EL CONOCIMIENTO LINGÜÍSTICO EN LA DESAMBIGUACIÓN SEMÁNTICA AUTOMÁTICA

Iulia Nica

A mi familia

Índice

Introducción

Línea de estudio. Este trabajo de investigación se inscribe en el marco de la Lingüística Computacional y, más en concreto, en el área de la Semántica Computacional. El objetivo de la semántica computacional es la interpretación semántica automática de los textos. Para ello, una tarea básica es la asociación de los sentidos correspondientes a cada una de las palabras de un texto, o sea la Desambiguación Semántica Automática (DSA). A la vez, la identificación de los sentidos supone unos fundamentos teóricos sobre el significado de las palabras, con lo cual nuestro estudio es tangencial a la semántica léxica.

Motivación. La polisemia es un problema controvertido para la comprensión del lenguaje y constituye uno de los temas centrales de estudio tanto en lingüística teórica como en lingüística computacional, dentro de teorías y orientaciones variadas. A pesar de la vasta bibliografía dedicada al respecto, la polisemia sigue siendo un problema teórico de difícil solución (Ravin y Leacock, 2000).

Últimamente, se resalta el diálogo necesario entre la lingüística teórica y la lingüística computacional: por un lado no es posible un progreso significativo en los aspectos computacionales de la polisemia sin avances serios en las cuestiones teóricas; por otro lado, la labor teórica puede beneficiarse de los resultados de la lingüística computacional y a la vez encontrar su comprobación en las aplicaciones del procesamiento del lenguaje natural (Pustejovsky y Boguraev, 1996; Ravin y Leacock, 2000). Nuestra investigación va encaminada a responder a esta necesidad de aproximación entre las perspectivas empírico-computacional y teórica en el estudio del significado y de la polisemia.
En el procesamiento del lenguaje natural, la polisemia se considera desde hace medio siglo como el mayor problema por resolver y las competiciones SENSEVAL 1, 2 y 3 de sistemas de DSA han revelado la inmensa dificultad de la tarea (Kilgarriff y Palmer , 2000; *Proceedings of SENSEVAL-2*, 2001; *Proceedings of SENSEVAL-3*, 2004). La investigación en la DSA, que se ha basado sólo en parte en los estudios de la lingüística teórica, y en particular en las teorías del sentido de la semántica léxica, debería tener una visión más consistente con la teoría lingüística reciente (Ide y Véronis, 1998).

La falta de resultados satisfactorios en la Desambiguación Semántica Automática reclama, por una parte, una incorporación mayor de conocimiento lingüístico en este proceso. Dicho conocimiento debería guiar la determinación de la información necesaria para la Desambiguación Semántica Automática que las fuentes léxicas deban contener, su representación y organización, así como también el refinamiento de los algoritmos ideados para esta tarea.

Las limitaciones de la tarea de Desambiguación Semántica Automática tienen a la vez implicación en el plano teórico: han determinado que se hayan replanteado algunas cuestiones que atañen a la polisemia y han hecho patente la necesidad de investigar en la línea de definir unos fundamentos teóricos y metodológicos para el tratamiento de los sentidos. A base de la experiencia acumulada en las técnicas de Desambiguación Semántica Automática y de las dificultades encontradas, se constata la necesidad de redefinir el concepto de sentido sobre unas bases lo más objetivas posibles, problema directamente relacionado con el de la construcción de fuentes de conocimiento léxico adecuadas para las aplicaciones del PLN.

Por otra parte, se constata la necesidad de un cambio radical de paradigma en la semántica léxica: la sustitución de la intuición en que se funda buena parte de la investigación por la evidencia ofrecida por los datos. La falta de evidencia objetiva hace imposible que la semántica cumpla el requisito de refutabilidad impuesto a las teorías científicas desde Popper. Pero el cambio de paradigma en la semántica léxica no es sólo una necesidad teórica sino también práctica. Las grandes discrepancias entre los diccionarios demuestran que en la lexicografía no existen criterios objetivos para la discriminación y agrupación de los sentidos. El tratamiento de la polisemia, implícito en las fuentes

léxicas, influye considerablemente en el diseño y en los resultados de los sistemas de Desambiguación Semántica Automática.

Una posición de bastante relevancia en el área de la Desambiguación Semántica Automática es la de quienes sostienen la falta de concordancia entre el tipo de conocimiento sobre los sentidos ofrecido por las fuentes léxicas y el necesario para desambiguar ocurrencias en el texto. Recordamos, en este contexto, propuestas como las de Kilgarriff (1997) y de Véronis (2001) para reorientar la lexicografía hacia las necesidades de la desambiguación semántica y definir los sentidos a partir de los usos de las palabras en los textos. Es decir, idear un modelo del significado y de fuente léxica adecuado para tareas del Procesamiento del Lenguaje Natural. La inherente dificultad de discriminar los sentidos de las palabras y la falta de criterios claros y objetivos, en la definición y delimitación de los sentidos, además de garantizar, un nivel de granularidad adecuado, son problemas claves de la DSA (Ide y Véronis, 1998; Ide, 2000; Palmer, 1998).

Una opinión creciente en la comunidad computacional es que el contexto desempeña un papel central en la resolución de la polisemia y por ello tiene que ser parte integrante de su solución (Ravin y Leacock, 2000). Varios experimentos desarrollados en este ámbito para entender mejor el papel que el contexto desempeña en la identificación del sentido adecuado de una palabra confirman esta opinión[1].

Objeto de estudio. Hasta el momento, el énfasis en los sistemas del tratamiento computacional del significado se ha centrado en el desarrollo y mejora de los algoritmos que asignan sentidos a cada una de las palabras del texto. Se trata de los algoritmos de Desambiguación Semántica Automática (DSA). Nuestro trabajo se va a centrar en los aspectos lingüísticos del problema.

Objetivos. En la presente tesis, nos proponemos dos objetivos principales, relacionados con dos líneas de la investigación. Desde la perspectiva de la Desambiguación Semántica Automática, miramos hacia la mejora cualitativa del nivel de DSA, específicamente la mejora de la precisión, e intentamos demostrar que la información lingüística permite este salto cualitativo. Vemos la mejora de la cobertura más bien como desarrollo posterior a este estudio prototípico. Desde la Semántica Léxica, nos interesa aportar evidencia empírica sobre los diferentes factores que influyen en el significado de una palabra en uso. En particular, investigamos los elementos del contexto que establecen una relación sintáctica con la palabra ambigua y las diferentes relaciones léxico-semánticas que los sentidos de la palabra entretienen en la estructura del vocabulario. Si tal investigación logra arrojar más luz sobre los mecanismos que regulan la relación entre el significado de la palabra y el significado del sintagma, esperamos que sea una contribución hacia la mejor comprensión de la competencia semántica[2] como conocimiento tácito e inconsciente, difícilmente accesible a la introspección.

Hipótesis de partida. Las limitaciones actuales en la Desambiguación Semántica Automática se deben en parte al uso de fuentes léxicas creadas para otros propósitos y por lo tanto no del todo adecuadas para esta tarea (Kilgarriff, 1997, 1998; Véronis, 1999, 2002; Agirre y Martínez, 2001a) y en general a una insuficiente explotación de la información lingüística. El uso intensivo del conocimiento lingüístico podría llevar a una mejora sensible en la calidad de Desambiguación Semántica Automática, con lo cual se pone de manifiesto la necesidad de integrar más conocimiento lingüístico en los sistemas de DSA (Manning y Schütze, 1999; Corazzari *et al*., 2000). Esta posición encuentra confirmación y justificación en algunos experimentos que ponen en evidencia la contribución más importante, en el proceso de DSA, de los atributos informativos frente a la de los algoritmos usados (Pedersen, 2002). El conocimiento lingüístico es necesario para ajustar y adecuar los elementos que participan en el proceso de Desambiguación Semántica Automática (fuentes de conocimiento léxico, algoritmos, texto por desambiguar al nivel de sentidos) a las necesidades de la tarea, y también para sistematizar y guiar el proceso en su conjunto.

[1] Cf. *Computational Linguistics*, **24 (1)**, 1998.

[2] El conocimiento de los significados de las unidades léxicas y de las reglas por medio de las cuales se combinan dichas unidades (Escandell Vidal, 2004: 23).

Enfoque empírico. En la actual investigación, hemos optado por no alinearnos con ningún marco teórico concreto del estudio del significado, aunque hemos tenido presentes resultados de la Semántica teórica y computacional. Nos basamos en la hipótesis de que el contexto modula el significado y que el contexto inmediato es relevante para la desambiguación. El estudio tiene un profundo carácter empírico y se basa en el análisis de corpus textual y en el uso de técnicas de DSA. Partimos de la idea que ambas modalidades de investigación empírica pueden aportar evidencia para el mejor conocimiento del significado de las palabras en uso. Hemos desarrollado la indagación empírica según resultados básicos de la teoría semántica, con lo cual el enfoque empírico interacciona con la perspectiva teórica.

Metodología y desarrollo de la investigación. Nuestra aproximación pretende aunar los trabajos de carácter teórico, por una parte, con el estudio de carácter empírico basado en el etiquetado semántico manual y automático, mediante el análisis de casos sobre corpus y la experimentación. La trayectoria que ha seguido esta investigación parte de la perspectiva teórica de diferentes enfoques sobre el estudio del significado, se desarrolla en el nivel empírico de la DSA y retorna al plano teórico con análisis y conclusiones de carácter lingüístico.

a) En el *plan teórico*, nos proponemos ante todo realizar el estado de la cuestión en el estudio y el tratamiento del significado en el marco de la DSA. Tenemos en cuenta las perspectivas de la lexicografía y de la teoría semántica como también de la Desambiguación Semántica Automática. Nos interesan los problemas que el significado plantea, puestos en evidencia por los estudios teóricos o por la labor de etiquetado semántico, y las dificultades que las propuestas actuales encuentran en solucionar dichos problemas. Para el caso específico de la Desambiguación Semántica Automática (DSA), damos una relación detallada de las fuentes léxicas usadas, del tipo de información relacionada con el significado que se ha utilizado y de la manera en que ésta ha sido explotada.

b) El *estudio empírico* está centrado en la incorporación de conocimiento lingüístico al proceso de DSA. Desarrollamos este estudio en varias fases, en que alternamos el análisis, la observación manual sobre el corpus y los experimentos automáticos con la correspondiente evaluación.

El estado de la cuestión en DSA nos permite realizar un análisis de los puntos críticos de la tarea. En base a este análisis, proponemos soluciones para las limitaciones señaladas desde una perspectiva lingüística, lo que nos lleva a delinear una estrategia diferente a la DSA: la desambiguación de una palabra ambigua se realiza considerándola integrada en cada uno de sus patrones léxico-sintácticos y no aisladamente. La integración de la palabra por desambiguar en patrones léxico-sintácticos se funda en la hipótesis de la "tendencia hacia un único sentido por patrón léxico-sintáctico" y en las siguientes dos reducciones: la contribución de los patrones léxico-sintácticos al sentido de la ocurrencia ambigua es independiente del resto de la oración (R1) y cada patrón léxico-sintáctico actúa sobre el significado de la ocurrencia ambigua independientemente de los otros patrones (R2). A partir de esta integración, se extrae información paradigmática del corpus relacionada con la ocurrencia ambigua, sobre la cual se aplica un algoritmo de DSA (la Marca de Especificidad (Montoyo, 2002)). Hemos realizado un estudio de caso de esta variante básica del método de DSA sobre ocurrencias del nombre *órgano* en el corpus de entrenamiento de Senseval-2. Este estudio nos ha revelado algunas limitaciones de nuestra propuesta, para cuya solución hemos mejorado el método, obteniendo una serie de heurísticas. La modalidad efectiva para la combinación de las heurísticas en un sistema de DSA complejo se establece por vía empírica, a través de la evaluación de las variantes. Las evaluaciones indican las variantes idóneas tanto para las heurísticas individuales como para el sistema de DSA en conjunto.

Además de los experimentos para el refinamiento del método, comprobamos, a través de la experimentación, la idoneidad de las dos reducciones adoptadas en nuestro método; los resultados confirman nuestra elección. También comprobamos la hipótesis de la "tendencia hacia un único sentido por patrón léxico-sintáctico", con resultados positivos.

El interés lingüístico en nuestra investigación nos determina realizar unos experimentos de carácter comparativo. Así, por una parte contrastamos los dos algoritmos de DSA, la Marca de Especificidad y la Prueba de Conmutabilidad, que hacen un uso distinto de la información relacional léxico-semántica de *EuroWordNet* (Vossen, 1998); por otra parte, comparamos la contribución de la información paradigmática y de la información sintagmática al proceso de DSA.

c) El estudio empírico y la experimentación desarrollados nos servirán de base para unas *conclusiones de órden teórico*. Por una parte, proponemos dos modelos de fuentes léxicas útiles para la DSA: Discriminadores de Sentido y una base de patrones léxico-sintácticos etiquetados con sentidos. Por otra parte, se emite la hipótesis de que el sentido de una palabra en contexto está principalmente determinado por sus relaciones sintácticas, antes que por la oración en su conjunto.

Estructura de la tesis. Organizamos la exposición en dos partes principales: en la primera, delineamos el marco de trabajo, la DSA, y sentamos las bases teóricas y metodológicas para la investigación que se presenta en la segunda parte. Así, la primera parte contiene una introducción a la problemática de la DSA (capítulo 1), una síntesis de los enfoques fundamentales al estudio del significado en la DSA (capítulo 2) y un estado de la cuestión en la metodología de la DSA (capítulos 3, 4, 5). La segunda parte se dedica a nuestra propuesta de DSA basada en la explotación de los patrones léxico-sintácticos de la ocurrencia ambigua. En el capitulo 6, se presentan el método de DSA, la metodología seguida y el desarrollo del mismo. En el capitulo 7, se detalla la experimentación realizada utilizando este método y se analizan los resultados de los distintos experimentos. Concluimos el trabajo con una visión de conjunto sobre los desarrollos en curso y futuros (capítulo 8) y con las conclusiones, en que reunimos los problemas abiertos en DSA y una síntesis de las aportaciones de la presente tesis.

I FUNDAMENTOS

1 La cuestión de la Desambiguación Semántica Automática (DSA)

El Procesamiento del Lenguaje Natural (PLN) encuentra uno de los mayores obstáculos en la ambigüedad del lenguaje, a todos los niveles de las lenguas: fonológico, morfológico, sintáctico, léxico-semántico, pragmático. Cualquier tarea de PLN implica de algún modo un proceso de desambiguación. La Desambiguación Semántica Automática es un caso específico orientado a la resolución de la ambigüedad al nivel léxico, específicamente la ambigüedad determinada por la polisemia de las palabras.

1.1 Definición de la DSA

La tarea de la Desambiguación Semántica Automática (DSA; en inglés, *Word Sense Disambiguation*, *WSD*)[3] consiste en asignar a las palabras el sentido con el cual se están usando en un texto[4].

Recordamos, en el contexto de la definición de la desambiguación semántica, unas distinciones necesarias para la delimitación del área de la DSA. Según Yarowsky (2000b), la DSA corresponde sólo a una de las variadas tareas de la solución de la ambigüedad léxica, que incluye también: generación texto-habla, clasificación de entidades discursivas, corrección ortográfica, etc. Por lo tanto, las denominaciones alternativas para la DSA, *desambiguación léxica* (en inglés, *lexical disambiguation*) o *resolución de la ambigüedad léxica* (en inglés, *lexical ambiguity resolution*) son demasiado amplias. En la presente tesis, usamos también estos dos términos, exclusivamente en la acepción estricta de 'DSA'. La DSA se puede ver como un proceso de asociación de información semántica a los textos. En este proceso, se suelen delimitar dos niveles de análisis y por consiguiente de etiquetado (Corazzari *et al*., 2000, siguiendo a Kokkinakis *et al*., 1999): *etiquetado con sentidos* (en inglés, *sense tagging* o *sense labelling*), es decir la asignación a las ocurrencias (*tokens*) de un texto del sentido adecuado con respecto de una fuente, y *etiquetado* o *anotación 273.48 -5 11.04Tw 363n estr2ooes0l a) de uj 3*

En nuestra tesis, siguiendo la distinción entre etiquetado con sentidos y etiquetado semántico, entendemos por DSA la asociación de sentidos a las palabras del contexto.

Para el etiquetado con sentidos, se usan comúnmente inventarios con sentidos provistos por fuentes léxicas estructuradas. En algunos casos, en vez de sentidos predefinidos de una fuente léxica, se usan clases construidas de modo automático a partir de corpus: las diferentes ocurrencias de una palabra se agrupan en clases (en inglés, *clusters*) en función de la similitud de los contextos en que ésta aparece. De aquí otra distinción que se hace en relación con la DSA: *etiquetado con sentidos* (en inglés, *sense labelling*), si se usan etiquetas explícitas y predefinidas de sentido, y *discriminación de sentidos* (en inglés, *sense discrimination*), si se usan clases derivadas del corpus, que corresponden a sentidos, y no sentidos, sin que se realice un etiquetado.

La distinción discriminación de sentidos *vs.* etiquetado con sentidos se encuentra relacionada con dos modalidades de entender la desambiguación semántica automática: la *DSA débil* y la *DSA fuerte*, respectivamente (Manning y Schütze, 1999). La opción entre una u otra variante depende de la aplicación secundaria de la tarea de WSD. En algunas aplicaciones, como la recuperación de información (RI; en inglés, *Information Retrieval*, IR), es suficiente la partición no etiquetada entre sentidos (DSA débil), mientras que para la gran mayoría (por ejemplo, la traducción automática) es necesario establecer una correspondencia con un inventario establecido de sentidos (DSA fuerte) (cf. Yarowsky, 2000b).

Recordamos, de paso, una línea de investigación que, aunque no incluida comúnmente en el dominio de la DSA, comparte con ésta la misma esfera de interés: el reconocimiento y la clasificación de entidades nombradas (en inglés, *named-entity recognition* y *named-entity classification*) que es básicamente análisis y desambiguación semántica de entidades discursivas (Yarowsky, 2000b).

1.2 Motivación de la DSA: necesidad y utilidad

La desambiguación semántica no es un fin en sí misma, sino que es una etapa necesaria para realizar varias tareas del PLN, como son el análisis sintáctico o la interpretación semántica, y para el desarrollo de aplicaciones finales. Obviamente identificar los sentidos adecuados de las palabras de un texto es esencial para tareas y aplicaciones que suelen incluir un proceso de comprensión del lenguaje: la interpretación de mensajes, la comunicación hombre-máquina, la extracción y clasificación de información sobre entidades específicas; traducción automática; recuperación de información y navegación hipertextual; el análisis temático y de contenido; procesamiento del habla; procesamiento del texto, etc.

Recordamos que la aparición misma de la DSA como línea de investigación se ha producido en el campo de la Traducción Automática (TA), que ha encontrado en la polisemia de las palabras su mayor dificultad (Weaver, 1955; Yngve, 1955).

1.3 Problemas que plantea la DSA

La DSA es una tarea no trivial (Suderman, 2000); toda la bibliografía sobre el tema resalta la dificultad de la tarea, lo que explica el carácter todavía no solucionado de la misma: la DSA es el gran problema abierto al nivel léxico del PLN (Resnik e Yarowsky, 1997). Entre las cuestiones teóricas y prácticas que plantea la DSA, cabe destacar:

- la existencia misma de los sentidos y de aquí la factibilidad del problema de la DSA;
- la definición de sentido;
- el número de sentidos para una palabra dada y la granularidad de los sentidos, es decir, la discriminación de sentidos;
- el lugar de la DSA dentro del PLN;
- los recursos o fuentes de información necesarios (y la cuestión implícita, el llamado “cuello de botella de la adquisición de conocimiento”; en inglés, *knowledge acquisition bottleneck*);
- el alcance de la DSA: general o restringida a dominios específicos (en este último caso, la dificultad de adaptación léxica; en inglés, *lexical tuning*);
- la cobertura, limitada a un unas pocas palabras seleccionadas o a todas sin restricción alguna; en inglés, *lexical-sample vs. all-words*;
- la evaluación de la DSA.

Por tales razones, el concepto mismo de *desambiguación* se considera poco preciso (Wilks *et al.*, 1996). Sobre algunos de estos problemas y sus implicaciones para la DSA intentaremos detenernos en la presente tesis, inventariando varios puntos de vista y soluciones propuestas.

1.4 Tarea intermediaria, no independiente

Si la DSA es una "tarea establecida" en el enfoque de base empírica del PLN (Basili *et al.*, 1998), Wilks (2000) añade que no es una tarea independiente. Está comúnmente aceptado que la DSA representa una "tarea intermedia" del procesamiento del lenguaje natural y no un fin a sí misma (cf. Wilks y Stevenson, 1997). Se trata de un proceso necesario para la realización de otras tareas del PLN, como hemos visto anteriormente. Wilks señala los límites de la DSA vista como tarea independiente: a diferencia del etiquetado morfosintáctico[7], los sentidos nuevos no se pueden prever. Es decir, al etiquetado semántico no se le puede aplicar el paradigma empírico estándar: marcar, modelizar/entrenar, y comprobar (en inglés, *mark-up, model/train, and test model*). Por lo tanto, Wilks se pronuncia a favor de un cambio de estrategia: no escindir el PLN en varias tareas parciales, para luego modelarlas y evaluarlas de manera autónoma, sino integrar las tareas en sistemas amplios, proyectados para desarrollar aplicaciones reales. Su conclusión es que la DSA no es una tarea parcial o autónoma más dentro del PLN. Es la perspectiva por la cual también se pronuncian Ide y Véronis (1998): como tarea independiente, la DSA es difícil y quizás imposible de valorar, de fijar en abstracto, por lo cual, es necesario incorporar los métodos de DSA en un marco más amplio. Podemos concluir que la DSA tiene que integrarse en el PLN general, aunque, de momento, se obtienen resultados interesantes usando sólo los enfoques más simplistas, limitados al aspecto práctico del problema (Agirre, 1998).

Visto que la DSA es una etapa intermedia, necesaria para otras tareas del PLN, como TA, RI, se plantea la cuestión de si la DSA tiene carácter preliminar a tales tareas o bien hay una interacción entre ésta y las aplicaciones. Palmer (2000) se pronuncia a favor de una interacción entre la DSA y las aplicaciones. Por ejemplo, algunos sistemas de RI que usan técnicas estadísticas realizan buena parte de la DSA. O bien podemos recordar que uno de los criterios invocados para la discriminación de los sentidos de una palabra es su equivalencia en otras lenguas con más unidades léxicas. Según Basili *et al*. (1998), hay evidencias sustanciales de que se aplica alguna forma de DSA en todas las tareas del PLN. En esta línea, se observa últimamente la tendencia a diseñar sistemas no para resolver exclusivamente la DSA, sino para realizar la DSA junto con otras tareas o aplicaciones.

1.5 La DSA y el PLN en otros niveles de la lengua

Si aceptamos la visión sobre el PLN como sucesión de análisis en varios niveles, ¿dónde se coloca la DSA? Como tarea que se ocupa de la ambigüedad léxica, pertenece al nivel de la interpretación semántica, como ya hemos mencionado. Pero el lugar mismo de la interpretación semántica en esta secuencia es objeto de discusión: ¿después del análisis sintáctico[8], o, junto con los factores pragmáticos, durante este proceso? Si la interpretación semántica ocurre enteramente después del análisis sintáctico, entonces tendrá que rechazar una multitud de posibles análisis; por eso, es más atractivo reunir parte del análisis semántico y pragmático para que actúe durante el análisis sintáctico[9].

Esta posición se aproxima al enfoque distributivo, la alternativa a los tradicionales sistemas de PLN con arquitecturas secuenciales, fáciles de usar y mantener, pero con limitaciones en el desarrollo del proceso. Este enfoque se basa en la cooperación entre los diferentes módulos autónomos y especializados[10]. En el caso concreto de la DSA, para una desambiguación léxica superior, son necesarios conocimientos conceptuales y estructurales, es decir hace falta una integración de niveles múltiples de conocimiento lingüístico.

---

[7] A lo largo de la exposición, usaremos igualmente el término inglés *POS tagging* para referirnos al etiquetado morfosintáctico.

[8] De aquí en adelante usaremos indistintamente análisis sintáctico o *parsing*, la terminología consacrada en la bibliografía de lengua inglesa.

[9] Cf. Asher (ed.) (1996), *s.v. Disambiguation: Role of Knowledge.*

[10] Para referencias, v. Tavares da Silva y Strube de Lima (1997).

