StringNet as a Computational Resource for Discovering and Investigating Linguistic Constructions.* En *The NAACL HLT Workshop on Extracting and Using Constructions in Computational Linguistics*, Los Angeles, USA, 2010.