DSA y el etiquetado morfosintáctico. Una palabra se puede usar como perteneciendo a varias categorías morfosintácticas. ¿Cómo se relacionan la DSA y el etiquetado morfosintáctico? Usar una palabra como categorías morfosintácticas distintas es claramente un uso diferente, con una significación implicada diferente. En la práctica, las dos cuestiones se han de distinguir, en parte por las diferencias entre la naturaleza del problema, en parte por los métodos desarrollados para resolverlos y por las informaciones necesarias (Manning y Schütze, 1999). Generalmente, se acepta que, en el caso de las palabras ambiguas, las partes de oración distintas implican sentidos distintos, por lo cual se delimitan los problemas de la desambiguación semántica y el de la desambiguación morfosintáctica, en otras palabras, se separan el etiquetado con sentidos y el etiquetado morfosintáctico. El problema de la DSA queda enfocado a los homógrafos con la misma categoría sintáctica. Desde este punto de vista, podemos decir que el etiquetado morfosintáctico resuelve una buena parte de la labor de DSA. Según Wilks y Stevenson (2000), el etiquetado morfosintáctico desambigua el 87% de las ocurrencias de las palabras ambiguas y reduce la ambigüedad para una parte de los restantes, si se considera la ambigüedad en un sentido amplio, incluyendo la ambigüedad categorial. Por lo cual, el etiquetado morfosintáctico tiene que preceder a la DSA en un sistema de PLN, es una etapa previa necesaria.
En la presente tesis consideramos la DSA limitada a la identificación de sentidos de una palabra con la misma categoría morfosintáctica, eventualmente divididos entre varios homónimos de la palabra.

1.6 Existencia y factibilidad de la DSA

Una pregunta fundamental dentro de la DSA, que cuestiona la tarea misma, es si es posible la *anotación semántica humana*. La importancia de la cuestión es obvia: si los humanos no tienen esta habilidad, es inútil intentar automatizarla. Según Wilks (2000), hay dos posiciones contrarias acerca de este tema.
La primera, ampliamente compartida y fundada en evidencias empíricas, es una posición positiva, que admite la existencia de la anotación semántica humana. Como tal, las máquinas también tienen que procurar hacerla: sí, se puede hablar de DSA dentro del PLN.
Los contraargumentos (p.ej., Kilgarriff, 1993, 1997) se formulan según dos críticas principales: la insuficiencia o incapacidad de los sentidos definidos en las fuentes léxicas para cubrir los usos de las palabras en los textos, y la vaguedad inherente para diferenciar entre los sentidos de una palabra dada. Wilks (2000) desarma estas críticas desde un punto de vista más bien teórico y las considera consabidas y asumidas en la labor de DSA.
El problema está identificado en (Gale *et al.*, 1992) como el límite superior (en inglés, *upper bound*) en la actuación (en inglés, *performance*) de los programas de DSA, y está definido como el nivel del etiquetado con sentidos realizado por los humanos. Los resultados contradictorios obtenidos en diferentes experimentos han contribuido a mantener cierta incertidumbre alrededor de la cuestión del etiquetado humano y su nivel real. Así, algunos experimentos indican límites bajos, del 68% (Jorgensen, 1990), o incluso del 57% (Ng y Lee, 1996), en cuyo caso la DSA sí sería una empresa destinada a fracasar. Sin embargo, la relevancia de tales resultados es limitada, dado que los experimentos no se propusieron maximizar el nivel de los resultados. Un intento de identificar este nivel es el estudio de Kilgarriff (1999), dentro de la competición Senseval de 1998, en directa relación con el establecimiento del estándar (*gold standard*) para la evaluación de los sistemas de DSA competidores. El estudio se llevó en unas condiciones que el autor considera esenciales para el fin propuesto, a saber, la obtención del más alto nivel de replicabilidad posible. Estas condiciones son las que siguen:

- la definición precisa de la tarea de desambiguación;
- la calificación idónea de los lexicógrafos;
- la doble anotación;
- la fase de comparación entre los resultados individuales (en inglés, *arbitration*).

El resultado obtenido en la anotación manual, que llega al 95% de acierto, sería, en la visión de Kilgarriff, una respuesta definitiva, fundada en pruebas empíricas, a la duda sobre la validez de la DSA. En Senseval-2, el estándar manual alcanza el 85.5%.
Reafirmando la importancia del nivel del etiquetado humana para la DSA, y por consiguiente la necesidad de que sea investigada rigurosamente, Véronis (2000) organiza un experimento -ya citado-

de relevancia en este contexto. El experimento consistió en dos pruebas destinadas, una, a la identificación de los sentidos y, la otra, a evaluar el nivel del etiquetado en base a los sentidos de un diccionario. Siguiendo una práctica ampliamente difundida en la estadística, se utilizó la medida estándar kappa (*k*) para medir el acuerdo entre los participantes. Los resultados obtenidos para la primera prueba indican que la tarea no fue considerada difícil, con un sólo un 4% de respuestas "no lo sé"; la mayor parte de las palabras fueron vistas como polisémicas, pero con una diferencia significativa entre las categorías: los nombres fueron interpretados como más polisémicos que los verbos, y estos, a su vez, más que los adjetivos. El desacuerdo entre los individuos fue bajo, con una media de k = 0.49. En la segunda prueba los anotadores evitaron las respuestas múltiples; el promedio de sentidos por categoría no es muy alto, pero sí difiere de una categoría a otra. La medida *k* varió entre 0.92 y 0.07, con valores medios por debajo del 50%, con lo cual es desacuerdo entre los jueces es considerable. Además, el intento de aclarar las posibles causas del desacuerdo relacionadas con el diccionario lleva a la conclusión de que no es una cuestión de granularidad en las distinciones de sentido, sino de falta de información sintagmática.

Esta observación, opinamos, se traduce en cierta relatividad en los resultados obtenidos en este experimento: el bajo acuerdo entre los jueces no es una prueba contundente, irrefutable en contra de la anotación humana. Otra conclusión que se impone es que los futuros experimentos sobre el nivel de la anotación humana se deben sentar sobre bases distintas, probablemente tomando un inventario de sentidos construido en base a criterios distribucionales.

La propuesta de Véronis (2000) se materializa en los experimentos de anotación manual desarrollados dentro de la preparación del ejercicio Senseval-3 para el español y el castellano (Taulé *et al*., 2004). La experimentación es importante bajo múltiples aspectos. Primero, se elaboran fuentes léxicas específicas para DSA. Segundo, se desarrolla una metodología para la evaluación de las fuentes léxicas en cuanto su utilidad para la DSA. Se parte de la idea que un buen grado de acuerdo entre los anotadores es índice de la calidad de la fuente usada en la anotación. Tercero, se aportan pruebas a favor de la información más aprovechable para la tarea de DSA que estas fuentes contienen.

En los experimentos, se etiquetan 800 ocurrencias de las palabras de Senseval-3 (en ejemplos del corpus EFE) con tres fuentes léxicas:

1) el DRAE (*Diccionario de la Real Academia Española*, XX), un diccionario de uso general, de referencia para el español;

2) el *MiniDir.2.1*. (el diccionario especialmente construido para Senseval-3, según los siguientes criterios: granularidad reducida, disyunción de los sentidos, información de tipo colocacional);

3) el *Modelo Véronis* (entradas elaboradas según el modelo de Véronis para la estructuración de las entradas, o sea usando información sintáctica, paradigmática (hiperonimia y sinonimia) y de coocurrencia (colocaciones)).

Cada ocurrencia se etiqueta por tres anotadores y la comparación entre las diferentes anotaciones se hace según varios grados de acuerdo: total (cuando las tres anotaciones son iguales), parcial (cuando no hay coincidencia total, pero hay solapamiento común entre todas las anotaciones), mínimo (cuando coinciden dos de los tres etiquetados) y desacuerdo (cuando no hay ninguna coincidencia entre las anotaciones). Además, se consideran otros criterios para la medición del acuerdo entre los anotadores: acuerdo total mínimo (que cuenta todos los casos de acuerdo total entre los anotadores), acuerdo total máximo (que cuenta todos los casos de acuerdo total y parcial entre los anotadores), acuerdo bilateral mínimo (que cuenta los casos de acuerdo total entre cada dos anotadores) y acuerdo bilateral máximo (que cuenta los casos de acuerdo parcial entre cada dos anotadores). Respecto de los resultados, el grado de acuerdo entre los anotadores es mayor en el caso de las dos fuentes creadas específicamente para fines de DSA, *MiniDir.2.1*. y el *Modelo Véronis*, respecto a la fuente de uso común, DRAE, demostrando así su idoneidad para tareas de DSA. Así, el acuerdo total mínimo obtenido es el siguiente, en el orden DRAE, *Minidir.2.1*., *Modelo Véronis*: para nombres, 51%, 86% y 95%; para adjetivos, 57%, 83%, 99%; para verbos, 64%, 81%, 95%. Mientras que el acuerdo total máximo obtenido es el siguiente, en el mismo orden *DRAE*, *Minidir.2.1*., *Modelo Véronis*: para nombres, 76%, 88% y 96%; para adjetivos, 74%, 88%, 1%; para verbos, 70%, 84%, 95%. Los resultados muestran un alto grado de acuerdo entre los anotadores para *MiniDir.2.1.* y el *Modelo Véronis*.

La variabilidad del acuerdo entre los anotadores humanos demuestra que éste depende en gran medida de la calidad de las fuentes léxicas usadas, de la delimitación de los sentidos que se haya realizado y de

la presencia de discriminadores explícitos de sentido. A la vez, la experimentación confirma las observaciones de Véronis (2000) de que es posible alcanzar un alto nivel en la anotación humana.
La existencia y la factibilidad de la DSA están relacionadas, por otra parte, con la *existencia de los sentidos* ya que, si los sentidos no existen, no sirve de mucho intentar "desambiguar". Atkins (cf. Hanks, 2000), seguida por Kilgarriff (1997), tienen una posición escéptica respecto de su existencia. Como hemos anticipado, para Kilgarriff el concepto de sentido no es suficientemente definido y coherente para que pueda constituir una unidad básica de significado utilizable en la investigación de DSA. Su posición se fundamenta en un experimento llevado a cabo con humanos, en que buena parte de las ocurrencias de palabras en un corpus no tenían cobertura con sentidos por parte del diccionario *LDOCE*. Con lo cual, propone definir los sentidos de las palabras a base de grupos (en inglés, *clusters*) de contextos del corpus (en inglés, *corpus citations*). Wilks (1997) se opone a Kilgarriff, aportando como prueba los resultados obtenidos ya en la DSA que demostrarían que la tarea es factible. Hanks (2000) replantea el problema a base de la evidencia ofrecida por los corpus, optando por una respuesta positiva: los sentidos existen, pero su descripción tradicional es engañosa. Se acerca, en esto, a las conclusiones de Véronis.
Sin embargo, las reservas mencionadas con respecto a la factibilidad de la DSA no impiden que el proceso siga en el foco de interés de la comunidad del PLN. Los avances en este área demuestran que la DSA es una tarea real, más allá de las limitaciones actuales.

2 El significado desde la DSA

El objetivo de este capítulo es sentar las bases teóricas que tomaremos como punto de partida en la presente tesis. Presentamos cuestiones relevantes para la DSA, enfocadas principalmente desde la lingüística computacional, aunque se harán, inevitablemente, referencias también a la lingüística teórica, sobre todo a la semántica léxica.

Los sistemas de DSA usan, para desambiguar, lexicones o bien corpus etiquetados a nivel de sentido. La labor de DSA en su forma actual parte, por lo tanto, de dos hipótesis: 1) el significado es representable (los modelos de significado adoptados se reflejan en las fuentes léxicas utilizadas[11]); 2) la tarea es factible a partir de las fuentes léxicas existentes.

La adopción de un inventario de sentidos en términos del cual se realice la tarea significa que en la base de la DSA hay una determinada concepción del significado y, por lo tanto, un determinado tratamiento de la polisemia[12]. El presente capítulo se centra en este tema.

En primer lugar, es necesario delimitar qué se entiende por polisemia (apartado 2.1.). Nos detenemos, posteriormente, en su tratamiento desde el área de la DSA. Así, la asignación de los sentidos de manera automática y en condiciones variables[13] hace que la DSA plantee además cuestiones relacionadas con el modelo adecuado para la representación de los sentidos, con el número y la delimitación de los sentidos y con su dependencia de las aplicaciones del PLN (apartado 2.2.). Seguimos con algunas consideraciones de la semántica teórica de alta relevancia para la DSA, referentes a la relación entre significado y contexto (apartado 2.3.) y la construcción del significado (apartado 2.4.). Estas consideraciones se asumen en la tarea de DSA, perfilando sus coordenadas lingüísticas (apartado 2.5.).

2.1 Delimitación de la polisemia

La *polisemia* es la propiedad de determinados ítems léxicos de tener más de un sentido (Cruse, 1986; Asher (ed.), 1994, *s.v. polisemy*).

Para el justo entendimiento de este problema, se impone una doble delimitación del fenómeno: 1) por una parte, la delimitación de la polisemia estricta con respecto a la ambigüedad y a la homonimia, y también con respecto a otras maneras de significar: la vaguedad y la indeterminación (subapartado 2.1.1.); 2) por otra parte, la correcta circunscripción de la polisemia entre los varios tipos de multiplicidad de significado (subapartado 2.1.2.). Este análisis se realiza desde la perspectiva de plantear la problemática de la delimitación de sentidos en la elaboración de fuentes léxicas, un recurso imprescindible para llevar a cabo DSA en cualquiera de sus modalidades. Dependiendo de la calidad de las fuentes léxicas usadas, se obtendrán mejores o peores resultados en la DSA. Se impone, por tanto, una reflexión sobre lo que es un sentido o simples matizaciones de un sentido.

2.1.1 Polisemia vs. ambigüedad, homonimia, vaguedad, indeterminación

Las expresiones lingüísticas pueden tener más de una interpretación en el contexto de uso; el término técnico para este fenómeno es *ambigüedad* (Asher (ed.), 1996, *s.v. ambiguity*). La noción de

[11] Ver el capítulo 3 sobre una presentación de las fuentes léxicas usadas en la DSA: diccionarios en formato electrónico, redes semánticas, lexicones computacionales, etc.

[12] Recordamos, en este ámbito, la aseveración de que no existe polisemia en el contexto, la polisemia es una invención de los lexicógrafos; las personas que comunican no tienen generalmente problemas de incomprensión de las palabras en el uso (Wilks, Slator y Guthrie, 1996).

[13] Nos referimos aquí a la variabilidad de la tarea de DSA respecto a: la aplicación de PLN que supone o incorpora la DSA (cf. capítulo 1), los corpus usados para el entrenamiento de los métodos de DSA supervisados (v. apartado 3.1.2.2.), el dominio al cual pertenece el texto por desambiguar, etc.

ambigüedad se puede aplicar a todos los niveles del significado[14]. Nuestro interés en la presente tesis, desde la perspectiva de la DSA, se centra en la *ambigüedad léxica*, es decir la ambigüedad de las palabras[15,16]. El término ambigüedad léxica se usa para diferenciar entre la ambigüedad de una secuencia lingüística que se debe a las palabras con interpretación múltiple en el contexto y la ambigüedad que resulta de una interpretación múltiple al nivel de la construcción de la secuencia, ésta última llamada *ambigüedad estructural*. Así, en el ejemplo:

Miro al hombre de la montaña con un telescopio.

la ambigüedad es estructural, causada por la existencia de dos estructuras sintácticas diferentes que se expresan con una misma secuencia lineal de palabras. En cambio, en el ejemplo:

No me gusta esta salsa (*salsa* 'baile'/'comida').

la ambigüedad es léxica y se debe a la polisemia del nombre *salsa*.

La relación entre la ambigüedad y la polisemia corresponde a la dicotomía *langue-parole* de Saussure (1916): la polisemia es un fenómeno que pertenece al plano de la lengua, al sistema lingüístico (un mismo ítem léxico, considerado aisladamente, tiene más de un significado asociado), mientras que la ambigüedad se da en el plano de las realizaciones (una expresión tiene más de una interpretación en el marco de un texto o de una comunicación oral) y es el contexto el que comporta habitualmente la selección de uno de los significados posibles.

En el marco de las aplicaciones computacionales, se utilizan indistintamente los términos polisemia y ambigüedad (léxica) (cf. Ravin y Leacock, 2000) probablemente debido al hecho de que en DSA y en las aplicaciones de PLN se trata siempre la polisemia en un contexto. Siguiendo la práctica computacional, en la presente tesis hablaremos indistintamente de palabras ambiguas *vs.* no ambiguas, y de palabras polisémicas *vs.* monosémicas. Existe, además, una motivación de coherencia terminológica con el nombre de la tarea: *Desambiguación* Semántica Automática.

Tradicionalmente se distinguen dos tipos de relaciones de significado múltiple: polisemia y homonimia. Se da *homonimia* cuando dos ítems léxicos coinciden en la forma, pero que en su etimología y en su significado son totalmente dispares. En el caso de la *polisemia*, el mismo ítem puede tener varias interpretaciones, manteniendo todas ellas una cierta relación. Como la polisemia, la homonimia está también relacionada con el plano de la lengua. La distinción entre los dos fenómenos sigue dos coordenadas: diacrónica y sincrónica. En perspectiva diacrónica, el criterio delimitador es la etimología: la misma en el caso de polisemia y distinta para homonimia. Desde la sincronía, la diferencia es de carácter semántico: cuando los diferentes significados de un mismo ítem léxico tienen una parte común no trivial, se considera un caso de polisemia, en caso contrario, de homonimia. Sin embargo, a veces es difícil de distinguir entre polisemia y homonimia. Dependiendo del criterio de la interrelación o no entre los significados, la distinción entre homonimia y polisemia es vaga; son más bien dos extremos de una escala. Ahora bien, desde el punto de vista de la relación entre forma y significado, la polisemia y la homonimia se pueden reducir a un solo fenómeno: un mismo significante con significados distintos. Esta visión común hace que a efectos representacionales se trate de una distinción no relevante y por lo tanto ignorada; se considera simplemente que en ambos casos intervienen ítems léxicos con más de una interpretación asociada, independientemente del proceso lingüístico que ha operado.

La distinción entre homonimia y polisemia está cubierta, a veces, por terminologías alternativas. Así, Weinreich (1964), seguido por Pustejovsky (1995), habla de *polisemia contrastiva* y de *polisemia complementaria* respectivamente. Pustejovsky y Boguraev (1996), desde una perspectiva sintagmática, prefieren denominar *ambigüedad contrastiva* a la homonimia y *ambigüedad complementaria* a la polisemia.

Desde la perspectiva del modo de significar de las expresiones lingüísticas consideradas en abstracto, fuera del contexto de uso, la polisemia se diferencia de la vaguedad y de la indeterminación.

---

[14] Para una definición y presentación de los niveles del significado (significado de expresión, significado de enunciación y significado de comunicación), véase, por ejemplo, Löbner (2002).

[15] A lo largo de esta tesis, usaremos indistintamente los términos *palabra*, *ítem léxico* o *lexema*, prescindiendo de las controversias terminológicas (ver Cruse, 1986).

[16] De aquí deriva el nombre de la tarea de PLN que hace objeto de nuestro estudio: Desambiguación Semántica Automática o, en variante inglesa, Desambiguación Semántica de las Palabras (*Word Sense Disambiguation*).

La *vaguedad* corresponde a una falta de especificidad, de precisión, en el contenido de las expresiones lingüísticas, lo que se traduce en una flexibilidad en el uso, mediante la adaptación al contexto[17]. La vaguedad se puede observar con todos los conceptos que dependen de propiedades que varían en una escala continua: colores, dimensiones, atributos, etc. y además tienen carácter relativo. En general, todos los adjetivos graduables (o sea, los adjetivos con forma de comparativo y de superlativo) tienen esta característica. La diferencia entre polisemia y vaguedad consiste en que mientras que un término polisémico se asocia como mínimo a dos significados delimitados, un término vago sólo se asocia a un significado, uno menos definido.

La *indeterminación* corresponde a una falta de especificación relacionada con un determinado componente del significado. La falta de especificación suele estar relacionada con la ausencia de una marca gramatical para los valores posibles, discretos, de este componente. Es el caso de *atleta*, *gato*, etc. donde no se especifica el género (como expresión del sexo). Por lo tanto, la indeterminación se puede ver como un caso intermedio entre la polisemia y la vaguedad, debido a que su ambigüedad es parcial, está limitada a un solo componente de significado, y no es indefinida, como en el caso de la vaguedad.

En las consideraciones previas, la polisemia (y la homonimia, asimilada a la polisemia, cf. *supra*) se toma en un sentido estricto, como la existencia de varios significados, discretos, para una misma palabra. Sin embargo, el término *polisemia* se puede usar de manera genérica para cubrir las diferentes formas de significar previamente discutidas, en cuyo caso remite de manera neutra a la multiplicidad de significado. Dentro de esta acepción genérica de la polisemia, la vaguedad y la indeterminación son casos extremos, asimilados formalmente a la monosemia, con un único sentido asociado al ítem léxico. En el siguiente subapartado tratamos los diferentes tipos de polisemia con el objetivo de delimitar qué entendemos por polisemia estricta.

2.1.2 Tipología de la polisemia. Polisemia estricta. Polisemia *vs.* monosemia

Un punto de partida útil para la tipología de la polisemia es la clasificación clásica de Deane (1988), basada en el carácter fijo u ocasional de los diferentes sentidos:

- *polisemia clásica*, cuando los sentidos de un término están fijados,
- *polisemia pragmática*, cuando los sentidos son derivados de forma ocasional a partir de un proceso metafórico o metonímico.

Este último tipo no puede ser codificado en un modelo del léxico aplicado a la descripción de una lengua, ya que no se ha fijado léxicamente. La presentación se va a limitar, por lo tanto, a la polisemia clásica**.**

Para la polisemia clásica, Cruse (1986) habla de grados o tipos de polisemia asociados a los procesos de selección de sentidos *vs.* modulación de sentidos, terminología que encuentra su equivalente en (Copestake y Briscoe, 1995): extensión del significado *vs.* polisemia construccional. La polisemia por *extensión del significado* se verifica para sentidos con carácter discreto, que se pueden enumerar y diferenciar, de manera que el hablante *selecciona* en cada contexto el sentido apropiado. Se da, en cambio, *polisemia construccional* si los sentidos no se pueden delimitar *a priori* y son de difícil enumeración, por lo cual la polisemia construccional no es un caso de polisemia en sentido estricto. Algunos autores denominan este fenómeno *vaguedad* o *indeterminación* ya que se trata de un significado general que se *modula* según el contexto.

En función de las relaciones entre el significante, el significado y el referente como también en función de sus características intrínsecas, el concepto de polisemia construccional o vaguedad se puede matizar en los siguientes subtipos (Kempson, 1977):

a- vaguedad referencial: cuando resulta difícil establecer la relación entre significante y referente es decir, resulta difícil decidir si el ítem léxico se puede aplicar o no a ciertos objetos;
b- indeterminación del significado: cuando el significado es poco específico;
c- genericidad de un elemento del significado: cuando se trata de un significado genérico, impreciso respecto de algún componente de significado;

[17] Una expresión es *vaga* si tiene casos límite en que ni se puede aplicar de manera definida ni falla claramente en aplicarse (Asher (ed.), 1996, *s.v. vagueness*).

d- disyunción del significado: cuando existen diferentes interpretaciones posibles, en relación de disyunción.

Martí (2002) reconsidera estos tipos de vaguedad con el objetivo de circunscribir el ámbito de estudio de la polisemia e identificar sus tipos.

Así, el caso *a* se identifica más bien con vacilaciones en la percepción y conceptualización del mundo, y no es pertinente para una teoría del significado lingüístico, de modo que no lo consideraremos dentro del ámbito de la polisemia.

La situación establecida en *b* equivale a un significado indeterminado y de carácter general, que se concreta o *modula* en cada uno de los contextos en subsentidos (variaciones), en función de los tipos de contexto que seleccionan (p.ej., el tipo semántico del contexto) o de algún componente del significado de la palabra con la cual se combina (p.ej., su estructura de *qualia*) (Pustejovsky, 1995). La enumeración de sentidos plantea aquí problemas de exhaustividad, debido a que en los textos hay situaciones en las que los significados están superpuestos: se usan de forma genérica, sin modular (cf. Apresjian, 1974). Cruse (1995) considera estos casos entre la polisemia suave y la monosemia. Es una característica que se da frecuentemente en la categoría adjetivo (también en algunos nombres abstractos) y, en menor medida, en el verbo. Los adjetivos son polisémicos de manera distinta a como lo son los nombres; resulta prácticamente imposible predecir los diferentes significados que puede adoptar un adjetivo y enumerarlos en un diccionario.

Para *c***,** se puede hablar de polisemia sólo en términos de una situación comunicativa concreta, con lo cual también esta situación queda fuera del análisis del significado léxico. La falta de especificación respecto a cierto componente de significado corresponde más bien a una clase más general como denotación que no a una multiplicidad de significado.

Por fin, *d* corresponde a los casos de vaguedad más controvertidos. La bibliografía sobre este tipo de polisemia es amplia. Pustejovsky (1995) los llama *objetos complejos* (en inglés, *dotted types)* con doble denotación: permiten la selección de una u otra faceta del significado, en función del contexto. Para Deane (1988), que habla de *alosemia***,** el significado se comporta como un objeto complejo con múltiples facetas; los diferentes sentidos presentan una relación de contigüidad, normalmente basada en la metonimia (*animal* y *carne*, *proceso* y *resultado*, *continente* y *contenido*, *lugar* y *población*, etc.). En la visión de Copestake y Briscoe (1995), se trata de *extensión de sentidos*, mientras que Apresjian (1974) denomina el fenómeno *polisemia regular*.

Apresjian (1974) establece la dicotomía polisemia regular (alosemia) *vs.* polisemia irregular. Se da *polisemia regular* cuando hay al menos dos palabras, A y B, con sentidos diferentes a1 y a2, b1 y b2 respectivamente, donde a1 y b1 y a2 y b2, no son sinónimas y se puede establecer entre ellas el mismo tipo de relación. Por ejemplo, *taza* y *vaso* comparten la doble acepción de expresar el continente y el contenido. En caso contrario, se da *polisemia irregular***.**

La diferencia entre los dos tipos obedece a dos criterios, el tipo de relación entre los sentidos y el grado de productividad: según el primer criterio, la polisemia regular es principalmente derivada a través de metonimia (los significados tienen algún elemento común en su definición), mientras que la polisemia irregular es metaforización (eliminación o substitución de algún componente del significado por otro); según el segundo criterio, la polisemia regular es más productiva. Apresjian (1974) identifica diferentes tipos de polisemia regular diferenciados por categorías: 32 clases nominales, 15 verbales y 23 adjetivales, cada tipo en correspondencia con una relación semántica específica.

Basándose en este análisis y prescindiendo de la oposición entre homonimia y polisemia, Martí (2002) restringe el ámbito de la polisemia por medio de opciones dentro de las principales distinciones mencionadas: se descarta la polisemia pragmática u ocasional, considerando sólo la polisemia clásica; dentro de la polisemia clásica, se toma en cuenta sólo la extensión de sentidos y algunos casos de polisemia construccional y vaguedad; dentro de la vaguedad, se excluye la vaguedad referencial (*a*) y la vaguedad de imprecisión de un elemento del significado (*c*), manteniendo sólo la indeterminación del significado (*b*) y la disyunción del significado (alosemia) (*d*). En síntesis, la polisemia en sentido estricto quedaría limitada a los casos siguientes:

- polisemia irregular, de base metafórica (el caso más claro de polisemia);
- la polisemia regular, de base metonímica;

- la polisemia hipotáctica (una palabra es su propio hiperónimo);
- la vaguedad (indeterminación del significado), característica de los adjetivos calificativos.

Refina la última delimitación Cruse (1995), esta vez frente a la monosemia. En el marco de la lingüística cognitiva, Cruse habla de un continuum entre la polisemia y la monosemia, entre los cuales establece tres grados intermedios: semipolisemia, cooperativismo y latencia. Para analizar este continuum, el autor usa dos criterios: el carácter antagónico *vs.* cooperativo de los diferentes sentidos, y el carácter discreto de las propiedades semánticas de las unidades léxicas. La *polisemia clara* se caracteriza por la oposición (antagonismo) y por el carácter discreto de los sentidos: *banco*. La *semipolisemia* se da entre sentidos semidistintos, antagónicos y de carácter discreto suave, y propone dos variantes:

- *sentidos locales* o conjuntos de extensiones metafóricas de un tipo ontológico común, basadas en una correspondencia relacional similar: *boca de un río, boca de una cueva*, etc.) y
- *subsentidos*, que disponen de un sentido superordinado funcional completo, a modo de hiperónimo: *cuchillo,* 'de cubertería', 'de caza', y 'general, de tipo indefinido'.

En el caso del *cooperativismo,* los significados de la palabra tienen carácter discreto pero no antagónico. Se distinguen dos tipos de cooperación: *paratáctica*, en el caso de que los sentidos se comportan como facetas o componentes discretos de un único sentido sin ningún elemento superordinado que los incluya: *madre*, 'progenitor', 'cuidadora-educadora'; e *hipotáctica*, cuando las palabras pueden funcionar a la vez como hiperónimos e hipónimos: *perro*, 'animal' genéricamente y 'macho'. Por fin, define *latencia* como un fenómeno de elisión, en el que el elemento elidido puede tener distintos referentes y es recuperable por el contexto; es el caso de los adjetivos (p.ej. *fuerte*), en cuya definición no se especifica a qué tipo de objetos se puede aplicar. Sintetizamos en el siguiente esquema[18] los diferentes estadios propuestos por Cruse en el continuum entre la polisemia y la monosemia.

	Polisemia	**Semipolisemia**		**Cooperativismo**		**Latencia**	**Monosemia**
Antagonismo	+	+/-		-		-	-
Carácter discreto	+	+/-		+		-/+	-
		sentido local	subsentido	paratáctico	hipotáctico		
Ejemplo	*banco*	*boca / cuchillo*		*libro / perro*		*fuerte*	*atleta*

Figura 2.1. *Continuum* entre polisemia y monosemia

Los cinco estadios del continuum se pueden agrupar en dos: los dos primeros bajo el concepto de polisemia y los últimos tres dentro del concepto de la monosemia. Desde esta perspectiva, la monosemia correspondería a un significado que puede quedar subespecificado en su uso en un contexto (Martí, 2002)[19].

Aunque no es el objetivo de este trabajo, se ha querido mostrar brevemente la dificultad de establecer criterios claros sobre la delimitación de sentidos en el continuum del significado. En último término, la DSA depende de las decisiones que se hayan tomado sobre qué se considera polisemia y de como se ha enfocado el problema en el desarrollo de fuentes léxicas. Desde la perspectiva de la DSA se pueden realizar aportaciones que ayuden a una mejor delimitación y tratamiento del problema.

2.2 Tratamiento de la polisemia desde la DSA

La polisemia constituye uno de los temas centrales de estudio tanto en lingüística teórica como en lingüística computacional, dentro de teorías y orientaciones variadas. Nos detenemos aquí en los modelos del significado adoptados en la DSA, modelos que están a la base de las fuentes léxicas usadas (diccionarios en formato electrónico, redes semánticas, lexicones generativos) o bien están derivados a partir del corpus (*clusters* de palabras basados en similitud, conjuntos de rasgos cuyos valores se extraen de ejemplos etiquetados) (subapartado 2.2.1.).

[18] Reproducimos el esquema de Martí (2002).
[19] El tratamiento de la polisemia en Cruse (1986) se refina en Cruse (2000b) (cf. subapartado 2.2.1.3.).

Entre los problemas que plantea la polisemia para la DSA, cabe destacar la identificación de los criterios para delimitar las fronteras entre sentidos o acepciones, la delimitación del sentido que se ha de considerar habitual y de los usos que constituyen sentidos diferentes del considerado habitual, como también el número y la granularidad de los sentidos. Desde el PLN, se plantea además la utilidad de los resultados de la DSA para diferentes aplicaciones y dentro de diferentes dominios, así como el tipo de información que se toma a la base de un modelo de significado basado en corpus. Se plantea, por lo tanto, la cuestión de la relatividad de los sentidos (subapartado 2.2.2.).

2.2.1 Modelos del significado adoptados en la DSA

En la línea del diálogo necesario entre la lingüística teórica y la lingüística computacional, Ide y Véronis (1998) aseveran la necesidad de una visión, dentro de la DSA, más consistente con la teoría lingüística reciente[20]. Aunque se ha basado sólo en parte en los estudios de la lingüística teórica, la investigación en la DSA debe hallar su fundamento en las teorías del sentido de la semántica léxica. De hecho, se observa una tendencia cada vez más clara, en el área de la DSA, a considerar el significado desde la teoría y desde las aplicaciones lexicográficas.

En la lingüística teórica, más precisamente en la semántica léxica, se pueden identificar tres perspectivas principales, que recordamos de paso con sus correspondientes modelos teóricos y representaciones (cf. Ravin y Leacock, 2000):

1. El *enfoque clásico*, *aristotélico*, resucitado por Katz y Fodor (1963). Entre los modelos clásicos, recordamos: listas de sentidos, campos semánticos, marcadores semánticos (Katz, 1972), nódulos de sentido (Cruse, 2000b), etc.
2. El *enfoque prototípico* (Rosch, 1977, con orígenes en Wittgenstein). Algunos de los modelos basados en el concepto de prototipo son: esquemas (en inglés, *frames*, Fillmore, 1982); categorías radiales (Lakoff, 1987); tendencias, recursividad en la extensión del significado (Taylor, 1989); extensión del significado por reglas generativas (Pustejovsky, 1995), etc.
3. El *enfoque relacional* (*WordNet* -Fellbaum (ed.), 1998-, anticipado por Pierce y Selz), reflejado en redes semánticas como *WordNet* (Fellbaum (ed.), 1998), *FrameNet* (Fillmore y Atkins, 2000) y *MindNet* (Dolan *et al*., 2000).

Últimamente, se asevera la necesidad de un cambio radical de paradigma en la semántica léxica: la sustitución de la intuición en que se funda buena parte de la investigación por la evidencia ofrecida por los datos. La falta de evidencia objetiva hace imposible que la semántica cumpla el requisito de refutabilidad impuesto a las teorías científicas desde Popper[21].

Desde la lexicografía, se sigue la tradición de enumerar los sentidos de las palabras polisémicas y agrupar a aquellos relacionados como subsentidos. Las grandes discrepancias entre los diccionarios demuestran que no existen criterios objetivos para la discriminación y agrupación de los sentidos.

En el procesamiento del lenguaje natural, a base de la experiencia acumulada en las técnicas de DSA y de las dificultades encontradas, se constata la necesidad de definir el concepto de sentido sobre unas bases lo más objetivas posibles, problema directamente relacionado con el de la construcción de fuentes de conocimiento léxico adecuadas para las aplicaciones del PLN. En cuanto a cuestiones de orden procedimental, se ha enfatizado la importancia del contexto y de los métodos estadísticos; así mismo, se han desarrollado enfoques orientados hacia diccionarios *vs.* enfoques orientados hacia los corpus; no siempre se diferencia entre polisemia (significados múltiples al nivel de una palabra) y ambigüedad (significados múltiples al nivel de estructuras sintácticas más complejas); la granularidad de los sentidos es controvertida y se perfila una visión dinámica sobre el significado como proceso de creación del sentido en el contexto.

[20] Ejemplos en esta línea de convergencia entre la lingüística teórica y la lingüística computacional en el estudio y tratamiento del significado son Pustejovsky y Boguraev (1996) o Ravin y Leacock (eds.) (2000). Se encuentran aquí, en simbiosis y en vivo diálogo, la investigación en aspectos teóricos (con un panorama de las principales teorías del significado y su tratamiento de la polisemia) y la experiencia en enfoques prácticos para resolver el problema en el ámbito computacional.

[21] Responden a esta necesidad algunos estudios recientes como Stubbs (2001) o Singleton (2000).

A continuación, detallamos la problemática de la definición y de la delimitación de los sentidos en cuanto a criterios y granularidad. Presentamos aportaciones de las más relevantes en el tratamiento de los sentidos en el ámbito del PLN, pero traemos igualmente en discusión propuestas de la semántica teórica, de interés desde la DSA.

La DSA es problemática entre otras razones a causa de la inherente dificultad de discriminar los sentidos de las palabras. De hecho, la cuestión más persistente en la DSA es ¿qué es un sentido de una palabra? ¿Cómo identificar los sentidos? (Ide, 2000). Recordamos que, desde la DSA misma, se plantea también la cuestión de la *existencia de los sentidos*, ya que, si los sentidos no existen, no sirve

Más recientemente, medir el parentesco o la *distancia semántica* entre los sentidos de las varias palabras contingentes en el texto permite ulteriormente guardar la combinación que minimiza la distancia global entre los posibles sentidos de las palabras. Atkins (1987) y Kilgarriff (1997) adoptan la visión de Harris (1954): cada sentido está reflejado en un contexto distinto. Una visión similar es la adoptada en los métodos basados en clases (en inglés, *class-based methods*) (Brown *et al.*, 1992, Pereira y Tishby, 1992, Pereira y Lee, 1993).

Como posibles argumentos a favor de considerar los usos de las palabras como criterio básico para la discriminación de los sentidos cabe citar el experimento de Véronis (2001). El experimento tenía como objetivo, entre otros, poner de manifiesto el desacuerdo existente tanto en la discriminación como en la asignación de sentidos a las palabras de un texto, así como comprobar si el desacuerdo obtenido entre los jueces en el etiquetado se debía a las distinciones de sentido demasiado finas del diccionario -observación hecha muchas veces en la labor de DSA-, o a la problemática inherente al propio significado. Con el fin de verificar esta hipótesis, se recalculó el acuerdo entre los jueces después de haber reducido las respuestas a las distinciones de sentidos de nivel más general (posible gracias a la estructura jerárquica de los diccionarios franceses); la reducción del desacuerdo fue poca, algo más notable en los nombres, pero sólo en un 25% respecto a la tarea de discriminación a nivel de granularidad fina.

Este último resultado arroja una nueva luz sobre el etiquetado de sentidos, que se funda en una crítica distinta a los diccionarios, no ya en términos de granularidad, sino orientada hacia el estilo y la organización misma de las entradas. La carencia principal de los diccionarios reside, afirma Véronis, respaldado por el experimento, en la falta de información sintagmática. El papel de esta información no es de simple acompañante de los sentidos, sino que se debe tomar como base en la discriminación de sentidos. Con lo cual Véronis asevera la necesidad de un cambio radical en la lexicografía: las entradas se deben dividir en clases de uso coherente -que se pueden ver como sentidos- usando sólo la información distribucional, sin análisis del significado o introspecciones del lexicógrafo.

2.2.1.3 Enumeración *vs.* generación de sentidos

Una propiedad del significado de una palabra es que más sentidos de la palabra puedan usarse simultáneamente o ser *coactivados* (Kilgarriff, 1993, 1997 y Schütze, 1997). Esto se debe a que las relaciones semánticas entre las palabras se comportan como conjuntos difusos más bien que como categorías discretas aristotélicas (Zadeh, 1982). Muchos casos de *coactivación*, fenómeno muy frecuente en el lenguaje ordinario, son casos de *polisemia sistemática*, que se puede representar mediante reglas léxico-semánticas que se aplican a una clase de palabras y cambian o amplían sistemáticamente su significado. Sobre la polisemia sistemática, tenemos tanto estudios teóricos (Apresjan 1974; Pustejovsky, 1991; Lakoff, 1987; Copestake y Briscoe, 1995) como computacionales (Buitelaar, 1998).

Una nueva visión sobre los sentidos de las palabras abre la propuesta del *lexicón generativo* (Pustejovsky, 1995). A diferencia del enfoque enumerativo, que asume un conjunto *a priori*, establecido de sentidos, con existencia independiente del contexto, el enfoque generativo desarrolla una representación del sentido en que se asumen sólo asignaciones de sentidos subespecificadas hasta que el contexto esté estudiado. Las dificultades de determinar un conjunto de sentidos adecuados para la DSA indican la necesidad de prestar más atención a la visión generativa en la DSA (Ide y Véronis, 1998). Por lo tanto, nos detenemos brevemente en la teoría de Pustejovsky[24].

El modelo tradicional, que Pustejovsky llama *lexicón de enumeración de sentidos*, no logra tratar de manera satisfactoria aspectos de la polisemia como el uso creativo del lenguaje (las palabras adquieren sentidos nuevos en contextos nuevos) o la permeabilidad de los sentidos (los sentidos de las palabras se sobreponen). Frente a esta aproximación estática al significado en el que las palabras tienen generalmente un significado fijado y los sentidos están separados y clasificados de manera neta, el lexicón generativo aporta una visión dinámica: intenta ofrecer una explicación a la manera en que las expresiones adquieren contenido y en que este contenido se modifica continuamente en contextos

[24] En la presentación del lexicón generativo hemos aprovechado consideraciones de Martí (2002) y de Espinal y Mateu (2002).

nuevos. En palabras de Pustejovsky, "lo que es característico del ser humano no es tanto un lenguaje extensional *per se* como su habilidad generativa a concebir el mundo tal como le está revelado a través del lenguaje y de las categorías que usa. El lenguaje es una manifestación de esta facultad para la categorización generativa y pensamiento composicional" (Pustejovsky, 1998: 290, *apud* Espinal y Mateu, 2002:120). Precisamente, la ventaja de este modelo de descripción léxica es que permite separar de manera bien definida el componente léxico como depósito de datos del componente generativo de la gramática, es decir la sintaxis.

En el lexicón generativo se adopta una metodología basada en: (a) la infraespecificación en la semántica de los elementos léxicos; (b) todo elemento de un sintagma puede funcionar *de manera activa* en la composición del sintagma; (c) la noción de sentido de palabra emerge sólo al nivel de la frase, es decir "su sentido en el contexto".

Así, en los modelos que adoptan un lexicón estático, el hecho de cambiar un sentido de una palabra equivale a crear un nuevo elemento léxico o un nuevo sentido. Pero, según Pustejovsky, el significado cambia en un contexto de maneras específicas. Para tratar esta variabilidad, Pustejovsky propone la *infraespecificación* como mecanismo que permite expandir los diferentes sentidos y usos creativos de las palabras a partir de lo que considera sentidos básicos, que se caracterizan por estar infraespecificados. O sea, en una entrada léxica se declaran todas las características de la palabra que permiten componer sentidos y usos específicos en un contexto de uso dado. Además, con el fin de generar los diferentes sentidos de una palabra, se definen diferentes tipos de entrada léxica y unos mecanismos de composición.

Así, en el lexicón generativo se postulan cuatro niveles básicos de representación léxica:

1) *estructura argumental*: especificación del número y tipos de argumentos;

2) *estructura eventiva*: definición del tipo de acontecimiento denotado (estado, proceso, etc.);

3) *estructura de herencia léxica*: identificación del modo en que una estructura léxica se relaciona con otras estructuras;

4) estructura de *qualia*: representación de los diferentes modos de predicación posibles de una unidad léxica.

El nivel de la *qualia* codifica la información semántica que hace referencia a la comprensión de un objeto o de una relación, mientras que los niveles de la estructura argumental y la estructura eventiva son *stricto sensu* más relevantes para la estructura gramatical. El significado de las palabras se estructura sobre la base de los cuatro factores generativos llamados *qualia*, que forman un "conjunto de restricciones semánticas mediante de las cuales entendemos una palabra en la lengua" (Pustejovsky, 1995:86). Estos papeles generativos son: el *papel formal* (lo que distingue un objeto dentro de un dominio más amplio); el *papel constitutivo* (lo que indica la relación entre el objeto y las partes que lo componen); el *papel télico* (lo que indica la función o finalidad del objeto); el *papel agentivo* (lo que hace referencia a los factores implicados en la creación o generación del objeto). Por ejemplo, la entrada léxica de *libro* contiene tres argumentos, uno para el objeto físico (*x*), uno para el contenido (*y*) y otro para una combinación entre los dos, *x·y*. Los *qualia* de *libro* determinan las relaciones que estos argumentos pueden tener uno con el otro o bien con otros componentes semánticos de su contexto. El *qualia* formal especifica que *x* contiene *y*, mientras que el *qualia* télico especifica que el propósito y la función del libro –ser leído por un agente– que se aplica al concepto combinado (*x·y*). La estructura de *qualia* constituye el núcleo de las propiedades generativas del lexicón, ya que nos proporciona una estrategia general para la creación de conceptos específicos con propiedades conjuntivas. Es decir proporciona una estructuración coherente para diferentes dimensiones del significado léxico.

A la vez, Pustejovsky define unos mecanismos de composición semántica:

1) *selective binding*: el tipo del argumento (polisémico o complejo, en inglés, *dotted*) está subsumido por el tipo esperado por el predicado del argumento;

2) *co-composición* (léxica): permite cambiar la semántica del predicado (verbo) en función del argumento (objeto) con el que interactúa;

3) *coerción de tipos*: una unidad rectora fuerza a otra unidad a cambiar su tipo semántico (pero no el tipo sintáctico).

Todos los tipos que se pueden derivar de la *qualia* definen la expansión generativa del tipo mediante la coerción de tipo. Los *qualia* proporcionan una estructura donde se aplican transformaciones semánticas que alteran la denotación de una palabra. El tipo asociado a las entradas

léxicas puede variar según el contexto a través de operaciones como la composición y la coerción de tipos.

En síntesis, el lexicón generativo define un conjunto de principios que controlan las condiciones de buena formación de los signos complejos (sintagmas y oración) y captan las relaciones entre las unidades que las componen. La aplicación de las reglas está condicionada por el entorno sintáctico-semántico en que aparece la pieza léxica. Pustejovsky defiende la imposibilidad de separar el significado de la estructura sintáctica que lo contiene.

Cruse (2000b)[25] adopta una posición que ubica entre el lexicón enumerativo y el lexicón generativo. El significado de una palabra es el contenido conceptual (o un tipo de agrupación de contenido conceptual) que se hace accesible por el uso de la palabra, como opuesta a otras palabras, en contextos particulares. Aunque en principio la polisemia se puede ver como infinitamente variable y dependiente del contexto, hay, sin embargo, regiones de mayor "densidad" semántica que forman espacios más o menos definidos de significado (*nódulos de sentido*), con estabilidad mayor o menor respecto al cambio del contexto. Un nódulo de sentido es una unidad de sentido relativamente autónoma, que puede jugar un papel independiente en diferentes procesos semánticos. Tiene propiedades de estabilidad (relativa) respecto del cambio de ciertas propiedades del contexto y riqueza y/o prominencia del contenido conceptual que se hace accesible. En principio, los nódulos se forman y se disuelven según cambie el contexto. Con esto no se quiere denegar la posibilidad de propiedades semánticas de las palabras, invariables en el contexto: Cruse expone una amplia gama de dependencias contextuales. La posición asumida, sin embargo, es que en general no es posible especificar propiedades semánticas de las palabras de una manera independiente del contexto.

Así, se habla de los siguientes factores de discontinuidad en el significado de las palabras: por una parte, las diferencias entre las entidades conceptuales[26], que determinan la división del significado de una palabra en subsentidos o bien en facetas; por otra parte, los puntos de vista sobre una misma entidad. Los subsentidos se asocian a tipos o variedades de las entidades, a una subclasificación (por ejemplo, para cuchillo, 'cuchillo de cocina', 'cuchillo de jardín', etc.). Parecen característicos de los nombres definidos a base de la función. Son casos de incompleta descontextualización del significado de las palabras. Las facetas son aspectos o componentes discretos del significado de una palabra, cada uno con un tipo ontológico radicalmente diferente. En otras palabras, son partes coordinadas del significado global de una palabra; por ejemplo, *país* como 'tierra', 'estado', 'nación', 'población', etc. Los puntos de vista son modalidades de construcción y suponen unidades de sentido; por ejemplo, *libro*, como 'volumen' o 'texto'. Los modos de ver se aproximan a los papeles de los *qualia* de Pustejovsky (cf. *supra*). Los subsentidos, las facetas y los puntos de vista corresponden a diferentes grados de discretitud entre monosemia y polisemia (cf. apartado 2.1.).

2.2.1.4. Potencial de significado

Hanks (2000), siguiendo la perspectiva centrada en el concepto de juego de Wittgenstein, trata los significados como eventos, no como entidades: los significados no existen fuera de los contextos transaccionales en que se usan (eventos de significado), los "significados en los diccionarios" no son significados, sino potenciales de significado (en inglés, *meaning potential*). El potencial de significado de cada palabra está constituido por un número de componentes semánticos (o cognitivos), que son separables, combinables, explotables, probabilísticos y prototípicos, no necesariamente compatibles, manteniendo entre sí una coexistencia pacífica como parte del potencial de significado de la palabra. Los componentes pueden ser activados cognitivamente por las otras palabras en el contexto en que se usa la palabra dada, de manera que los potenciales de significado están dispuestos jerárquicamente en una serie de opciones por defecto: cada interpretación por defecto se asocia con una jerarquía de normas sintácticas. Esta red provee la base de toda la semántica del lenguaje, con enorme potencial para decir cosas nuevas y relacionar lo desconocido con lo conocido. El proceso de desambiguación supone una competición entre los distintos componentes o conjuntos de componentes. En este marco,

[25] En este trabajo se retoman y desarrollan las ideas de Cruse (1985) al cual hemos hecho referencia en el apartado 2.1.

[26] Cruse denomina *individuadores* a los factores de discontinuidad atribuibles a diferencias de entidades conceptuales.

el problema tiene que reformularse: no se trata tanto de qué sentido tiene una determinada palabra en un contexto sino de cuál es la contribución específica de una determinada palabra al significado del texto. La desambiguación y discriminación del sentido de una palabra se convierte, de este modo, en la selección de la combinación de aquellos componentes de su potencial de significado activados por los "disparadores" (en inglés, *triggers*) contextuales. En este caso, la lexicografía computacional tendría como objetivo identificar los componentes de significado de las palabras, el modo en que se combinan (jerarquía y relaciones), los vínculos con los componentes de significado de las palabras relacionadas semánticamente y las circunstancias oracionales en las cuales están activados.

2.2.1.5. Modelo relacional

Una de las fuentes léxicas más usadas actualmente en DSA es *WordNet* (Fellbaum, 1998)[27], que se basa en un modelo relacional del lexicón[28]. En los modelos relacionales, vinculados con dominios semánticos, las palabras están organizadas en redes semánticas, en base a relaciones semánticas, explícitas, que establecen entre ellas. Las relaciones semánticas que se usan en los modelos relacionales se pueden dividir en los dos tipos principales de relaciones que Saussure (1916) establece entre las unidades del sistema lingüístico: sintagmáticas y paradigmáticas. En los modelos relacionales, se consideran relacionadas sintagmáticamente las palabras que coocurren frecuentemente, incluyendo aquí también las colocaciones y las restricciones de selección. Se consideran, en cambio, relacionadas paradigmáticamente las palabras que aparecen en un contexto similar. Con pocas excepciones (por ejemplo, Ahlswede y Evans, 1988), los lexicones relacionales contienen exclusivamente relaciones paradigmáticas: hiponimia/hiperonimia, meronimia/holonimia, sinonimia, antonimia, troponimia, etc. La hiponimia/hiperonimia corresponde a la relación case-subclase: el hipónimo equivale al individuo de una clase o a la subclase, y el hiperónimo, a la clase[29]. La meronimia/holonimia es la relación entre una parte y un todo: el merónimo corresponde a la parte, el holónimo, al todo. Dos palabras se consideran sinónimas si su similitud semántica es más relevante que sus diferencias[30].

Desde esta perspectiva, conocer el significado de una palabra equivale a conocer su ubicación en el espacio semántico cubierto por la red. Aunque no descomponen el significado de los conceptos, los modelos relacionales mantienen la división del significado de las palabras en sentidos distintos. La representación de la polisemia regular es más difícil en el ámbito de un modelo relacional. Para codificar la relación semántica regular entre nodos con la característica *a* y nodos con la característica *b*, Miller (1998, *apud* Ravin y Leacock, 2000) propone relacionar los nodos *a* y *b* en un nivel superior de la jerarquía taxonómica. Por ejemplo, codificando la relación entre el nodo *árbol* y el nodo *madera*, la relación se transmitirá para abajo, entre todos los árboles y su madera.

WordNet, en particular, está construido desde una perspectiva diferencial sobre el significado: se asume que los usuarios ya conocen el concepto con sus significados y que la representación de los significados del concepto debe ayudar sólo a diferenciarlos[31]. Así, la referencia a un sinónimo o cuasisinónimo permite distinguir de qué concepto se trata. Por lo tanto, los conjuntos de sinónimos (*synsets*, en terminología de Miller y Charles, 1991) corresponden a conceptos; no los definen, pero establecen su existencia y su identificación posible. En *WordNet* se representan igualmente conceptos no lexicalizados, sin una forma léxica; en este caso, se recurre a una definición que los identifique. Se supone que el léxico mental está organizado según una serie de relaciones semánticas entre los significados. Como en *WordNet* los significados se expresan mediante *synsets*, las relaciones semánticas se conciben entre *synsets*. El papel de las relaciones semánticas es doble: indican la manera bajo la cual se relacionan los significados y explicitan las palabras que están relacionadas.

[27] Presentamos las características de *WordNet* y de otras fuentes derivadas de ésta en el capítulo 3.
[28] Recogemos observaciones de Ravin y Leacock (2000) y de Martí (2002).
[29] Cruse (1986) distingue la hiponimia/hiperonimia de la taxonomía. No insistimos aquí en la diferencia, debido a que en *WordNet* (que usamos en nuestra investigación), no se hace esta delimitación y ambos fenómenos se registran como hiponimia/hiperonimia.
[30] Para una presentación más detallada de las relaciones paradigmáticas, ver Cruse (1986, 2000a).
[31] En Miller y Fellbaum (1991, *apud* Martí, 2002), se contraponen, en el ámbito de la semántica léxica, las teorías diferenciales a las teorías constructivistas. Éstas últimas asumen que la representación del significado debe contener información suficiente para la reconstrucción precisa de los conceptos.

2.2.2 Problemas que plantea la polisemia para la DSA

2.2.2.1. Delimitación de los sentidos. Criterios

La representación semántica es una de las grandes limitaciones de las técnicas del PLN (en inglés, *bottleneck of semantic representation*). Palmer (2000) opina que la clave para su solución es hallar qué constituye una clara separación de sentidos y cómo pueden caracterizarse y distinguirse los sentidos desde una perspectiva computacional. Reclama además, como necesario, un consenso sobre las líneas maestras de los lexicones computacionales. Los intentos de delimitar los sentidos tienen que cumplir dos requisitos: criterios explícitos y flexibilidad de la representación.

Si bien existe cierta base psicológica para la noción de sentido (Simpson y Bengess, 1988; Jorgensen, 1990, *apud* Ide y Véronis, 1998), no existe un acuerdo comúnmente aceptado sobre el número y la delimitación de los sentidos de cada palabra. A pesar de las dificultades que entraña este problema tanto desde un punto de vista filosófico como lingüístico, en el PLN se han hecho esfuerzos en la línea de encontrar medios prácticos para distinguir los sentidos de las palabras, por lo menos a un nivel que los haga útiles para las tareas del PLN.

Así, se han propuesto varias técnicas de agrupamiento de los contextos de una determinada palabra para definir sus sentidos. Estas técnicas presentan, sin embargo, la desventaja de que muchas veces es difícil asignar un sentido único o encontrar un sentido adecuado entre las variantes (Yarowsky, 1992): no está claro cuando los sentidos se tienen que agrupar o dividir, dado que en muchos casos el significado se considera un continuo a lo largo del cual se distribuyen los diferentes matices de significado (Cruse, 1986), y los puntos donde los sentidos se agrupan o se diferencian pueden variar de manera muy significativa. Uno de los modelos más extensamente adoptados en la distinción de sentidos es el de *banco***,** que permite diferenciar entre *banco-dinero/banco-asiento*. Se ha tratado de extender y generalizar este modelo para cualquier distinción de sentido, pero evidentemente no es aplicable a todas las palabras.

En el mismo afán de buscar medios prácticos para la distinción de sentidos, útiles en las tareas del PLN, se han propuesto varios criterios para la determinación de los sentidos en un contexto[32], entre los que destacan el comportamiento sintáctico y el conocimiento semántico y pragmático. Otros criterios propuestos son la coocurrencia de la palabra dentro de relaciones sintácticas (Hearst, 1991; Yarowsky, 1993) o las palabras coocurrentes en un contexto global (Gale *et al.*, 1993; Yarowsky, 1992; Schütze, 1992, 1993). Sin embargo, de momento no se ha logrado establecer ningún criterio claro de discriminación, por lo que el problema sigue dominando la investigación en el área de la DSA (Ide, 2000). A continuación, recordamos algunas de las propuestas de carácter programático.

Coherente con los dos requisitos mencionados para los lexicones computacionales, Palmer (2000) sugiere "criterios consistentes" para las distinciones de los sentidos en el caso concreto de los verbos: cambios en la estructura argumental; cambios en los esquemas (*frames*) sintácticos y/o en las restricciones sobre la clase semántica; y coocurrencia léxica.

Como posibilidad de obtener la deseada flexibilidad de la representación, sugiere (para el caso específico de los verbos) añadir a las entradas del *WordNet* los cambios sistemáticos en los esquemas (*frames*) de subcategorización y las preposiciones o sintagmas adverbiales/preposicionales regidos, lo que permitiría el establecimiento de extensiones regulares en el significado. Este nuevo tipo de estructuración sería una solución a las críticas de *WordNet* (granularidad excesiva de los sentidos) ya que permitiría el desplazamiento entre las distinciones de sentido groseras y las finas, dado que las distinciones basadas en conocimiento del mundo son problemáticas.

Corazzari *et al*. (2000) proponen una perspectiva distinta: discriminar los sentidos a partir del análisis de corpus anotados semánticamente (en inglés, *semantically tagged*). Este análisis ofrece una visión más completa y más precisa sobre la semántica de los lemas; como tal, permite decidir acerca de las distinciones de sentido (cuántas y cuáles) basándose en los datos proporcionados por la evidencia, más fiables que la simple intuición. Los autores toman en cuenta diferentes parámetros (en inglés, *clues*) para la DSA o pruebas (en inglés, *tests*), como son los indicadores sintácticos y semánticos, para la identificación de los sentidos:

- un esquema sintáctico específico permite seleccionar un determinado sentido;

[32] Véase una presentación en Ide y Véronis (1998).

- un dominio semántico de uso ayuda a seleccionar un determinado significado para un lema dado;
- un modificador específico selecciona o muestra preferencias por un sentido particular;
- una clase específica de sujeto, de objeto directo o de objeto indirecto pueden ayudar a seleccionar un determinado significado;
- sinónimos y/o antónimos distintos seleccionan sentidos distintos (Cruse, 1986);
- dos sentidos distintos de un lema no pueden ser seleccionados simultáneamente por el mismo contexto (Cruse, 1986).

Además, señalan las posibilidades que ofrece disponer de grandes cantidades de corpus anotados semánticamente para el análisis del impacto de las distintas claves (*clues*) sobre la DSA. Por otra parte, resaltan la necesidad de integrar los diccionarios tradicionales o los lexicones computacionales con los sentidos atestados en los corpus e identificar el carácter inadecuado de determinadas distinciones de sentidos que figuran en los diccionarios tradicionales o los lexicones computacionales corrientes.

Comparte esta posición Véronis (2000), que, a partir de una serie de experimentos sobre discriminación y notación de sentidos, se pronuncia a favor del uso de información sintagmática extraíble a partir de corpus no sólo como mero anexo a las entradas en los diccionarios actuales sino como fundamento para la delimitación de los sentidos.

Recordamos en este contexto la línea de investigación en desambiguación automática que prescinde del uso de alguna fuente de conocimiento sobre los sentidos de las palabras, la DSA no supervisada, y que construye los sentidos a base exclusivamente de los contextos (p.ej. Schütze, 1992). Tal definición de los sentidos, que lleva a la mera DSA débil, o discriminación de sentidos, es suficiente para algunas aplicaciones de PLN, como la recuperación de información.

Una línea de trabajo más reciente en el área se funda en la idea de que la comparación entre las lenguas es útil para la DSA. Se parte en este caso de la suposición de que la correspondencia (en inglés, *mapping*) entre las palabras y los sentidos varía de manera significativa de una lengua a otra (Ide, 2000). Resnik y Yarowsky (1997) sugieren que, para los propósitos de la DSA, los varios sentidos de una palabra podrían determinarse considerándose sólo las distinciones de sentido lexicalizadas entre lenguas (en inglés, *cross-linguistically*)**.** Más preciso, proponen identificar un conjunto de lenguas de destino (en inglés, *target*) y restringir las distinciones de sentido necesarias en las aplicaciones y en la evaluación del PLN a las realizadas léxicamente en un subconjunto mínimo de aquellas lenguas[33].

2.2.2.2. Número y granularidad de los sentidos

Uno de los problemas principales para la DSA es determinar la adecuada granularidad, el nivel de generalidad de los sentidos. No hay un acuerdo general sobre el número adecuado de sentidos para las entradas léxicas. Entre las propuestas teóricas, en los extremos se situarían, por un lado, Wierzbicka (1989) para quien las palabras tienen esencialmente un único sentido y, por otro, Pustejovsky (1995), para quien las palabras pueden asumir un número de sentidos potencialmente infinito.

Desde una perspectiva computacional, se resalta que el número y la granularidad de los sentidos dependen del tipo de información usada para su descripción, es decir**,** de la fuente de conocimiento (Ide y Véronis, 1998; Corazzari *et al*., 2000; Yarowsky, 2000b, etc.) utilizada para la DSA. Veremos a continuación algunos casos representativos.

Los diccionarios suelen contener distinciones excesivamente finas para la DSA; los que han sido comprobados como más adecuados para la tarea de desambiguación semántica son los diccionarios monolingües para el aprendizaje de segundas lenguas (en inglés, *learner dictionaries*), ya que suelen contener sentidos básicos. Éste es el caso del LLOCE (*The Longman Lexicon of Contemporary English*) aunque, si bien ofrece la granularidad adecuada para la DSA, presenta, en cambio, la desventaja de tener un vocabulario limitado y una cobertura reducida sobre los sentidos de las palabras.

Los *thesauri* presentan una organización de los sentidos en un número limitado de categorías semánticas. El *Roget's* (*Roget's International Thesaurus of English Words y Phrases*, 3ª edición, 1995, 2000) tiene una jerarquía con tres niveles, donde el nivel superior o *top* contiene 6 clases, el nivel

[33] Adoptan esta perspectiva, entre otros, Ide (2000), Tufis (2002, 2004), Diab y Resnik (2002), Chugur *et al.* (2002), etc.

mediano, 39 secciones, y el inferior, 1000 categorías. El LLOCE, que reúne las características de un diccionario y las de un tesauro, organiza los sentidos principalmente según el argumento (en inglés, *subject*), en una jerarquía con tres niveles: sujeto (14), tópico (129), conjuntos (conjuntos tópicos de palabras) (2500).

El *WordNet* (Fellbaum (ed.), 1998) también hace distinciones de sentidos de granularidad demasiado fina.

Los problemas principales relacionados con la granularidad tienen que ver primero con las divisiones demasiado finas, las cuales crean dificultades para la DSA ya que:

- introducen importantes efectos combinatorios;
- requieren elecciones de sentido extremadamente difíciles;
- aumentan la cantidad de datos necesarios para los métodos supervisados de DSA[34] a proporciones no realistas (Slator y Wilks, 1987).

Las distinciones hechas en muchos diccionarios van más allá de la capacidad de los humanos mismos (Kilgarriff, 1992, 1993).

Entre las variadas soluciones propuestas, Manning y Schütze (1999) recuerdan la estrategia de restringir el problema, de definir uno más fácil, como en el caso del análisis sintáctico limitado al análisis superficial (en inglés, *shallow parsing*). En el caso específico de la DSA, la aplicación de la estrategia se ha hecho de dos maneras:

1. Se toman en consideración sólo las distinciones más básicas (en inglés, *coarse-grained*), como por ejemplo, las que se manifiestan a través de diferentes lenguas, partiendo de la hipótesis de que muy probablemente la polisemia sistemática es similar en todas ellas. Estas soluciones tienen su utilidad ya que muchas ambigüedades de traducción son muy generales, por lo que un sistema limitado a distinciones básicas de sentidos es suficiente.
2. Se parte de enfoques de agrupamiento (en inglés, *clustering approaches*) con la idea de que la agrupación automática encontrará sólo grupos de uso que se pueden distinguir con éxito. Dolan (1994) aplica un método para desambiguar los sentidos de los diccionarios a través de su combinación en sentidos más amplios. Otra variante es usar las divisiones de sentido más generales de *thesauri* como *Roget's* (cf. Ide y Véronis, 1998).

Chen y Chang (1998) proponen un compromiso entre un diccionario (*LDOCE*, *The Longman Dictionary of Contemporary English*) y un *thesaurus* (LLOCE): la unificación de los sentidos de un diccionario en *clusters* de acuerdo con la granularidad de un *thesaurus*, con distinciones de sentido más básicas, es decir el agrupamiento automático de los sentidos de un diccionario accesible por ordenador[35] a base de la información de un *thesaurus*.

Por otro lado, el sólo hecho de agrupar los sentidos de los diccionarios no resuelve el problema debido a que el grado de granularidad requerido depende de la tarea (Ide y Véronis, 1998; Yarowsky, 2000b, etc.). Por ejemplo, para la síntesis del habla y para la recuperación de los acentos de un texto se necesita sólo la distinción entre homógrafos, mientras que para la traducción automática suelen hacer falta distinciones de sentido más finas, a veces más finas que las de un diccionario. Pero no hay una estricta correspondencia entre una tarea dada y el grado de granularidad requerido: así, para la traducción a veces las distinciones de sentido no son necesarias ya que existe una correspondencia entre los varios sentidos a través de las lenguas (por ejemplo, el inglés *mouse* y el francés *souris*) mientras que para la recuperación de información la distinción entre los sentidos, en este caso, sería importante.

Además, los diccionarios son generalmente genéricos, mientras la labor práctica de DSA es para subdominios. Para el problema de los dominios específicos, una solución es partir con un lexicón genérico y adaptarlo de manera automática con palabras y sentidos de especialidad (Wilks y Stevenson, 1997).

Una alternativa para resolver el problema, procedente del área de la DSA no supervisada, es construir sistemas en los cuales el número de sentidos de las palabras -sentidos construidos como grupos de contextos similares con respecto a una métrica dada- sea variable, un parámetro ajustable en función de las necesidades de la aplicación (Manning y Schütze, 1999).

[34] Presentados en el apartado 4.3.1.

[35] O diccionarios en formato electrónico. Ver el capítulo 3 para una presentación más detallada.

2.2.2.3 Relatividad de los sentidos. Factores que condicionan el concepto de sentido en el marco del PLN

Las consideraciones previas han puesto de relieve la relatividad de los sentidos en términos de los cuales se realiza la tarea de DSA. A continuación, detallamos más la dependencia del concepto de sentido en el marco del PLN con respecto a factores como el tipo de información usada para la DSA, la aplicación a la cual está subsumido el proceso de DSA, el dominio del corpus sobre que se entrenan los métodos basados en aprendizaje, etc.

El tipo de información usada en el proceso de desambiguación por las técnicas de DSA define de manera implícita una determinada concepción de sentido (Manning y Schütze, 1999). Así, tenemos criterios de discriminación de sentidos basados en la coocurrencia (el modelo bolsa de palabras, en inglés, *bag-of-words*), en la información relacional (sujeto, objeto, etc.), en la información gramatical (clase morfosintáctica de la palabra), en las colocaciones (la hipótesis de un sentido por colocación, en inglés, *one sense per collocation*), en el discurso (la hipótesis de un sentido por texto, en inglés, *one sense per discourse*), etc. Por ejemplo, si se usa la información de coocurrencia, se reconocen sólo las distinciones de sentido tópicas, como son los sentidos asociados a varios dominios.

Las fuentes léxicas mismas difieren ampliamente en cuanto al número y a la manera de definir los sentidos[36]. Más allá de las distinciones inherentes de orden tipológico entre estos recursos, cabe señalar aquí las diferencias marcadas incluso dentro de una misma clase de fuentes léxicas, como son los diccionarios.

Se puede concluir, con Wilks y Stevenson (1997, 2000), que los sentidos de las palabras no son absolutos, sino relativos respecto de una determinada fuente. Es muy difícil asignar las ocurrencias de las palabras a clases de sentidos de alguna manera que sea a la vez general y precisa. Como tal, la DSA depende siempre de las opciones de sentido disponibles en la fuente, por lo que su nivel de calidad no se puede valorar de manera absoluta; a la vez, la asignación de un sentido a una palabra tiene un alto grado de relatividad.

Una opción muy frecuente es el uso de algoritmos de DSA entrenados sobre corpus. En este caso, la desambiguación se realiza no en relación a un conjunto bien definido de sentidos, sino con respecto a un conjunto de sentidos *ad hoc* (Habert *et al*., 1997), lo que cuestiona la utilidad de tal método (Wilks y Stevenson, 1997). Varios experimentos realizados últimamente ponen de relieve la relatividad del conocimiento relacionado con los sentidos adquirido sobre un corpus, conocimiento que depende de las particularidades de género y de tópico del corpus (Krowetz, 1997; Martínez y Agirre, 2001; Escudero *et*., 2000b).

Otro factor importante que influye en la definición de sentido asumida en un sistema de PLN es su tarea final: RI, TA, etc. Se habla en este caso de nociones de sentido orientadas hacia las aplicaciones (Manning y Schütze, 1999). Su ventaja es que se pueden evaluar fácilmente a través de los resultados de la aplicación en que la desambiguación participa, frente a la evaluación directa, independiente de la tarea, que es más difícil[37]. El potencial de la DSA varía según la tarea: las mayores aplicaciones del lenguaje difieren en su capacidad de hacer un uso satisfactorio de la buena información sobre el sentido de las palabras (Resnik y Yarowsky, 1997).

Uno de los casos más evidentes es la traducción automática, para la cual hacen falta las distinciones de sentido que reflejen las diferentes traducciones que una palabra de una lengua pueda tener a otra (Yarowsky, 2000b; Ide, 2000, etc.). Otro ejemplo es el de la incorporación de un módulo de DSA en sistemas de recuperación de información (Mihalcea y Moldovan, 2001), donde la desambiguación está orientada a la pregunta (en inglés, *query*) y a los documentos de donde en que se busca la información.

Cabe mencionar aquí la posición escéptica de Kilgarriff (1997), que afirma la no existencia de los sentidos en abstracto, independientemente de alguna tarea de PLN. No es posible una lista de sentidos que sea válida para diferentes aplicaciones a la vez. Los usos de las palabras serían no estándar, los sentidos serían dictados por los corpus.

---

[36] Como hemos ya mencionado en el apartado 2.2.2. Nos remitimos al apartado 3.1.1. para otros detalles.
[37] Véase el capítulo 5.

El dominio tiene también consecuencias para la definición del sentido. Íntimamente vinculado con el tema, el tratamiento de dominios específicos nos lleva al problema de la adaptación del lexicón a corpus de un determinado dominio (en inglés, *lexical tuning*) (v., p.ej., Basili *et al*., 1998). La importancia de esta operación es obvia: un discriminador entrenado sobre un dominio específico es menos efectivo sobre un texto de otro dominio, y al revés, un discriminador entrenado sobre diferentes tipos de texto, es decir un discriminador genérico, no tiene un rendimiento particularmente alto (Wilks y Stevenson, 1997). Los experimentos de carácter comparativo sobre sistemas de DSA basados en corpus[38] demuestran, con la relatividad del conocimiento adquirido, la necesidad de adaptación de los algoritmos para su aplicabilidad a otros dominios (Escudero *et al*., 2000b). En esta línea, Moldovan y Girju (2000) proponen la extensión automática de *WordNet* con conocimiento sobre dominios a partir de corpus. Otros desarrollos de *WordNet* con información de dominio pertenecen a Magnini y Cavaglià (2000) (*WordNet Domains*) o a Agirre y López de Lacalle (2004).

Hemos intentado poner de relieve, en estos dos primeros apartados, la relación dialéctica existente entre el estudio de los sentidos y de la polisemia, por un lado y, por otro, la desambiguación semántica automática. Si la tarea computacional se nutre de la perspectiva teórica sobre la polisemia, también es verdad que las aplicaciones y el dominio en que desemboca el proceso de desambiguación influyen en las opciones sobre los sentidos; además, las dificultades y la casuística que la DSA revela ofrecen la reclamada evidencia objetiva necesaria para la semántica léxica. A pesar de la vasta bibliografía dedicada al respecto, la polisemia sigue siendo un problema teórico de difícil solución (Ravin y Leacock, 2000).

2.3 Significado y contexto

Según Cruse (1986), en el enfoque contextual se considera que el significado de una palabra está totalmente reflejado en sus relaciones contextuales e incluso construido por éstas. O sea se asume que las propiedades semánticas de un ítem léxico están plenamente reflejadas en aspectos adecuados de las relaciones que el ítem contrae con el contexto presente y con los contextos potenciales. En consecuencia, se buscará derivar información sobre el significado de una palabra a partir de sus relaciones con los contextos lingüísticos presente y potenciales.

La variabilidad del significado en el contexto se presenta en términos de tres posibles efectos del contexto sobre el significado de una palabra: selección, coerción, modulación. La *selección* opera mediante la supresión de las lecturas que producen un tipo de conflicto semántico con el contexto, y la selección de sólo una lectura, no conflictiva. La selección se da entre los posibles grados de distintividad dentro del significado de la palabra: sentidos, subsentidos, perspectivas, facetas[39]. En el caso de la *coerción*, ninguna de las lecturas establecidas es compatible con el contexto. Esto supuestamente dispara una búsqueda entre posibles extensiones de significado, como metáforas o metonimia, para una lectura que sea compatible con el contexto. Si se halla una, ésta se considerará la lectura que se quería transmitir y decimos que el contexto ha obligado a una nueva lectura. La *modulación* cubre las variaciones de significado que aparecen como resultados de efectos contextuales dentro de los límites de un único sentido. Hay dos grandes variedades: enriquecimiento y empobrecimiento, según añaden o quitan significado (Cruse, 2000b:120-121).

Para desarrollar la cuestión de la interpretación consistente de una expresión lingüística compleja o sea para dar cuenta de cómo influyen las palabras del contexto el significado de una palabra, hay que estudiar las relaciones sintagmáticas de sentido[40]. La buena formación semántica de un discurso o de una estructura sintáctica está determinada por las relaciones semánticas que se establecen entre sus unidades léxicas: relaciones entre unidades discursivas más amplias y relaciones al nivel léxico. Nuestro interés recae en las últimas relaciones que corresponden a una interacción sintagmática entre los elementos significativos de un discurso, interacción gobernada por la sintaxis[41]. Las buena formación de las expresiones lingüísticas complejas se expresa en términos de su normalidad semántica. Cruse identifica dos tipos de anormalidad: *conflicto*, cuando los significados de los

[38] Presentados en el apartado 4.3.
[39] Hemos definidos estos conceptos en el apartado 2.1.
[40] A continuación, sintetizamos las ideas sobre las relaciones sintagmáticas expuestas en Cruse (2000a: 219-235).
[41] En el primer caso, se trata de interacción pragmática, basada en el conocimiento enciclopédico.

constituyentes no van bien juntos, y *pleonasmo*, cuando un significado no aporta nada nuevo al otro. Desde esta perspectiva, las unidades sintagmáticas imponen condiciones semánticas sobre sus colaboradores (en inglés, *partners*) sintagmáticos. Si las condiciones están satisfechas, el resultado es bien formado y la combinación interpretable. En caso contrario, el conflicto puede provocar una transformación semántica para que la lectura sí satisfaga las condiciones. Estas condiciones se llaman *preferencias de coocurrencia*. Entre ellas, las *preferencias colocacionales* son consecuencia inherente del contenido oracional, mientras las *preferencias de selección* no lo son. Nos centramos, por lo tanto, en éstas últimas.

A diferencia de las relaciones paradigmáticas, que son generales, independientes de contexto, las relaciones sintagmáticas están ligadas a construcciones gramaticales particulares. Cruse establece tres tipos básicos de relaciones sintagmáticas: de *normalidad* (entre filónimos), de *conflicto* (entre xenónimos) y de *pleonasmo* (entre tautónimos). Las restricciones sobre la coocurrencia entre ítems léxicos suelen tener propiedades direccionales. Así, la direccionalidad de una relación sintagmática -que supone una relación de selección- significa que un ítem hace de seleccionador y el otro de seleccionado. Según Cruse, hay dos nociones de selección: 1) selección entre un conjunto de lecturas polisémicas u homónimas, o sea en el plano paradigmático; 2) selección entre los significados de los ítems en una construcción gramatical. En el primer caso, la selección es bidireccional, mientras que en el segundo caso, se trata de un proceso unidireccional y la dirección en que opera la selección está correlacionada con la gramática. Así, los adjetivos seleccionan a los nombres núcleo y los verbos a sus complementos. En general, los nombres están siempre seleccionados. En términos lógicos, el predicado selecciona y los argumentos están seleccionados.

En cuanto al papel semántico de los constituyentes de una construcción gramatical, el núcleo semántico determina las relaciones semánticas de la combinación como todo con los ítems externos[42], mientras que el elemento no núcleo debe aportar información no disponible en el núcleo. El núcleo semántico es el componente dependiente y el elemento no núcleo es el independiente de una combinación semántica. Las *restricciones de coocurrencia*, nos referimos aquí a las restricciones de selección, se establecen, en una visión clásica, en forma de categorías semánticas a las cuales los ítems léxicos deben pertenecer. En los casos en que la anomalía se puede resolver reinterpretando la expresión lingüística compleja como metáfora, la alternativa viable es la visión prototípica de las categorías: las combinaciones no son normales o anormales, sino más o menos normales. Se habla, desde esta perspectiva, de preferencias y no de restricciones de selección.

Las vinculaciones entre las palabras coocurrentes se reflejan en la frecuencia relativa de asociación, como índice de su afinidad[43]. Además de los factores semánticos previamente analizados, la coocurrencia de las palabras es resultado de otros parámetros: factores extralingüísticos (frecuencia en el mundo extralingüístico, la importancia de un aspecto: cuanto más significativo sea, más se habla de ello); combinaciones estereotipadas, clichés (patrones por defecto), restricciones colocacionales arbitrarias (incompatibilidad entre palabras, dentro de una lengua dada, y no entre significados), afinidades no composicionales (idiomáticas).

Recogemos a la vez unas observaciones acerca de la noción de *significado de una palabra posible* o sea significados que se pueden construir y por lo tanto pueden algún día recibir una etiqueta. Así, los significados de una palabra posible están bajo la restricción de las dependencias semánticas: los elementos que constituyen el significado de una palabra deben formar una cadena continua de dependencia: debe haber una relación de dependencia entre los elementos y no debe haber vacíos en la cadena que se deban llenar por elementos semánticos de fuera de la palabra. De aquí, una combinación como N ADJ puede ser un significado de palabra, pero ADV N no lo puede ser (Cruse, 2000a: 91-92). Nosotros reinterpretamos las consideraciones de Cruse como una indicación más sobre la construcción del significado de expresiones complejas.

Concluimos con algunas ideas de Cruse de interés para la DSA, sobre la vinculación entre las relaciones sintagmáticas y las relaciones paradigmáticas. Cruse (2000a:219-235) establece la

[42] As, en una combinación entre un nombre y un adjetivo, son las propiedades semánticas del nombre las que determinan si la combinación de esta construcción con otros elementos es normal.

[43] La afinidad sería la ración entre la coocurrencia real de las dos palabras y su coocurrencia prevista en base de sus frecuencias individuales en el lenguaje. La noción de afinidad nos remite al término de ratio de asociación (en inglés, a*ssociation ratio*) usado en el PLN.

sistematicidad de las conexiones entre las relaciones sintagmáticas y las relaciones paradigmáticas y su funcionamiento en colaboración: las relaciones sintagmáticas delimitan el espacio en que las relaciones paradigmáticas operan. Así, las relaciones paradigmáticas reflejan las opciones disponibles en un punto estructural en una oración. Las relaciones sintagmáticas se establecen entre ítems que ocurren en la misma oración, particularmente entre los que se hallan en una relación sintáctica íntima. Una oración bien formada se puede ver como una cadena de elementos elegidos, cada uno, de un conjunto de posibilidades ofrecidas por la lengua. Este conjunto de posibilidades no es completamente libre sino que están restringidas por lo otros elementos de la oración. Las relaciones sintagmáticas son expresión de las relaciones de coherencia. Las relaciones paradigmáticas operan dentro de los conjuntos de opciones. Cada conjunto de opciones es la modalidad en que la lengua articula o divide un área conceptual. Hay en estas áreas conceptuales un grado mayor o menor de estructuración sistemática. Las relaciones paradigmáticas son expresión de tal estructuración.

2.4 Construcción del significado

La *construcción del significado* tiene que ver con la hipótesis, central en la semántica, de que el significado es estructurado[44]. Según el principio de Composicionalidad, el significado de las expresiones complejas[45] (nos limitamos aquí sólo a unidades superiores a la palabra, o sea a sintagmas y oraciones) es una función del significado de las expresiones simples, guiada por la estructura sintáctica.

Entramos, por lo tanto, en el terreno de la *semántica compositiva*[46]. Preferimos este término al de semántica oracional, por incorporar de manera más explícita otras estructuras lingüísticas, mayores o menores a la oración.

Sintetizamos, a continuación, unas consideraciones básicas acerca de la construcción del significado[47]. Según la semántica composicional, el significado de las expresiones lingüísticas complejas se puede explicar en base a dos ideas básicas: la productividad gramatical y la hipótesis de la composicionalidad. La productividad gramatical designa la capacidad combinatoria de las reglas de un sistema lingüístico, basada en dos propiedades de la gramática de las lenguas: la existencia de un número finito de reglas de construcción de expresiones bien formadas y la posibilidad de aplicar recursivamente estas reglas (Escandell Vidal, 2004). Según el principio de la composicionalidad**,** una parte decisiva del significado de las expresiones complejas depende, de manera estable y sistemática, de la estructura sintáctica. La construcción del significado de las expresiones complejas responde a pautas estables. Para cada regla sintáctica hay una regla de composición correspondiente[48]. Junto con el conocimiento léxico, estas reglas pertenecen a nuestro conocimiento lingüístico. Mientras una expresión compleja está formada de acuerdo con las reglas gramaticales, se puede interpretar de manera composicional.

El proceso de composición semántica se concibe como un proceso ascendente (en inglés, *bottom-up*): procede desde las unidades más pequeñas hacia las más amplias. En el proceso intervienen: el significado léxico de las expresiones básicas, las formas gramaticales de las expresiones básicas y la estructura sintáctica de las expresiones complejas. Así, los significados léxicos de las unidades más

[44] Kearns (2000, *apud* Escandell Vidal, 2004) llama *significado estructural* el significado que deriva de la organización sintáctica.

[45] Una expresión compleja es una unidad lingüística formada por la combinación de unidades simples, de acuerdo con las reglas y los principios de la gramática. Es imprescindible que cumpla el criterio de la gramaticalidad, es decir que esté construida siguiendo los patrones que marca la gramática. (Escandell Vidal, 2004:18).

[46] La Semántica Composicional se ocupa del significado de las expresiones complejas. Partiendo del paralelismo entre sintaxis y semántica (cada categoría sintáctica, cada tipo de relación estructural tiene un correlato semántico específico, la Semántica Composicional estudia cómo se proyecta la sintaxis en la semántica, cómo contribuyen al significado de las expresiones complejas los diferentes elementos que configuran la estructuras sintácticas: relaciones de dependencia, clases de palabras, unidades con significado gramatical, elementos de conexión, etc. (Escandell Vidal, 2004).

[47] Seguimos a continuación, en esta síntesis, principalmente a Escandell Vidal (2004), Cruse (2000a) y Löbner (2002).

[48] En caso contrario, la gramática produciría cadenas de palabras que sería imposible de interpretar.

pequeñas sirven como entrada para las reglas del significado gramatical, cuya salida es la entrada para las reglas de combinación asociadas a las reglas sintácticas.

El principio (en la versión fuerte) incorpora tres supuestos: 1) el significado de una expresión compleja está completamente *determinado* por los significados de sus constituyentes[49]; 2) el significado de una expresión compleja es completamente *predecible* a través de reglas generales a partir de los significados de sus constituyentes; 3) cada constituyente gramatical tiene un significado que contribuye al significado del todo (Cruse, 2000a:67).

Sin embargo, el principio de la compocisionalidad no es universalmente válido. El principio encuentra sus límites en la existencia de expresiones que se escapan a la composicionalidad, o sea cuyos constituyentes gramaticales no contribuyen con un significado identificable al significado global: a) expresiones no composicionales (*idioms*, metáforas fosilizadas, colocaciones, clichés); b) aspectos no composicionales de las expresiones composicionales (nombres compuestos, zonas activas, categorías complejas)[50]. Por lo tanto, Cruse considera que el principio se debe ver más bien como una asunción por defecto y matiza el principio de la siguiente manera: el significado de una expresión compleja es una función composicional de los significados de sus constituyentes semánticos, es decir aquellos constituyentes que parten el conjunto de manera exhaustiva, y cuyos significados, cuando se juntan adecuadamente, llevan al significado global[51].

Las consecuencias del Principio de Composicionalidad se constituyen en una doble ventaja, de orden empírico y teórico. En plan empírico, según Larson y Segal (1995), se da cuenta de la sistematicidad y de la productividad de la comprensión. En plan teórico, el principio hace posible la explicación del significado de un conjunto infinito de expresiones complejas, a través de la combinación del Principio de Composicionalidad con la productividad gramatical. De esta manera, se sientan las bases de una *visión algorítmica del significado* de las expresiones complejas.

El Principio de Composicionalidad no especifica lo que es el significado o hasta dónde llega la especificidad del significado que está ligada a la gramática. Cruse (2000b:77-78) identifica tres enfoques respecto del Principio de Composicionalidad o, mejor dicho, acerca de las relaciones entre composicionalidad y significado:

a) El modelo de bloques de construcción (en inglés, *building-block*) o bien, las "teorías de las listas de control" (en inglés, *check-list theories*). Este modelo está íntimamente relacionado con las variantes más fuertes de la composicionalidad: el significado de una expresión se puede definir de manera finita y total por medio de procesos de composición estándar que actúan sobre los significados de sus componentes, que están también totalmente determinados.

b) El modelo del armazón (en inglés, *scaffolding*) o del 'esqueleto semántico'. Desde este punto de vista, la composicionalidad provee el esquema básico de la estructura semántica de una expresión compleja, que se completa luego por otros medios pragmáticos menos predecibles, usando conocimiento enciclopédico, contexto, etc. Esto se puede ver como una versión más débil de la composicionalidad.

c) El modelo holístico. Es también una versión fuerte de la composicionalidad. Requiere que el significado de cada ítem sea una entidad indefinidamente que consiste en sus relaciones con todas las demás unidades de la lengua. En cierto sentido, todos los efectos de la combinación con otros ítems están ya presentes en el significado: sólo hace falta tomar la parte relevante en cada caso[52].

En otras palabras, el enfoque de los "bloques de construcción" postula que la totalidad de los aspectos de la interpretación de una expresión tienen que derivarse composicionalmente a partir de un conjunto limitado de rasgos. En cambio, el enfoque del "armazón semántico" considera que el significado lingüístico representa sólo una parte de la interpretación, que se ve luego complementado por otros procesos que tienen que ver con la integración de informaciones contextuales y extralingüísticos (Escandell Vidal, 2004:32-33).

---

[49] O sea que los significados de las expresiones complejas son *totalmente* determinadas por las tres fuentes mencionadas, es decir sólo por la entrada lingüística. Por lo tanto, en particular, el proceso no se basa en conocimiento sobre el contexto extra-lingüístico (Löbner, 2002).

[50] Para detalles, véase Cruse (2000a:77-79).

[51] Para la definición de los constituyentes semánticos, véase Cruse (2000a:70-72).

[52] Hemos seguido, en parte, la traducción de la cita de Cruse (2000a) en Escandell Vidal (2004).

El principio de composicionalidad tiene su complemento en el *Principio de la Interpretación Consistente*: una expresión compleja se interpreta siempre de tal manera que sus partes vayan bien juntas y que toda la oración encaje en el contexto[53] (Löbner, 2002: 46-47, 52-53)[54]. En el caso de palabras con significado no aclarado dentro de una expresión lingüística compleja, su significado se puede inferir, en un proceso descendiente (en inglés, *top-down*), a partir del significado de la expresión más amplia, previamente derivado de manera composicional. La interpretación en contexto elimina, por lo tanto, las lecturas de las expresiones complejas con contradicciones internas y resuelve así el significado de los elementos ambiguos o conflictivos. El hecho de que la interpretación de las palabras y de las expresiones lingüísticas complejas obedece al Principio de la Interpretación Consistente determina los cambios de significado y también de las restricciones de selección. La resolución de los casos de ambigüedad y de conflicto se resolvería en este tiempo descendiente de la construcción del significado y no durante el movimiento ascendiente (en inglés, *bottom-up*) de la composición semántica en sí misma, que sería ciega a las necesidades de consistencia del contexto. En otras palabras, la construcción del significado sería un proceso a la vez ascendente y descendiente.

2.5 Coordenadas linguísticas de la DSA

Sintetizamos en este apartado el espacio teórico que fundamenta la tarea de la DSA, mediante una serie de opciones dentro de algunas distinciones básicas establecidas en la semántica teórica en relación con el significado. Por una parte, delimitamos el espacio teórico en que nos movemos y circunscribimos más estrictamente el objeto de estudio. A la vez, sintetizamos algunos resultados fundamentales de la semántica que asumimos en la tarea de DSA.

(i) Semántica. Así, la DSA trata el significado estrictamente lingüístico y no el significado pragmático. El significado lingüístico es el significado codificado gramaticalmente, mientras el significado pragmático es el significado delimitado por parámetros tanto gramaticales como no gramaticales. En otras palabras, estudia el significado de las expresiones lingüísticas en su sentido general[55], fuera de un particular contexto comunicativo. La determinación del significado de las expresiones requiere una abstracción a partir del uso de las expresiones en contextos concretos. Más bien, lo que se intenta capturar es el *potencial* de las expresiones. Esta opción sitúa la DSA dentro de la semántica en sentido estricto, independientemente de la pragmática.

(ii) Semántica léxica. La DSA supone la identificación de los sentidos que las palabras tienen en un uso concreto. Por lo tanto, de las diferentes expresiones lingüísticas, focaliza el significado de las palabras. Nos situamos, con esta opción, dentro de la semántica léxica.

(iii) Interacción del significado de las palabras con el contexto lingüístico. Además, en la DSA, la asignación de sentidos a las palabras del texto se hace en base de la información provista por el contexto. Nos aproximamos al significado de las palabras desde su interacción con el contexto lingüístico. Esto significa, primero, que tenemos en cuenta el significado de las unidades superiores a la palabra, lo que nos lleva a la cuestión de la construcción del significado (abajo, en (iv)), y, segundo, que adoptamos un enfoque contextual a la polisemia (abajo, en (v)).

Se imponen, en estas circunstancias, unos comentarios previos acerca de la noción de contexto, noción controvertida, con múltiples acepciones, y por lo tanto engañosa[56]. La investigación en la DSA actual se limita al contexto lingüístico y por lo tanto no trata el contexto pragmático[57].

---

[53] Nosotros restringiremos aquí el contexto al contexto lingüístico.

[54] La modificación de significado durante el proceso de interpretación se suele llamar también *coerción*.

[55] Lo que Löbner (2002: 4) llama *significado de las expresiones* y considera uno de los niveles de significado (los demás serían el nivel de enunciación y el nivel comunicativo).

[56] En términos de Hirst (2000).

[57] Para la definición del contexto pragmático, véase, por ejemplo, la discusión en Levinson (1983:5-35). Así, el contexto pragmático se define como aquellos rasgos que son culturalmente y lingüísticamente pertinentes en cuanto a la producción e interpretación de enunciados (Van Dijk, 1976). Estos rasgos, además de los principios universales de la lógica y del uso del lenguaje, serían los conocimientos sobre los parámetros de la

(iv) Enfoque contextual a la polisemia. En la DSA, la asignación de sentidos a las palabras del texto se hace en base de la información proveída por el contexto. Nos aproximamos, por lo tanto, al significado desde su interacción con el contexto. Es decir, adoptamos un enfoque contextual a la polisemia, considerando que el significado de una palabra está totalmente reflejado en sus relaciones contextuales y construido por éstas.

(v) Variante fuerte del Principio de Composicionalidad. En las condiciones en que DSA se suele definir como la selección de entre un conjunto de sentidos discretos y en la medida en que consideramos exclusivamente la contribución del contexto lingüístico al significado de las expresiones lingüísticas[58], se adopta una variante fuerte del Principio de Composicionalidad, el modelo de bloques de construcción[59].

(vi) Enumeración de sentidos como modelo del significado. La DSA, tal como se define actualmente, supone la selección de uno de los sentidos discretos provistos por la fuente léxica usada. Esta manera de operar se aplica tanto en los métodos "basados en conocimiento", que eligen el sentido consultando directamente la fuente, como en los métodos basados en corpus, que aprenden "los sentidos" a partir de un corpus etiquetado previamente con sentidos de una fuente léxica. Por lo tanto, a la base de la tarea de DSA en su forma actual está el modelo de enumeración de sentidos (Pustejovsky, 1995), un modelo que corresponde a una visión estática del léxico.

comunicación: el papel y la posición de los hablantes; la situación espacial y temporal; el de formalidad; el medio o sea el vehículo de la información transmitida; el contenido adecuado; el campo adecuado (Lyons, 1977).
[58] No tenemos en cuenta, por ejemplo, la información de dominio.
[59] Cf. el apartado 2.4.

3 Metodología de la DSA (I): información usada

La dificultad de la tarea de DSA explica el desarrollo de una gran variedad de métodos para su solución. Igual que otras áreas del PLN, la DSA aplica modelos y técnicas procedentes de otros campos de investigación, principalmente de la Estadística y de la Inteligencia Artificial. La elección de un método u otro depende, en parte, de la evolución del área misma y está relacionada con la disponibilidad de recursos de conocimiento útil para la resolución de la ambigüedad léxica.

El problema de la DSA se encuadra dentro del área de la Inteligencia Artificial (IA, en inglés, *Artificial Intelligence*) y, en concreto, se considera como un problema de IA completo (en inglés, *AI-complete*). Así, los primeros enfoques a la DSA -en los años setenta y ochenta- proceden de esta área de conocimiento. En este período, el conocimiento necesario para la DSA se codifica a mano y se limita a conjuntos restringidos de palabras. Los modelos que se usan son desarrollados a propósito para la tarea de resolución léxica: la semántica de las preferencias de Wilks (1972, 1973, 1975), los expertos de palabras de Small (1980) y Rieger (1982) o las palabras polaroid de Hirst (1987)[60]. Los años ochenta, pero especialmente los años noventa están marcados por la aparición de lexicones computacionales (diccionarios en formato electrónico, ontologías, etc.) y de corpus textuales. La gran cantidad de datos almacenados electrónicamente permite la utilización de técnicas de tipo estadístico. Por una parte se explotan las fuentes de conocimiento estructurado[61]; por otra, empiezan a proliferar los métodos basados en corpus, sobre todo a partir del momento en que se dispone de corpus etiquetados semánticamente (*SemCor*, *DSO*, *line*, etc.). Los corpus anotados permiten la derivación de conocimiento necesario para la resolución léxica y para este propósito se aplican técnicas del Aprendizaje Automático, que inducen el modelo preestablecido para la DSA a partir del corpus[62]. A partir de los años noventa, se investigan distintos tipos de algoritmos y de fuentes de información, con el objetivo de mejorar cualitativamente el proceso de desambiguación.

En síntesis, en el área de la DSA se encuentran enfoques desde ámbitos diversos: modelos de Inteligencia Artificial (modelos simbólicos, como los árboles de decisión o las reglas, y subsimbólicos, como las redes neuronales), modelos estadísticos (los clasificadores bayesianos, las cadenas ocultas de Markov, el modelo de la Máxima Entropía, etc.) y técnicas de Aprendizaje Automático (aprendizaje basado en ejemplos, técnicas de la Teoría del Aprendizaje Computacional, etc.).

En los tres capítulos siguientes (3, 4, 5), ofrecemos una imagen de conjunto sobre la investigación en el área de la DSA. El factor decisivo en la elección de un método u otro es el tipo de fuente de conocimiento que se usa en el proceso de resolución léxica. Además, últimamente se opina que la información tiene una importancia superior a la que aportan los algoritmos en el proceso de DSA (Pedersen, 2002; Mihalcea, 2002; Yarowsky y Florian, 2002). Por tal motivo, empezamos la presentación de la metodología desarrollada para la DSA con un inventario de las fuentes léxicas y del conocimiento de éstas explotado para la DSA (capítulo 3). Seguimos con los métodos de DSA (capítulo 4) y sobre éstos, a continuación, ofrecemos una perspectiva evaluadora y comparativa (capítulo 5).

Dedicamos el presente capítulo a la información usada en la asignación de sentidos. La DSA está condicionada por la disponibilidad de conocimiento semántico; por eso, el problema de los recursos léxicos es fundamental para la investigación en el área. Interesan en este contexto:

[60] Cf. apartado 4.2.1.
[61] Cf. apartado 4.2.2.
[62] Cf. apartado 4.3.
[65] Ver el apartado 3.1.1.3.

- los tipos de fuente usados, y su correspondiente existencia o necesidad, con las relativas técnicas de obtención;
- el conocimiento de las fuentes explotado por los algoritmos de desambiguación;
- las necesidades informativas de los sistemas de DSA y
- las tendencias actuales en la solución de los varios problemas asociados a las fuentes de información para la DSA.

En concreto, aquí nos centramos en las fuentes de conocimiento léxicas (apartado 3.1.) y en las clases de información que contienen (apartado 3.2.), todo ello desde la perspectiva de su aplicación a la DSA.

3.1 Fuentes de conocimiento léxico

Entre las fuentes de conocimiento léxico que se usan en DSA, distinguimos principalmente las fuentes de información estructuradas preexistentes (apartado 3.1.1.), las fuentes de información no estructuradas (apartado 3.1.2.) y las combinaciones de ambos tipos de fuentes. A la primera clase pertenecen obras lexicográficas elaboradas generalmente fuera del ámbito de la DSA, de uso general, que suelen tener una estructura, en la cual los sentidos de las palabras tienen asociados varios tipos de información. La segunda clase corresponde a los corpus. A diferencia de las fuentes léxicas de conocimiento estructuradas, éstos no tienen una organización en términos de unidades relacionadas a las palabras o los sentidos. Además, los sentidos de las palabras pueden o no tener información lingüística asociada.

3.1.1 Fuentes de información estructuradas

Las fuentes léxicas estructuradas se pueden clasificar en los siguientes tipos:

(a) diccionarios o lexicones;
(b) tesauros;
(c) ontologías o redes semánticas;
(d) lexicones generativos.

Se trata de una clasificación de referencia, ya que algunas fuentes reúnen las características de más de un tipo. Así, los tesauros se pueden ver como un tipo particular de lexicones; las ontologías y los lexicones a veces se diferencian poco, aunque los lexicones suelen dar prioridad a la descripción del significado, mientras que las ontologías dan prioridad a las relaciones entre los sentidos. *WordNet*[65], por ejemplo, se puede considerar desde las dos perspectivas: como lexicón o como ontología, según se focalicen las definiciones de los sentidos o bien las relaciones léxico-semánticas que existen entre los mismos.

3.1.1.1 Diccionarios accesibles por ordenador

Los diccionarios en soporte electrónico empiezan a aparecer en los años ochenta. El término más usual que se emplea para dichas fuentes es el de diccionarios accesibles por ordenador o máquina (DAO) (en inglés, *Machine Readable Dictionary*, *MRDs*)[66]. Los DAO representan la fuente léxica más estudiada y utilizada en la investigación relacionada con la DSA, y en general para la adquisición de conocimiento léxico necesario para las tareas de PLN[67]. La unidad básica de los diccionarios es el sentido, al cual le corresponden el campo de la definición y otros campos opcionales, como etimología, tema, ejemplos, sinónimos, contrarios, etc.

Los DAO tienen como mayores desventajas la inconsistencia de las definiciones, la granularidad a menudo inadecuada para la DSA y la falta de información sintagmática requerida para la determinación del sentido.

Parece que los diccionarios monolingües para estudiantes de segundas lenguas son los más idóneos para la DSA, debido a que ofrecen información sintáctica, semántica y pragmática más amplia que los demás diccionarios. A la vez, para los investigadores que tienden a dividir el significado de una palabra en unos pocos sentidos, más generales, estos diccionarios tendrían la ventaja de una

[66] Para la terminología castellana, seguimos a Rodríquez y Martí (1998).

[67] Hay una vasta bibliografía sobre la utilidad de los DAO en la lexicografía computacional y en la DSA. Nos remitimos, entre otras fuentes, a Boguraev y Briscoe (ed.) (1989), Wilks *et al*. (1996), Rodríguez y Martí (1998), Ooi (1998). Un tema recurrente en toda la bibliografía es el tema de las ventajas e inconvenientes que presenta el uso de los DAO en el PLN y en particular en la DSA.

granularidad menos fina, y así más adecuada a las tareas de PLN y a la DSA. El prototipo de esta clase de diccionarios es el *LDOCE* (*Longman's Dictionary of Contemporary English*), y es el que más se ha utilizado para la DSA. El interés que ha tenido en el ámbito de la lexicografía computacional se debe a la variedad de información que contiene -códigos temáticos, información de subcategorización, preferencias de selección básicas, etc.- y al hecho de que, por estar orientado a estudiantes de inglés como L2, tiene entradas estructuradas, de baja granularidad y con definiciones elaboradas con un vocabulario controlado.

Otra clase de diccionarios usados para la desambiguación léxica son los bilingües. Los diferentes equivalentes de traducción a otras lenguas se pueden explotar, dentro de DSA, como discriminadores de sentido.

3.1.1.2 Tesauros

Un tesauro es esencialmente una organización jerárquica de listas de palabras parcialmente sinónimas. Suele estar organizado alfabéticamente y se distingue de los diccionarios en tanto que no define los sentidos sino que se limita a proveer sus sinónimos y relaciones de hiponimia e hiperonimia[68]. El tesauro más usado en DSA es el *Roget's International Thesaurus*, que fue puesto en formato electrónico en los años cincuenta. *Roget's* provee una jerarquía explícita de entidades, con hasta ocho niveles progresivamente refinados, a base de las nociones aristotélicas clásicas de género y especie. Está constituido por 1000 categorías y cada una de las categorías contiene un número de palabras relacionadas entre sí. Generalmente, la ubicación de una misma palabra en diferentes categorías del tesauro corresponde a diferentes sentidos de la palabra. Por lo tanto, las categorías suelen discriminar los sentidos de las palabras (Yarowsky, 1992).

El uso de los tesauros para la DSA encuentra un problema en que son a menudo inadecuados para dominios particulares.

3.1.1.3 Redes semánticas[69]

Constituyen representaciones del conocimiento estructurado, organizadas como grafos con enlaces etiquetados entre nodos. Los nodos son sentidos de las palabras o clases abstractas de sentidos, mientras que los enlaces son relaciones semánticas entre los sentidos. Una de las relaciones prototípicas es la que se da entre clase y subclase. Este tipo de representación permite la herencia de propiedades, como por ejemplo las restricciones de selección, lo que es importante para la DSA. En este formalismo se pueden realizar dos tipos de operaciones: deducción simple -reglas de inferencia asociadas a las relaciones declaradas en la red- y la búsqueda de relaciones entre objetos.

Las redes semánticas se diferencian en cuanto a: a) su alcance, ya sea general o para un dominio específico; b) sus unidades, limitadas sólo a etiquetas terminológicas o bien estructuradas internamente; c) el tipo de relaciones que se definen entre sus unidades; d) el tipo de semántica, más o menos precisa y bien definida, de sus unidades y relaciones, como la herencia y otros mecanismos de inferencia.

A continuación, presentamos algunas redes semánticas representativas que se utilizan en la DSA.

WordNet (*WN*, Fellbaum (ed.), 1998). *WordNet* es una base de datos léxico-conceptual del inglés estructurada en forma de red semántica y construida manualmente. Es el lexicón relacional en formato electrónico más completo y extenso existente, comparable sólo con el diccionario bilingüe para el japonés y el inglés *EDR Electronic Dictionary*[70]. La unidad básica en que se estructura *WordNet* es el *synset*[71], un conjunto de sinónimos representando un concepto. *WordNet* se basa en un concepto débil de sinonimia, la sinonimia en contexto[72]. Los synsets se relacionan entre sí mediante relaciones

[68] La relación clase-subclase (cf. apartado 2.2.).

[69] Un término alternativo, pero no equivalente, es el de *ontología*. Según Rodríguez *et al.* (1998), no hay un consenso sobre qué es una ontología. Se puede ver como un esquema conceptual exhaustivo y riguroso de un determinado dominio. Nosotros preferimos usar el término de *red semántica.*

[70] Para detalles, consúltese el sito: http://www.iijnet.or.jp/edr/.

[71] De *synonym set* 'conjuntos de sinónimos'.

[72] Dos palabras se consideran sinónimas en contexto si se pueden sustituir sin que se modifique substancialmente el significado.

semánticas básicas, como son la hiponimia/hiperonimia[73] y la meronimia/holonimia[74]. La versión *WN 1.5* contiene 126.000 entradas, repartidas entre nombres (el 70%), adjetivos (el 15%) y verbos (el 10%), mientras que la última, *2.0*, tiene un inventario de 144.309 palabras, colocadas en 115.424 synsets.

WordNet 2.0. ha supuesto una serie de cambios respecto a las variantes anteriores, destinados a aumentar la conectividad en tres direcciones: morfología derivativa, desambiguación de términos en las glosas y agrupaciones temáticas. Así, hasta el momento se han incluido 42.000 enlaces entre pares de nombres y verbos relacionados derivativa y semánticamente. En una futura versión se añadirán enlaces morfológicos entre otras categorías sintácticas. La versión incluye también nueva terminología y una organización temática (en inglés, *topical*) para muchas áreas, que clasifica los *synsets* por categoría, región, uso, glosa. En consecuencia, los sentidos de palabras relacionados se asignan a un dominio. En cuanto a la desambiguación de términos en las glosas, se incluirán enlaces que indiquen el sentido adecuado en el contexto para cada palabra de clase abierta de las glosas.

El uso de *WN* en varias tareas o aplicaciones computacionales ha puesto de relieve un grado de polisemia excesivo, es decir la granularidad demasiado fina (Palmer, 2000; Gayral y Saint-Dizier, 1999). Esto ha motivado el desarrollo de técnicas de compresión de sentidos (p.ej., Agirre y López de Lacalle, 2004).

WordNet se ha convertido en una herramienta de referencia para la investigación computacional, siendo la más usada en los últimos años, debido fundamentalmente a su cobertura y a su potencial carácter multilingüe, a través del proyecto *EuroWordNet*. La Global *WordNet* Association (GWN)[75] ofrece el marco para estandarizar la construcción de *wordnets* para otras lenguas; un ejemplo notable es *Balkanet*[76], que extiende *EuroWordNet* con otras seis lenguas europeas: el checo, el rumano, el griego, el turco, el búlgaro y el serbio. A la vez, GWN alberga proyectos variados, diseñados para el enriquecimiento cuantitativo y sobre todo cualitativo de *WordNet* o de *EuroWordNet* (ver a continuación, en este apartado), a través de incorporación de nuevos tipos de información: *Euroterm,* que extiende *EuroWordNet* con terminología especializada en el medio ambiente; *Meaning*, ideado para la ampliación de *EuroWordNet* con corpus etiquetados a nivel de sentido, extraídos de la *WorldWideWeb* y con módulos de desambiguación; *Hamburg Metaphor Database*, para el estudio y la codificación de relaciones metafóricas en *WordNet*.

Una línea de investigación derivada del creciente interés para recursos como *WordNet* es la construcción automática de *wordnets*; un ejemplo relevante en esta dirección es la propuesta de Mihalcea y Moldovan (2001).

EuroWordNet (*EWN***,** Vossen (ed.), 1998). El propósito de *EuroWordNet* ha sido construir una base de datos léxica multilingüe para diferentes lenguas europeas, siguiendo la metodología de *WordNet*. *EuroWordNet* es una base de datos multilingüe, con *wordnets* para varias lenguas (holandés, italiano, español, inglés, alemán, francés, estoniano y checo), compatibles entre sí en cobertura e interpretación de las relaciones. Al igual que *WordNet*, se trata de una base de datos semántica de tipo relacional, en la cual el significado de cada palabra está descrito principalmente mediante sus relaciones con los significados de otras palabras. El diseño lingüístico se basa en el de *WordNet 1.5*: se construye alrededor de la noción de *synset*, con relaciones semánticas entre los *synsets*. Las relaciones semánticas son, en su mayor parte, las de *WordNet*, pero hay también cambios importantes, en vista de un modelo más general, con relaciones intercategoriales y subcategorizadas, extraíbles de manera semiautomática de diccionarios y útiles para las aplicaciones de PLN. Los *wordnets* locales están construidos de manera relativamente independiente de las estructuras específicas de cada lengua, y están conectados a un índice común (*Inter-Lingual-Index*, *ILI*). El índice interlingua es un fondo no estructurado de *synsets* -principalmente basado en *WordNet 1.5*- llamados *ILI-records.* Los *synsets* específicos de las varias lenguas que están vinculados al mismo *ILI-record* se consideran como equivalentes. La compatibilidad entre los *wordnets* fue asegurada por un núcleo de 1300 conceptos comunes (en inglés, *base concepts*), definidos a mano y que han sido cubiertos por cada lengua,

[73] V. nota 3.
[74] La relación entre una parte y un todo (cf. apartado 2.2.).
[75] http://www.globalwordnet.org/.
[76] Para detalles, consultar, por ejemplo, http://www.ceid.upatras.gr/Balkanet/files/balkanet-elsnet-ko-accept.pdf.

obteniéndose así unos *wordnets* nucleares (en inglés, *core-wordnets*) totalmente equivalentes. Luego, éstos han sido extendidos, en una estrategia descendiente (en inglés, *top-down*) y usando técnicas semiautomáticas, a partir de los siguientes recursos: principalmente diccionarios accesibles por ordenador, pero también taxonomías, diccionarios bilingües y herramientas de extracción de información de diccionarios[77].

Extended WordNet (*XWN*, Harabagiu *et al.*, 1999)[78]. Debido a que *WordNet* ha sido construida como una base de datos léxica, hay limitaciones en su uso para ciertas aplicaciones del procesamiento del conocimiento; por ejemplo, no es posible extraer palabras relacionadas temáticamente. El propósito del proyecto Extended *WordNet* es transformar *WordNet* en un formato que permita la derivación de relaciones semánticas y lógicas adicionales. La primera variante fue construida sobre *WordNet* 1.7., mientras que la última variante, *XWN2.0-1.1*, está basada en *WordNet 2.0*. La idea es explotar la rica información que está contenida en las glosas de las definiciones, la principal información que se usa por los humanos para la identificación del sentido correcto de las palabras. Para alcanzar este objetivo, el proyecto se desarrolla en tres etapas: (1) analizar las glosas a nivel morfológico y sintáctico (en inglés, *POS tagging* y *parsing*); (2) transformar las glosas en formas lógicas; (3) etiquetar, al nivel de sentido, los nombres, verbos, adjetivos y adverbios de las glosas. De esta manera, aumenta la conectividad entre *synsets* y se facilita el acceso a un contexto más amplio para cada concepto. La forma lógica es un paso intermediario entre el análisis sintáctico y la forma semántica profunda. Para su obtención, en la segunda etapa, se ha partido del árbol sintáctico de una oración y se han asignado los predicados y los argumentos a las palabras. Se han codificado los sujetos, los objetos, los adjuntos preposicionales, los nombres complejos y los modificadores. La desambiguación semántica de las glosas, en la tercera etapa, corresponde a la desambiguación de 637,067 palabras de clase abierta. Se ha usado igualmente la anotación humana y automática. El etiquetado humano se ha efectuado sobre un conjunto prefijado de glosas, en la idea de obtener un estándar para la comprobación de la calidad de la desambiguación automática. En el etiquetado automático, se han combinado dos sistemas de DSA, uno diseñado especialmente para la desambiguación de glosas y otro para la desambiguación de un texto de cualquier tipo, y se han etiquetado los casos cuando ha habido coincidencia en la asignación de sentido. El primer sistema combina una serie de heurísticas y desambigua el 64% de las glosas, con una calidad de 75%; el resto de las palabras se han etiquetado con el primer sentido. El segundo sistema desambigua las glosas con el 100% de cobertura y el 70% de precisión. Se estima que, para aproximadamente el 10% de las palabras etiquetadas con el mismo sentido por ambos sistemas, la precisión en la desambiguación es de 90%. La anotación de las glosas se ha realizado en formato XML.

WordNet Domains (Magnini y Cavaglià, 2000)**.** Otra extensión de *WordNet* (en la variante 1.6.) consiste en etiquetas de dominio, de manera similar al uso de códigos para campos temáticos en los diccionarios. En *WordNet Domains*, los *synsets* se han anotado con una o más etiquetas de dominio, seleccionadas de un conjunto jerarquizado de aproximadamente 200 etiquetas. Un *dominio* se define en esta investigación como un conjunto de palabras entre las cuales hay fuertes relaciones semánticas. El etiquetado con dominios es complementario a la información ya contenida en *WordNet*: un dominio puede incluir diferentes *synsets* o diferentes categorías sintácticas, sentidos de diferentes subjerarquías de *WordNet* o también puede agrupar sentidos de una misma palabra en un *cluster* temático. La agrupación de sentidos tiene como efecto la reducción del nivel de ambigüedad, si se desambigua respecto de un dominio. La agrupación de sentidos de *WordNet*, es de hecho, un tema saliente en la DSA (Palmer *et al*., 2002, *apud* Magnini *et al.*, 2002). Las aproximadamente 200 etiquetas de dominio se han elegido a partir de varios diccionarios y luego fueron estructuradas en una taxonomía en conformidad con la Dewey Decimal Classification (DDC). Para la anotación de los *synsets* de *WordNet* con una de estas etiquetas, se ha anotado primero, a mano, un número reducido de *synsets* de alto nivel en la jerarquía. Luego se ha usado un procedimiento automático, basado en la herencia a través de las relaciones léxico-semánticas de *WordNet* para propagar este etiquetado a todos los

[77] Para una visión completa sobre el *EuroWordNet*, nos remitimos al número especial de *Computers and the Humanities*, **32,** 1998, del cual hemos seguido aquí los artículos de Alonge *et al*., Rodríguez *et al*., Vossen *et al.*.
[78] http://xwn.hlt.utdallas.edu/

synsets alcanzables, con ciertas restricciones. Así, para las palabras que no pertenecen a un determinado dominio sino que aparecen en textos relacionados con cualquier dominio, se ha creado la etiqueta FACTOTUM. Este dominio incluye *synsets* genéricos, difícil de clasificar en un dominio particular, o bien *synsets* que aparecen con frecuencia en diferentes contextos (por ejemplo, números, días de la semana, colores, etc.), llamados *stop sense synsets*. Para el propósito de la desambiguación, se ha elegido un conjunto de 43 de etiquetas, las del segundo nivel de la jerarquía, que permiten un buen nivel de abstracción sin perder información relevante y, a la vez, reducen el problema de escasez de datos en el caso de dominios insuficientemente representados en los textos.

Una labor próxima es la de Buitelaar y Sacaleanu (2001), que determinan la relevancia de los *synsets* en *GermaNet* con respecto de un dominio determinado. A partir de *WordNet Domains*, (Vázquez *et al*., 2003) derivan un nuevo recurso léxico, Dominios Relevantes (cf. apartado 4.2.2.3.), con la misma finalidad de la DSA. Otra línea de investigación está orientada hacia la adaptación de *WordNet* o de *EuroWordNet* a las necesidades de lenguajes especializados o de diferentes tareas de PLN (por ejemplo, Basili *et al*., 1998).

WordNet con marcas temáticas y ejemplos (Agirre y López de Lacalle, 2004). Las *marcas temáticas*[79] (en inglés, *topic signatures*) son vectores de contexto asociados a sentidos de palabras o conceptos, en los cuales para cada palabra del vocabulario hay asignado un peso. La dimensión de los vectores coincide con el número de palabras en el vocabulario y, dentro del vector correspondiente a un sentido dado, los pesos de cada palabra intentan capturar su relación con el sentido. Para su obtención, se adquieren primero ejemplos de Internet asociados a cada concepto, usando el método de las palabras monosémicas relacionadas en *WordNet*, de Leacock *et al.* (1998). Así, para cada concepto, con la ayuda de estas palabras monosémicas relacionadas, se construyen preguntas (en inglés, *queries*) y se extraen los contextos reducidos (en inglés, *snippets*) ofrecidos como respuesta por Google. Se prefieren estos *snippets* a los documentos enteros en conformidad con los resultados de los experimentos previos, de (Agirre *et al*., 2001), según los cuales los contextos oracionales proporcionan marcas temáticas más precisas que los documentos enteros. Los contextos extraídos se filtran según los siguientes criterios: longitud mínima de seis palabras; número de caracteres alfanuméricos superior a la mitad de número de palabras; número de mayúsculas inferior al de minúsculas. De los ejemplos así obtenidos para un sentido dado, se extraen las palabras y su frecuencia, que se compara con los datos similares obtenidos para los demás sentidos. Se guardan las palabras que tienen una frecuencia notable; estas palabras y sus frecuencias constituyen la marca temática para el sentido. Finalmente, de manera opcional, se filtran las palabras de las marcas temáticas obtenidas para los sentidos de cada palabra a través de la marca temática de la palabra correspondiente. Esta marca es calculada sobre un corpus amplio y balanceado, en concreto *British National Corpus*. Se eliminan así palabras raras o de frecuencia baja, no relevantes para la palabra fijada. El recurso obtenido es de libre acceso[80] y uso, e incluye, para cada sentido de un nombre en *WordNet* 1.6, los ejemplos extraídos (una media de 3500 oraciones) y las marcas temáticas construidas a partir de estos ejemplos (con aproximadamente 4500 palabras).

Finalmente, mencionamos otras propuestas para el enriquecimiento de *WordNet* o de *EuroWordNet*, como la incorporación de preferencias de selección (Agirre y Martínez, 2002) y de colocaciones (Wanner *et al*., 2004).

MindNet. Una alternativa a *WordNet* y *EuroWordNet* es *MindNet* (Dolan *et al.*, 2000). Aunque tiene un núcleo derivado a partir de diccionarios, la red se construye a base de oraciones nuevas de un corpus que, una vez analizadas, se incorporan a la red. En la visión de sus autores, la dinamicidad y la continua ampliación, permitirán a *MindNet* perfilarse como un sistema de amplia cobertura. Para cada oración o fragmento de texto de entrada, el analizador produce árboles sintácticos y formas lógicas que se registran en una base de datos. Las formas lógicas son grafos orientados y etiquetados que operan una abstracción a partir de la estructura sintáctica superficial, para ofrecer una descripción de las dependencias semánticas entre las palabras de clase abierta[81]. Durante el análisis y la creación de

[79] La traducción al español del término *topic signature* es nuestra.
[80] A la dirección: http://ixa.si.ehu.es/Ixa/resources/sensecorpus.
[81] Siguiendo la bibliografía mayoritariamente de lengua inglesa, llamaremos palabras de clase abierta a las clase

formas lógicas, se identifican alrededor de veinticinco tipos de relaciones semánticas. Posteriormente las formas lógicas se invierten completamente y se propagan a través de toda la base de datos de *MindNet*, para establecer conexiones con cada una de las palabras que contienen. Debido a que se invierte toda la estructura de la forma lógica y no sólo las tripletas relacionales, *MindNet* almacena un contexto lingüístico complejo para cada palabra de clase abierta de un corpus. Esta representación codifica simultáneamente las relaciones paradigmáticas (hiperonimia, sinonimia, etc.) y sintagmáticas (localidad, propósito, objeto lógico, etc.). Las estructuras invertidas de las formas lógicas permiten el acceso a relaciones directas e indirectas entre la palabra raíz de cada estructura (en el caso de las entradas de los diccionarios, sería la palabra que se define, el *headword*) y cualquier otra palabra contenida en las estructuras. Una o más relaciones semánticas conectadas constituyen caminos entre dos palabras. A partir de dos caminos dentro de dos estructuras de formas lógicas invertidas se construye un camino extendido. Estos caminos extendidos, si están construidos de manera adecuada, tienen un gran potencial para determinar relaciones entre palabras que de otra manera no hubieran sido conectadas. A los caminos se les asignan pesos de acuerdo con su prominencia; se da preferencia a las relaciones semánticas que ocurren con una frecuencia media.

Según sus autores (Dolan *et al.*, 2000), *MindNet* difiere de *WordNet* y de las cadenas de hiperonimia "conceptual" desambiguadas, derivadas a partir de DAO. Dichas fuentes seguirían una línea de la IA que considera las palabras como unidades no útiles para el procesamiento semántico. *MindNet*, en cambio, es un objeto fundamentalmente lingüístico: sus contenidos son representaciones lingüísticas para oraciones o fragmentos de oraciones reales, obtenidas durante el análisis del corpus. Estas representaciones aportan información extremadamente aprovechable sobre el uso del lenguaje, ya que reflejan directamente las opciones léxicas y sintácticas de los autores de los textos. *MindNet* representaría la convergencia de dos líneas de investigación: la labor de tipo simbólico, de análisis sintáctico de diccionarios para la extracción de conocimiento estructurado de semántica léxica, y los enfoques estadísticos, orientados hacia la discriminación de sentidos a base de usos similares de las palabras en los corpus. A lo largo del desarrollo del sistema, sus autores se han convencido de que la desambiguación explícita de los nodos es innecesaria y no deseable, de manera que la variante actual de *MindNet* contiene, como nodos, palabras (usos) y no sentidos. La hipótesis fuerte que subyace a este diseño es que el contexto definido por un texto de entrada, junto con sus pesos dentro de la red, provee suficiente contexto de desambiguación como para filtrar los caminos incorrectos. Tal como está construido, *MindNet* permite el uso de técnicas estadísticas sobre un texto que combina la información paradigmática y la sintagmática; el resultado es que las estructuras obtenidas describen a la vez las relaciones sintácticas y semánticas entre palabras.

FrameNet (*FN*)[82]. *FrameNet* tiene sus orígenes en la teoría de la gramática de los casos (Fillmore, 1968) y se construye alrededor del concepto nuclear de *esquema* (en inglés, *frame*). El *esquema* es una representación semántica de las situaciones que involucran varios papeles conceptuales. Los papeles conceptuales se consideran elementos del esquema. De esta manera, el esquema es un constructo intuitivo que permite la formalización de las relaciones entre la sintaxis y la semántica. La semántica de los argumentos de una palabra predicativa corresponde a los elementos del esquema o de los esquemas asociados a la palabra (Fillmore y Baker, 2001)

Desarrollado en la Universidad de Berkeley, el proyecto *FrameNet* consiste en la construcción de una fuente léxica compleja para el inglés, compuesta de un lexicón, una base de datos de esquemas (*frames*) y una base de datos de anotaciones. El lexicón provee la descripción de las valencias de las palabras en la cual los componentes semánticos están expresados en términos de estructura de tipo "esquema". La base de datos de esquemas contiene descripciones de los esquemas semánticos incluyendo el nombre de sus elementos y conexiones. La base de datos de anotaciones es una colección de oraciones en que se marcan las palabras focalizadas en *FrameNet* y los constituyentes se anotan con respecto a su participación en la semántica y combinatórica sintáctica de la palabra focalizada.

La construcción de esta fuente léxica compleja supone: 1) el estudio de las palabras; 2) la descripción de los esquemas y de las estructuras conceptuales que hay detrás de éstos; 3) el examen de las oraciones que contienen estas palabras, dentro de un corpus amplio del inglés contemporáneo (en *FrameNet I*, se

---

de palabras con contenido léxico: nombres, adjetivos, verbos y adverbios.

[82] Ver www.icsi.berkeley.edu/~framenet.

ha usado *The British National Corpus*[83] para el inglés británico; en *FrameNet II*, se usan también los corpus LDC de *North American Newswire* para el inglés americano y se preve añadir otros corpus más, como *The American National Corpus*[84]); 4) el almacenamiento de las modalidades en que la información asociada a los esquemas está expresada en cada oración. La anotación es en parte automatizada[85].

El nombre de *FrameNet* fue modelado según el de *WordNet*, al cual se quería aprovechar en un principio en la construcción de *FrameNet*. *FrameNet* cumula las propiedades de un diccionario y de un tesauro, como también *WordNet*. Sin embargo, FN difiere con respecto a *WordNet* bajo varios aspectos: indica las relaciones entre palabras mediante esquemas semánticos; se basa en ocurrencias de corpus; incluye ejemplos de corpus para cada sentido de las palabras, en cada una de sus variantes gramaticales; dentro de un esquema, reconoce relaciones entre palabras de diferentes categorías.

Las entradas léxicas cubren un lema en una sola categoría gramatical (nombre o verbo) y las subentradas representan cada sentido suyo, o sea corresponden a unidades léxicas. Una subentrada léxica contiene los siguientes componentes:

1) el *headword*;
2) el dominio/esquema: el camino hacia el esquema individual de referencia (en inglés, *the individual background frame*, por ejemplo, "Communication/Argument" -en inglés, *Communication domain*, *Argument frame*-);
3) una definición (del *Concise Oxford Dictionary*);
4) una tabla con las realizaciones de los elementos del esquema: la lista de las modalidades sintácticas, en términos de funciones gramaticales y tipos de sintagmas, en que los elementos del esquema se expresan en las oraciones anotadas;
5) una tabla con los patrones de valencia: la lista de las agrupaciones de elementos de esquema y de su realizaciones sintácticas tal como se han encontrado en las oraciones anotadas;
6) oraciones anotadas: cada oración está anotada para una sola palabra focalizada y para los papeles semánticos cubiertos por los sintagmas vecinos a la palabra.

Desde la perspectiva de la DSA, *FrameNet* tiene la ventaja de ofrecer información sintáctica y colocacional, muchas veces asociada a un único sentido.

3.1.1.4 Lexicones generativos

Generalmente los sistemas de DSA explotan conocimiento de fuentes léxicas en las cuales hay una enumeración de los sentidos. En la última década se ha empezado a trabajar en DSA explotando lexicones generativos. Siguiendo la propuesta de Pustejovsky (1995), en estos lexicones la polisemia regular (metonimia, meronimia, etc.) se distingue de la homonimia y, para la primera, los sentidos relacionados no son enumerados sino generados a través de reglas que captan las regularidades en la creación de sentidos.

Corelex (Buitelaar, 1998)[86]. Corelex es una ontología y una base de datos semántica con 126 tipos semánticos, que cubre unos 40.000 nombres y define un amplio número de clases polisémicas derivadas mediante un análisis minucioso de las distribuciones de los sentidos de *WordNet*. Los tipos semánticos son representaciones subespecificadas basadas en la teoría del lexicón generativo. Corelex está construido a mano. Sin embargo, Buitelaar (idem) describe modalidades de generar entradas para Corelex a partir de un corpus, con el objetivo de crear lexicones para dominios específicos. Se demuestra así el potencial de obtener recursos de tipo generativo a una escala superior.

3.1.2 Fuentes de información no estructuradas: corpus

El uso de los corpus como fuente de información está relacionado con la evolución de la investigación empírica en lingüística. Se puede hablar de análisis manual de textos ya a finales del siglo XIX, pero en lingüística se empiezan a usar los corpus a mediados del siglo XX (Boas, 1940, Fries, 1952, *apud* Ide y Véronis, 1998). Los corpus se tratan como fuente de ejemplos y facilitan el desarrollo de los

[83] http://info.ox.ac.uk/bnc/
[84] http://www.cs.vassar.edu/~ide/anc/
[85] Actualmente se están investigando estrategias para el etiquetado totalmente automático mediante técnicas de aprendizaje automático a partir de ejemplos anotados a mano (Gildea, 2001; Gildea y Jurafsky, 2000).
[86] http://www.cs.brandeis.edu/~paulb/CoreLex/corelex.html

modelos numéricos del lenguaje. El vínculo estrecho con los métodos empíricos explica su período de declive alrededor de los años sesenta, con la aparición de la teoría de Harris (1954) y sobre todo de la gramática generativa-transformacional de Chomsky. El trabajo basado en corpus resucita en los años ochenta, debido precisamente a la aparición de los corpus de grandes dimensiones en soporte electrónico.

Los *corpus* son colecciones de textos –las usadas en el área del PLN, accesibles por ordenador-, construidas para servir una determinada función, y según unos criterios explícitos de acuerdo con un determinado objetivo[87]. Debido a que ofrecen conjuntos amplios de ejemplos para un determinado hecho lingüístico, los corpus permiten el desarrollo de modelos numéricos del lenguaje y, en consecuencia, el uso de métodos empíricos.

En el área de la DSA, los corpus constituyen una fuente de información en la línea de investigación llamada precisamente *DSA basada en corpus*. En este caso, la desambiguación se realiza mediante un algoritmo que no usa la información explícita de una fuente léxica, sino que adquiere conocimiento sobre los sentidos de las palabras a partir de un corpus. La fase de aprendizaje de los algoritmos se puede desarrollar sobre corpus anotados (aprendizaje supervisado) o no anotados (aprendizaje no supervisado)[88], de donde procede la diferencia entre DSA supervisada y DSA no supervisada. A continuación, explicamos cada uno de estos tipos de corpus con más detalle.

3.1.2.1 Corpus no etiquetados

En este caso, las unidades léxicas del corpus no tienen asociada información lingüística de ningún tipo. Aparecen ya en los años setenta. Entre los más importantes del primer período cabe mencionar: el *Brown Corpus*, publicado por Kucera y Francis en 1967, con un millón de palabras; el *Trésor de la Langue Française* (Imbs, 1971), y el *LOB (Lancaster-Oslo-Bergen)* de 1980. Otros corpus de referencia son: para el inglés británico, *British National Corpus*, una colección de 100 millones de palabras de lengua escrita y hablada, adquirida explorando una gran variedad de fuentes; para el inglés americano, *Wall Street Journal*, 1986-1992, de 550 millones de palabras; para el castellano, *Lexesp*, con 5,5 millones de palabras (Sebastián *et al.*, 2000) y *EFE* -el corpus de la agencia EFE-, de más de 70 millones de palabras.

A partir de los corpus no etiquetados a nivel de sentido, se pueden inducir automáticamente agrupaciones (en inglés, *clusters*) semánticas como etiquetas de sentido efectivas (por ejemplo, Schütze, 1992). Estas agrupaciones se pueden alinear con inventarios tradicionales de sentido de los diccionarios. Sin embargo, pueden igualmente funcionar sin esta correspondencia, sobre todo si se usan para aplicaciones secundarias como la recuperación de información, para la cual es más importante la partición de los sentidos que la elección de la etiqueta (Yarowsky, 2000a).

3.1.2.2 Corpus etiquetados

En esta categoría, incluimos corpus de tipo heterogéneo en cuanto a la metodología de anotación, el inventario de sentidos y las correspondientes etiquetas usadas para los sentidos. Una primera distinción se suele establecer entre corpus etiquetados (3.1.2.2.1.) a mano y corpus etiquetados automáticamente (3.1.2.2.2.). Una clase más de corpus etiquetados es la que se podría llamar "corpus artificiales" (Stevenson, 2003); en éstos se induce, de manera voluntaria, una ambigüedad que no corresponde a la discriminación entre sentidos (3.1.2.2.3.). Presentamos a continuación algunos casos representativos de corpus etiquetados y la metodología empleada en su desarrollo.

3.1.2.2.1. Etiquetado manual

Entre los corpus etiquetados manualmente más utilizados se hallan *SemCor*, "*line*", "*interest*" y *DSO*.

SemCor (de *SEMantic COncoRdance*, Miller *et al.*, 1993). Es el corpus más amplio anotado con sentidos. Construido sobre un fragmento del corpus *Brown* y de la novela *The Red Badge of Courage* de Stephen Craig, dentro del proyecto *WordNet*, tiene actualmente aproximadamente 350.000 palabras

[87] Para la tipología y la metodología de construcción de los corpus, nos remitimos, entre otros, a McEnery (1996).

[88] Los algoritmos de aprendizaje automático serán objeto de nuestra atención en el apartado 4.1.2.

y cada palabra está etiquetada con un concepto de *WordNet*. La metodología del etiquetado, manual, está presentada en (Landes *et al*., 1998) y en (Fellbaum *et al*., 1998). *SemCor* es el único corpus libremente disponible con todas las palabras de clase abierta etiquetadas. Este etiquetado posibilita la evaluación de los algoritmos de DSA para todas las palabras. Sin embargo, aunque cubre un gran número de palabras, contiene un conjunto muy bajo de ejemplos para cada una.

El "corpus *line*" (Leacock *et al.*, 1993). Es de dimensiones reducidas: contiene 2094 ocurrencias de la palabra *line* etiquetadas con sentidos. La limitación de este corpus es contraria a la de *SemCor*: tiene un gran número de ejemplos pero para una sola palabra.

El "corpus *interest*" (Bruce y Wiebe, 1994). Está formado por las ocurrencias etiquetadas de *interest* en 2369 oraciones, con uno de sus sentidos en *LDOCE*. Las autoras subrayan la dificultad de anotar con un conjunto de etiquetas correspondientes a una distinción de sentidos de granularidad fina.

El *DSO* (Ng y Lee, 1996). Está construido a partir de los corpus *Brown* y *Wall Street Journal*, de donde se han seleccionado 191 palabras polisémicas (121 nombres y 70 verbos) y aproximadamente 1.500 oraciones para cada una de ellas, con un total de 192.800 de oraciones. En estos ejemplos se han etiquetado manualmente sólo las ocurrencias de las 191 palabras de partida, consideradas por los autores las más frecuentes y más ambiguas en inglés.

La fiabilidad del etiquetado manual, como fuente de referencia en la anotación con sentidos, es relativa. Por ejemplo, los corpus *DSO* y *SemCor* tienen en común un fragmento del corpus no anotado del *Wall Street Journal*, y el etiquetado sobre este fragmento, con sentidos de *WordNet*, coincide sólo en el 57% (Stevenson, 2003).

3.1.2.2.2 Etiquetado automático

La obtención de corpus etiquetados a mano está limitada por el elevado coste y la dificultad de la anotación: es el llamado "cuello de botella de la adquisición de conocimiento" (en inglés, *knowledge acquisition bottleneck*) (Gale *et al*., 1992). Superar este obstáculo, o sea abrir este "cuello de botella", es una cuestión central hoy en día en la investigación relacionada con la DSA. Se han propuesto varias técnicas para la obtención automática de corpus o ejemplos etiquetados, que permitan a los clasificadores que se desarrollen y entrenen sin la necesidad de datos etiquetados manualmente[89]. Presentamos, a continuación, algunas de estas propuestas.

Se pueden identificar dos estrategias en el etiquetado automático: se anotan corpus ya existentes o bien se coleccionan conjuntos de ejemplos para cada uno de los sentidos de las palabras prefijadas.

Yarowsky (1992) usa el *Roget's International Thesaurus* (cuarta edición, Chapman, 1977) y además un corpus, la *Grolier's Encyclopedia* (en la versión electrónica de 1991), de 10 millones de palabras. La obtención de ejemplos etiquetados se hace con la ayuda de las 1.042 categorías semánticas debajo de las cuales se encuentran todas las palabras del tesauro. Para una categoría del tesauro, se extraen todas las oraciones del corpus.

En (Yarowsky, 1995), se obtienen ejemplos correspondientes a los diversos sentidos de una palabra partiendo con un conjunto reducido de colocaciones etiquetadas y utilizando un enfoque de *bootstrapping*[90]. En concreto, a partir de un corpus de 460 millones de palabras, se extraen ejemplos con las colocaciones semilla. De estos ejemplos se extraen nuevas colocaciones y de nuevo se buscan en el corpus oraciones con las nuevas colocaciones obtenidas. Además, se prueban diferentes modalidades automáticas o semiautomáticas para el etiquetadode las colocaciones iniciales. Para doce nombres, se adquieren, con este procedimiento, más de 47.000 ocurrencias, con un promedio de casi 4.000 por palabra.

[89] Hay una amplia bibliografía dedicada a métodos de adquirir información y construir bases de conocimiento de manera automática: Manning y Schütze (1999), Boguraev y Pustejovsky (1996), etc.
[90] Para la definición del concepto de *bootstrapping*, ver apartado 4.1.4.

Leacock *et al.* (1998) proponen un método basado en palabras monosémicas de *WordNet*. En el caso de una palabra dada, las palabras monosémicas relacionadas proveen claves para buscar en un corpus oraciones de entrenamiento relevantes. De esta manera se construyen conjuntos de oraciones para cada sentido. El método permite obtener resultados en la DSA superiores al uso de material etiquetado a mano, pero está limitado por la existencia necesaria, para los diferentes sentidos de la palabra, de palabras monosémicas relacionadas. Restringiendo las relaciones semánticas a los sinónimos, y a los hipónimos o hiperónimos directos, los autores encuentran que, aproximadamente, el 64% de las palabras de *WordNet* tienen palabras relacionadas en el corpus de 32 millones de palabras del *San José Mercury News.* Además, mediante un test sobre un conjunto de 1.100 palabras polisémicas, demuestran que sólo un 25% de los sentidos de las palabras polisémicas tienen palabras monosémicas relacionadas.

*Internet como corpus***.** Entre las modalidades más explotadas últimamente para obtener de manera automática datos necesarios en el entrenamiento de los sistemas supervisados está el uso de la red (*World Wide Web*). En esta línea, la estrategia cambia: no se trabaja sobre corpus previamente construidos, sino que se extraen ejemplos de la red, es decir oraciones con palabras polisémicas que corresponden a los sentidos de estas palabras. Las pocas técnicas desarrolladas hasta el momento se valen del *WordNet* en el proceso de búsqueda y extracción de estos ejemplos.

Una propuesta en esta línea es la de Mihalcea y Moldovan (1999, 2001). Los autores intentan superar las limitaciones del algoritmo de Leacock *et al.* (1998) usando otros tipos de información: las definiciones de las glosas de *WordNet* y un corpus amplio constituido por textos electrónicos almacenados en Internet. La idea básica es determinar una clave léxica, formada por una o más palabras, que identifique de manera inequívoca un sentido determinado de la palabra, y luego encontrar ejemplos que contengan esta clave. Una clave léxica asociada a un sentido dado de una palabra se crea usando los sinónimos monosémicos de la palabra o bien la definición asociada al syns*et* al cual pertenece el sentido. La adquisición de los ejemplos tiene que ser seguida por una validación manual. A partir de 10 palabras (cuatro nombres, cuatro verbos, un adjetivo y un adverbio), con un total de 75 sentidos, los autores han obtenido 49.115 ejemplos, frente a los 1.432 ejemplos para las mismas palabras existentes en *SemCor*. La comprobación manual de los ejemplos obtenidos se ha realizado sobre una muestra de 658 de ellos, con un porcentaje de 92% respuestas correctas. Por otra parte, los autores advierten que el número de ejemplos obtenido mediante este método no está relacionado con la frecuencia de los sentidos y, por lo tanto, los clasificadores que utilizan este tipo de corpus tienen que establecer probabilidades *a priori.*

Agirre y López de Lacalle (2004b) también obtienen, de Internet, un corpus para los nombres de *WordNet,* utilizando el mismo procedimiento de Leacock *et al*. (1998) con palabras monosémicas relacionadas en *WordNet*. Hemos presentado la metodología para obtener el corpus en el apartado 3.1.2.

GenCor (Mihalcea, 2002a, 2002b) es un corpus etiquetado a nivel de sentidos obtenido a través de un proceso de generación, lo que motiva su nombre. Para su construcción, se ha utilizado un algoritmo de *bootstraping*, que combina y extiende los enfoques de Yarowsky (1995) y de Mihalcea y Moldovan (1999) en la obtención de ejemplos anotados con sentidos. El algoritmo es iterativo y consiste en tres pasos principales que detallamos a continuación. 1) Se crea un conjunto-semilla formado por sintagmas nominales y construcciones verbo-objeto o sujeto-verbo, que se extraen de *SemCor* y de ejemplos anotados automáticamente con el método de Mihalcea y Moldovan (1999) usando palabras monosémicas relacionadas en *WordNet*. Las construcciones están etiquetadas al menos en una de las dos palabras de contenido léxico: el verbo y/o el nombre. 2) Se hace una búsqueda en la red para cada secuencia de palabras en el conjunto semilla. 3) En los documentos obtenidos, se desambiguan las palabras en una pequeña ventana de texto alrededor de la semilla de búsqueda (en inglés, *snippet*), aplicando el algoritmo de Mihalcea y Moldovan (2000). Para ello, se usa la vinculación de las palabras con las demás palabras en su vecindad inmediata, explotando las relaciones de hiperonimia-hiponimia de *WordNet*. Los sintagmas nominales y las construcciones de verbo seguido de nombre con las palabras desambiguadas en este paso (3) se añaden al conjunto semilla de partida y se repite el paso 2,

obteniéndose así nuevos textos que permiten la extracción de nuevas semillas. Se aplica iterativamente el proceso hasta que se obtiene un número preestablecido de ejemplos etiquetados. Evaluado en las mismas condiciones de las dos tareas de Senseval-2 para el inglés, *all words* y *lexical sample* (v. apartado 5.2.), el algoritmo ha proporcionado la obtención de corpus de entrenamiento que se han demostrado útiles para el entrenamiento de los sistemas de DSA[91].

Reetiquetación de un corpus anotado con sentidos. Ilustramos esta línea de investigación con el corpus producido por Stevenson y Wilks (2000), mediante la adaptación de dos recursos léxicos existentes: *SemCor* y *Sensus*[92]. La finalidad de este corpus es el aprendizaje y la evaluación en el caso de DSA de vocabulario no restringido. *Sensus* es una taxonomía terminológica de 70.000 nodos, concebida como un marco para la inserción de conocimiento adicional. Es una extensión y reorganización de *WordNet*, con elementos de otras ontologías y de la taxonomía extraída del diccionario *LDOCE*. La alineación entre las distintas ontologías utilizadas se ha realizado mediante unos algoritmos especiales y sus sentidos se han puesto en correspondencia, de tal forma que las palabras de contenido léxico etiquetadas con sentidos de *WordNet* han sido "traducidas" a etiquetas de *LDOCE*. La ventaja de la anotación con etiquetas de *LDOCE* es la reducción de la granularidad excesiva de los sentidos de *WordNet*. El corpus resultante contiene 36.869 palabras etiquetadas con sentidos de *LDOCE*, que, a pesar de ser bastante reducido, tiene una dimensión razonable para una evaluación de este tipo de tarea (en inglés, *all-words*), y es algunas veces más grande que los corpus de evaluación utilizados en DSA y considerados de vocabulario amplio (en inglés, la tarea de DSA de vocabulario amplio se llama *large vocabulary WSD*).

La obtención automática de datos etiquetados con sentidos plantea el problema de la fiabilidad del etiquetado. Las evaluaciones de los corpus o colecciones de ejemplos etiquetados al nivel de sentido son pocas hasta el momento. Presentamos a continuación las conclusiones de dos experimentos realizados con este objetivo. Agirre y Martínez (2000) aplican el método de Mihalcea y Moldovan (1999b), entrenando el algoritmo con listas de decisión[93] sobre el corpus obtenido usando el método mencionado e implementando luego el algoritmo sobre el corpus *SemCor*. Contrariamente a las expectativas, los resultados obtenidos en el experimento son más bien bajos e inducen cierta reserva entre los autores acerca de la utilidad real de los ejemplos etiquetados que se adquieren a través de estos métodos de etiquetado automático. Sin embargo, se resalta la necesidad de analizar las causas de este nivel bajo y de explorar modalidades para la mejora de la calidad en la adquisición automática de ejemplos etiquetados. Un análisis con esta orientación proponen los mismos autores en (Agirre y Martínez, 2004). La evidencia empírica obtenida por vía experimental muestra que los datos etiquetados con sentidos automáticamente adquiridos están altamente afectados, igual que los corpus anotados a mano, por la distribución de los sentidos de las palabras[94]: el cambio de esta distribución entre el corpus de entrenamiento y el corpus de prueba lleva a un descenso dramático en la calidad de la desambiguación.

3.1.2.2.3 Corpus artificiales

a) *Pseudo-palabras***.** Con el objetivo de obtener material sobre que se puedan entrenar y evaluar los sistemas de DSA, Yarowsky (1992) propone una estrategia para crear falsas ambigüedades en un corpus. Así, sugiere unir dos palabras fijadas en una nueva, llamada "pseudo-palabra"; posteriormente las ocurrencias de las dos palabras originales se sustituyen con la pseudo-palabra. Se crea así en el texto una ambigüedad deliberada que el algoritmo debe aprender a resolver, sustituyendo la pseudo-palabra por la palabra original adecuada. Es decir, se mira el texto con pseudo-palabras como el texto fuente ambiguo, y el original como texto con las palabras desambiguadas. Una de las alternativas para crear pseudo-palabras consiste en la eliminación de los acentos prosódicos en lenguas como el español o el francés. Igualmente para la construcción de pseudo-palabras, en (Yarowsky, 1993), se usan los

[91] En el apartado 4.4.3., presentaremos el sistema de DSA de Mihalcea y Moldovan (2001), que explota, además de *SemCor*, el corpus *GenCor*.
[92] Para detalles, consúltese el sito: www.isi.edu/natural-language/resources/sensus.html.
[93] Ver el apartado 4.3.1.2.
[94] El *bias*, en términos de los autores.

homófonos, mientras que, en (Yarowsky, 1994), se usan los homógrafos creados quitando acentos a las palabras en español y francés.

Según Yarowsky (2000b), el método de las pseudo-palabras tiene un potencial ilimitado de crear datos etiquetados para el entrenamiento y la evaluación. Además, permite obtener datos etiquetados que tengan un grado variable de ambigüedad, a través de la selección de pares de palabras con el grado adecuado de similitud semántica, distribución temática y frecuencia. Sin embargo, Wilks y Stevenson (1997) consideran que estas ambigüedades léxicas no son verdaderas distinciones de sentido, ya que no corresponden a la polisemia real de los textos. La labor sobre este tipo de ambigüedades habría comprometido buena parte de la investigación en DSA supervisada (Manning y Schütze, 1999).

b) Distintas traducciones en corpus bilingües alineados. Un ejemplo de referencia es el corpus *Canadian Hansard*, constituido por las actas del parlamento canadiense, que se publican simultáneamente en francés y en inglés. El corpus ha sido usado para la DSA, entre otros, por Brown *et al.* (1991) y Gale *et al*. (1992). Efectivamente, en (Gale *et al*., 1992) se buscan primero palabras inglesas con dos traducciones distintas al francés; luego se coleccionan separadamente los contextos de las ocurrencias que la palabra inglesa tiene cada una de las dos traducciones; finalmente se obtienen dos colecciones de contextos para los dos sentidos de la palabra inglesa que reflejan las traducciones.

El uso de corpus bilingües para crear colecciones de datos se enfrenta, primero, con la falta de corpus bilingües y, aún más, alineados. Segundo, la traducción no es siempre un buen indicador para la discriminación entre sentidos porque los equivalentes en otra lengua de los distintos sentidos de una palabra pueden ser iguales, o sea la polisemia puede ser paralela en diferentes lenguas y transferirse con la traducción (Leacock *et al*., 1998). Este problema encuentra una solución en el uso de lenguas de familias muy diferentes (Yarowsky, 2000b) o bien en el uso de más lenguas a la vez. Un ejemplo en esta última dirección es Tufis (2004), que aprovecha el corpus multilingüe (ocho lenguas) construido para la novela *1984* de George Orwell.

c) Construcción de fuentes adecuadas de conocimiento léxico. Debido a los problemas que plantea el uso de las fuentes léxicas existentes para la DSA, una tendencia actual es construir lexicones computacionales o bases de conocimiento léxico adecuados a las necesidades de la tarea de desambiguación. Agirre y Martínez (2001a) identifican dos orientaciones en la labor de DSA:

1) partir de las fuentes léxicas existentes y extraer de éstas el conocimiento que pueda ser útil para la DSA: diccionarios (DAO), ontologías, o bien corpus, anotados o no;

2) partir de los tipos de conocimiento considerados útiles o necesarios para la DSA como líneas-guía en la construcción de bases de conocimiento léxico para sistemas de DSA.

Si bien todavía domina la primera orientación, se nota, según Agirre y Martínez (2001a), un desplazamiento desde sistemas basados en fuentes de información hacia sistemas basados en fuentes de conocimiento, con un correspondiente interés en las posibilidades para el enriquecimiento semiautomático de las bases de conocimiento léxico a partir de recursos existentes. Sin embargo, esta tendencia tiene que fundarse en un análisis contrastivo de los resultados obtenidos a partir de diferentes tipos de conocimiento y de algoritmos propuestos para la DSA. Este análisis es útil para entender qué tipos de conocimiento son útiles para la DSA y por qué algunos sistemas de DSA tienen mejores resultados que otros. Esto permitiría, a la vez, como alternativa a las bases tradicionales de conocimiento léxico elaboradas principalmente a mano[95], construir de manera semiautomática fuentes que contengan el conocimiento requerido por tales sistemas. Desde esta perspectiva, la sistematización propuesta en (Agirre y Martínez, 2001a) (cf. apartado siguiente, 3.2.) sería una base de partida en idear tales procedimientos semiautomáticos.

3.2 Tipos de conocimiento útiles para la DSA

Una vez presentadas las diferentes fuentes de conocimiento utilizadas por los sistemas de DSA y los métodos automáticos para su obtención, trataremos en este apartado la información útil que se puede obtener de cada una de ellas para el desarrollo del proceso de desambiguación.

[95] Lo que significa una tarea ingente y que plantea muchos problemas, dada la semántica profunda y rica que estos sistemas deberían contener.

El proceso de DSA necesita información sobre las palabras ambiguas. Se suelen usar dos tipos principales de conocimiento: conocimiento general (sobre el significado) y conocimiento específico (sobre los usos de la palabra). En otras palabras, el conocimiento utilizado en la DSA se clasifica, según la relación o no con el contexto, en dos categorías:

a) conocimiento independiente del contexto, que corresponde a una caracterización general de los sentidos;

b) conocimiento dependiente del contexto, o sea información sobre la categorización de los sentidos para apariciones particulares de una palabra.

En el primer caso, las fuentes de información suelen ser recursos léxicos estructurados con conocimiento léxico explícito (diccionarios, tesauros, diccionarios bilingües, lexicones); en el segundo caso se explotan los corpus para adquirir tal conocimiento que se encuentra implícitamente en el texto.

Según el tipo de conocimiento utilizado, se delimitan dos enfoques a la DSA: uno fundado en fuentes léxicas estructuradas, llamado *DSA basada en conocimiento* (en inglés, *knowledge-based WSD*), y otro orientado hacia colecciones de textos, llamado *DSA basada en corpus* (en inglés, *corpus-based WSD*)[96]. Sin embargo, los dos enfoques no se excluyen mutuamente, ya que hay casos que combinan los dos tipos de conocimiento (sistemas híbridos). Dicha estrategia de usar información de más recursos léxicos permite reducir la cantidad de corpus necesario para el entrenamiento (cf. EAGLES, 1998). Al igual que en otras tareas del PLN, como el etiquetado morfosintáctico o el análisis sintáctico, la DSA se m2a osistemaResnik e Yarowsky, -120; -12.72 TD -0.0104 Tc 440107 Tw 4

4c. *Tema*, como entre *bat* ‘palo’ y *baseball.*
4d. *Preferencias de selección***.** Son relaciones verbo-argumento, básicamente entre agente y/o tema y predicado. Por ejemplo, entre *dog* ‘perro’ y *bite* ‘morder’ en *the dog bite the postman* ‘el perro ha mordido al cartero’.

5. *Claves (*en inglés, *clues) sintácticas***.** Se trata de información de subcategorización, p.ej. transitivo *vs.* intransitivo en el caso de los verbos. Así, *eat* ‘comer’ es intransitivo en el sentido de *‘take a meal’*, construcción que correspondería a **comer una comida*, pero es transitivo en otros sentidos.

6. *Dominio*. Se hace referencia a áreas concretas de conocimiento. P.ej., en el dominio de los deportes, se prefiere el sentido específico de *racket* ‘raqueta’ como *‘tennis racket’* ‘raqueta de tenis’.

7. *Frecuencia de los sentidos***.** Sobre los cuatro sentidos de *people* ‘gente‘, el sentido general corresponde al 90% de las ocurrencias en *SemCor*.

Fiabilidad del conocimiento según la fuente de procedencia. Agirre y Martínez (2001a) realizan una serie de experimentos para analizar la utilidad de los diferentes atributos para la DSA, cuyos resultados sintetizamos a continuación:

- las *colocaciones* son indicadores fuertes si se aprenden de corpus etiquetados manualmente;
- la *información taxonómica* es muy débil;
- las *asociaciones semánticas* relacionadas con un tópico o una situación son fuertes cuando se aprenden de corpus etiquetados manualmente, aunque de difícil delimitación entre ellas; también son útiles las asociaciones aprendidas de DAO;
- las *claves sintácticas* son fiables cuando se aprenden de corpus etiquetados manualmente;
- las *preferencias de selección* son fiables cuando se aprenden de corpus etiquetados a mano, pero su aplicabilidad es relativamente baja; queda por comprobar si en el caso de que fueran adquiridas de corpus no etiquetados los resultados cambiarían;
- el *sentido más frecuente* es igualmente un indicador fuerte que depende de la disponibilidad de datos etiquetados.

Las conclusiones parecen confirmar las observaciones de McRoy (1992) de que las colocaciones y las asociaciones semánticas de palabras son los tipos de conocimiento más importantes para la DSA y, por otra parte, de que las restricciones de selección son aplicables en menor grado a la DSA. Además, Agirre y Martínez (2001a) observan que las claves sintácticas son igualmente útiles.

Respecto de las fuentes léxicas que proporcionan dichos tipos de conocimiento, Agirre y Martínez (2001a) consideran que los corpus etiquetados a mano parecen ser la mejor fuente para la adquisición automática de todos los tipos de conocimiento estudiados en los experimentos desarrollados (colocaciones, asociaciones semánticas, claves sintácticas, restricciones de selección, el sentido más frecuente), excepto el conocimiento taxonómico, extraído de ontologías.

El conocimiento sobre qué tipo de información es el más adecuado para la DSA permitirá, por una parte, enriquecer adecuadamente las fuentes existentes y, por otra parte, organizar las fuentes léxicas utilizadas de manera adecuada para obtener combinaciones más potentes en la tarea de desambiguación.

3.3 La información contextual en la DSA

Varios autores han estudiado cuál es el papel del contexto en la tarea de DSA. Así, Ide y Véronis (1998) ven la DSA como la mejor asociación (en inglés, *match*) entre el contexto de la palabra que se quiere desambiguar y la fuente de información. Pero, según ellos, el papel del contexto en el proceso de DSA es una cuestión todavía por resolver. En el enfoque contextual, la idea es hallar, para una palabra dada, el sentido que más se aproxima al sentido de sus vecinos. La base del enfoque es la inducción del sentido desde la similitud contextual. Por lo tanto, hay que medir el parentesco o la distancia semántica entre los sentidos de las palabras que aparecen en el texto y guardar la combinación que minimiza la distancia global (Habert *et al.*, 1997; Schütze, 1998). Una variante es la codeterminación de los sentidos en un enfoque global de la DSA (Resnik, 1995 o Agirre y Rigau, 1996).

El enfoque tiene como justificación la existencia de cierta evidencia: la similitud contextual también juega un papel en la categorización humana (Schütze, 1998). Según Miller y Charles (1991), los humanos determinan la similitud semántica de las palabras a base de la similitud de los contextos en que se usan. Los últimos autores formulan dos hipótesis que rigen este proceso:

> a. *Hipótesis contextual fuerte.* Dos palabras son semánticamente similares en la medida que sus representaciones contextuales son similares, es decir, la similitud semántica está determinada por el grado de similitud entre los conjuntos de contextos en los cuales las palabras se pueden usar.
>
> b. *Hipótesis contextual para los sentidos* (que puede entenderse como una extensión de la hipótesis anterior). Dos ocurrencias de una palabra ambigua pertenecen al mismo sentido en la medida que sus representaciones contextuales son similares. Es decir, los sentidos se basan en la similitud contextual: un sentido es un grupo de ocurrencias (*tokens*) de la palabra en cuestión con contextos similares, o sea un grupo de ocurrencias de una palabra que son similares contextualmente.

El contexto es el principal medio para la identificación del sentido de una palabra polisémica; como tal, toda la labor de DSA se funda en el contexto de la ambigua (en inglés, llamada también *target*) con el fin de proveer información para su desambiguación (Ide y Véronis, 1998). Además, para los métodos guiados por los datos (en inglés, *data-driven*), principalmente los basados en corpus, la información sobre el uso de los sentidos de una palabra en contexto se usa como información de referencia para la desambiguación y se obtiene fundamentalmente mediante el entrenamiento sobre un corpus. O sea, los contextos de una palabra en un corpus de entrenamiento proveen también el conocimiento previo con el cual el contexto particular de una ocurrencia de la palabra está comparado para realizar la desambiguación.

Según señalan Ide y Véronis (1998), el uso del contexto en la DSA es una de las cuestiones abiertas de esta tarea. Intentamos a continuación presentar la problemática del contexto en el área de la DSA, siguiendo principalmente a los autores citados. Se suelen distinguir dos clases principales de contexto que se usan para la selección del sentido: contexto local y contexto global**.** Nos detenemos en presentar los diferentes tipos de información, relacionados con cada clase de contexto, que se pueden usar en la DSA.

3.3.1 Contexto local

El *contexto local* o *microcontexto* es el más usado en DSA. Consiste en una ventana de unidades léxicas en un texto alrededor de la ocurrencia de la palabra que se quiere desambiguar; varía desde algunas pocas palabras hasta toda la oración. El contexto local ofrece información variada, de utilidad en la DSA. Algunos parámetros que se han utilizado en la DSA son: la distancia, las colocaciones y la información sintáctica. Detallamos estos parámetros a continuación.

El concepto de *distancia* está relacionado con el número de palabras (*n*) que se incluyen en el contexto. Se han dado respuestas muy variadas sobre el valor óptimo de *n*, es decir, la distancia. Así, los experimentos de Kaplan (1950), Choueka y Lusingan (1985) (*apud* Ide y Véronis, 1998, y Miller y Leacock, 2000) han evidenciado que los contextos de dos palabras son altamente fiables para la desambiguación, incluso los contextos de una palabra son fiables en ocho casos de cada diez. Pero hay pocos estudios sobre el problema frente a la gran variedad de la dimensión de las ventanas en la labor de DSA. Yarowsky (1993, 1994) examina varias ventanas de microcontexto y pares de palabras de conjuntos cerrados (en las posiciones -1 y -2, -1 y +1, +1 y +2 respecto de la palabra focalizada). Su conclusión es que el valor óptimo de *k* para la dimensión de la ventana contextual varía con el tipo de ambigüedad: la ambigüedad local necesita el valor de *k* igual a 3 o 4, mientras que la ambigüedad relativa a un tema (en inglés, *topic-based*), requiere de 20-50 palabras. Por esta razón, no se propone ninguna medida: para palabras ambiguas distintas, las relaciones de distancia distintas más eficientes serán distintas. En cambio, para Leacock *et al.* (1998), los distintos tests muestran los mejores resultados con una ventana local con tres palabras de clase abierta (nombres, verbos, adjetivos, adverbios) en ambas direcciones, derecha e izquierda.

b) Las *colocaciones* se están usando ampliamente en la labor de DSA[100]. Según Allen (1995) o Ide y Véronis (1998), hay evidencia psicológica de que las colocaciones están tratadas de manera diferente a otras coocurrencias. En el caso de las colocaciones frecuentes con palabras ambiguas, las palabras que aparecen primero introducen las palabras ambiguas, es decir forman el *contexto asociativo* (p.ej. *iron-steel*) de dichas palabras ambiguas, activándolas en las tareas de decisión léxica. En cambio, las palabras preparativas que constituyen el contexto temático, es decir, relaciones determinadas por la situación, escenario o guión (p.ej. *plane-gate*), no facilitan la decisión léxica de los sujetos para elegir el sentido adecuado. Yarowsky (1993) trata explícitamente el uso de colocaciones en la tarea de DSA[101]. Examina una variedad de relaciones de distancia y, tomando en consideración la adyacencia (p.ej., el primer nombre a la izquierda), observa que en caso de ambigüedad binaria hay un único sentido por colocación (la hipótesis de *one sense per collocation*). Es decir, en un determinado contexto, una palabra se usa con un solo sentido con una probabilidad de un 90-99%[102]. Martínez y Agirre (2000) matizan la hipótesis de Yarowsky, afirmando que funcionaría dentro de un mismo corpus, pero no de un corpus a otro, o sea se vería afectada por la variación de género o temática del corpus.

c) *La información sintáctica* que se utiliza para la DSA es variada: categoría morfosintáctica de las palabras del contexto, relaciones sintácticas, sintagmas en que participa la palabra ambigua, etc. Las relaciones sintácticas, usadas ya por Earl (1973), se limitan comúnmente a relaciones argumentales; excepciones notables son Stetina *et al*. (1998) o Martínez *et al*., (2002) entre otros. El procesamiento del texto a nivel sintáctico no es necesariamente completo, sino que se pueden utilizar técnicas de análisis parcial o superficial (en inglés, *shallow/partial parsing*). Por ejemplo, Hearst (1991) segmenta el texto en SN, SP, SV, eliminando todo el resto de información sintáctica y examina los ítems dentro de segmentos de tres sintagmas a la izquierda y a la derecha desde la palabra ambigua. En unos experimentos más amplios, Martínez *et al.* (2002) muestran que la información sintáctica permite una alta calidad en la asignación de sentidos.

Yarowsky (1993) determina varios comportamientos basados en la categoría sintáctica: los verbos derivan más información de sus objetos que de sus sujetos; los adjetivos derivan casi toda la información de los nombres a los cuales modifican y los nombres están desambiguados por los adjetivos o por los nombres inmediatamente adyacentes. Recientemente, se suele utilizar la información que proporciona la categoría sintáctica junto con otros tipos de información (McRoy, 1992; Bruce y Wiebe, 1994; Leacock *et al.*, 1998) y se hace evidente la necesidad de recurrir a diferentes tipos de procedimientos de desambiguación, idea que ya encontramos en el enfoque con expertos en palabras, en función de la categoría sintáctica y de otras características de la palabra ambigua (Yarowsky, 1993; Leacock *et al.*, 1998). Los resultados de Mihalcea (2002b) parecen indicar también la complejidad variable de la tarea de DSA entre las diferentes categorías morfosintácticas: los nombres necesitarían el menos número de atributos, los adjetivos más y los verbos un número aún superior.

La utilidad del contexto local para la asignación de sentido fluctúa de un experimento a otro; sin embargo, los experimentos amplios realizados en los últimos años (Martínez y Agirre, 2000; Martínez *et al*., 2002) parecen indicar que el contexto local es muy efectivo y permite la obtención de una buena calidad en la asignación de sentido. Por otra parte, Mihalcea (2002b) afirma que los atributos más usados en el proceso de desambiguación han sido la forma flexiva de la palabra ambigua dentro del

[100] En Firth (1951), la colocación no es una simple coocurrencia, sino que debe tener un carácter habitual, usual. Halliday (1961) da una definición más adecuada desde una perspectiva computacional: la colocación es la asociación sintagmática de ítems léxicos, cuantificable, en el texto, como la probabilidad de que los ítems*a,b,c,*… aparezcan a *n* desplazamientos (una distancia de *n* ítemsléxicos) desde un ítem *x*. Berry-Rogghe (1973) usa el término *colocación significativa* (en inglés, *significant collocation*): una asociación sintagmática entre ítemsléxicos, donde la probabilidad de que un ítem *x* co-ocurra con los ítems*a,b,c,*… es mayor que la simple casualidad (cf. Ide y Véronis, 1998).

[101] Yarowsky define la colocación como coocurrencia de dos palabras en alguna relación definida.

[102] En variadas ocasiones, los experimentos de DSA han ofrecido evidencias que contradicen esta hipótesis. Véase a continuación, en este mismo apartado, y en el apartado 4.1.4.3.

contexto, su categoría sintáctica, las palabras en la ventana contextual y las colocaciones locales. Los varios experimentos realizados hasta la actualidad indican también que el uso del microcontexto ofrece mejores resultados para los verbos o adjetivos que para los nombres (Schütze, 1998; Leacock *et al*., 1998). Finalmente, la ventaja que presenta el microcontexto sobre otros tipos de información es que su uso es menos costoso que el de los conocimientos que requieren un procesamiento más complejo (Yarowsky, 1992).

3.3.2 Contexto global

Se trata, en este caso, de palabras de contenido léxico que coocurren con un sentido dado de una palabra dentro de una ventana amplia (Ide y Véronis, 1998). El contexto global se puede diferenciar en contexto temático y contexto de dominio[103].

Los métodos que se basan en el *contexto temático* (en inglés, *topical context*) explotan la redundancia en un texto, o sea el uso repetido de palabras semánticamente relacionadas en un texto sobre un determinado tema. El contexto temático se ha usado primero en la recuperación de información y se ha incorporado recientemente en la labor de DSA. Las dimensiones de las ventanas usadas varían de un método a otro (Ide y Véronis, 1998). Así, Yarowsky (1992) analiza una ventana de 100 palabras para derivar clases de palabras relacionadas como contexto a la palabra polisémica que se trata; Voorhess *et al.* (1995) consideran ventanas de dos oraciones; Gale *et al.* (1993) tratan un contexto de 50 palabras a la izquierda y 50 a la derecha, y afirman que las palabras más cercanas contribuyen más a la desambiguación, pero que se obtiene una mejora de los resultados de entre el 86% y el 90% ampliando el contexto de seis palabras a 50 para cada parte de la palabra por desambiguar. Según Ide y Véronis (1998), no está claro si la distinción microcontexto *vs.* contexto temático es significativa. Es posible que sea más útil tratarlos como un todo y considerar el papel y la importancia de la información contextual como una función de la distancia con respecto a la palabra por desambiguar.

El *dominio* fue usado por primero vez para la DSA en los microglosarios desarrollados en los trabajos iniciales de Traducción Automática. La DSA basada en el dominio es implícita en varios enfoques procedentes de la Inteligencia Artificial. En 1977, Schank basándose en los *guiones* (en inglés, *scripts*), asocia palabras a sentidos basándose en el guión activado por el tema general del discurso, aunque asocia sólo el sentido de una palabra relevante al dominio del discurso. Granger (1977) utiliza la información en los guiones y la representación de la dependencia conceptual entre las oraciones para determinar el significado de las palabras del texto totalmente desconocidas. Es decir, examina el dominio y la evidencia contextual para determinar el significado, de manera semejante a los métodos usados en muchos casos de DSA basados en la IA (cf. Ide y Véronis, 1998).

Ambas visiones sobre el contexto global tienen como fundamento la asunción de que se da un único sentido por texto (en inglés, *one sense per discourse*) (Gale *et al.*, 1

(Skorochod'ko, 1972; Morris, 1988; Morris y Harris, 1991), de aquí la aparición de métodos explotando esta observación a la segmentación de un texto en subtemas (cf. Leacock *et al.*, 1998).

El uso del dominio en la desambiguación léxica conoce un nuevo impulso a partir de Magnini y Stapparava (2000) y Magnini y Cavaglià (2000). En estos trabajos, los dominios se asocian a *WordNet* y se propone la desambiguación de los dominios de las palabras (*Word Domain Disambiguation, WDD*) como una variante de la DSA. Este enfoque tiene como fundamento la asunción que el dominio establece relaciones semánticas entre los sentidos de las palabras útiles en el proceso de desambiguación y, como implicación, la reducción de la granularidad de los sentidos. El uso de etiquetas de dominio asignadas a los *synsets* de *WordNet* permite alcanzar un buen nivel de desambiguación (Magnini *et al.*, 2002). Vázquez *et al*. (2003) aportan una mejora a la calidad de la desambiguación usando los Dominios Relevantes[104].

3.3.3 Modalidades de explotar el contexto en la DSA

En cuanto a las modalidades de explotar el contexto en la DSA, hay dos aproximaciones, que corresponden a dos importantes enfoques teóricos en el NLP estadístico: el enfoque "bolsa de palabras" y el enfoque relacional.

En el enfoque "*bolsa de palabras*" (en inglés, *bag-of-words*), el contexto está constituido por algunas palabras en una ventana alrededor de la palabra focalizada (en inglés, *target*). Se trata de un simple grupo, sin tener en cuenta las relaciones con la palabra focalizada ni la estructura sintáctica. En este caso, un algoritmo muy usado es la clasificación bayesiana (propuesta por Gale *et al*., 1992) que consiste en tratar el contexto como una bolsa de palabras donde el algoritmo no selecciona ningún atributo (en inglés, *feature*), sino que combina la evidencia de todos los atributos.

Justeson y Katz (1995) hablan de la insuficiencia e inadecuación del modelo “bolsa de palabras” para muchas distinciones de sentido: si se usa como información la simple coocurrencia de las palabras de un texto, se reconocen sólo sentidos asociados con diferentes dominios. La insuficiencia de este modelo para muchas distinciones de sentido, no limitadas a dominios particulares, está compensada, con buenos resultados, a través de la combinación de la información tópica con la colocacional (Leacock *et al.*, 1998).

En el enfoque basado en la *información relacional*, el contexto es considerado en base de determinadas relaciones con la palabra focalizada (en inglés, *target*): distancia hasta dicha palabra, propiedades sintácticas, preferencias de selección, propiedades ortográficas, colocaciones (en inglés, *phrasal collocations*), categorías semánticas, etc. Un algoritmo de referencia es el propuesto por Brown *et al.* (1991). Este enfoque procede de la teoría de la información y, a diferencia de la clasificación bayesiana, mira sólo un atributo informativo en el contexto, que puede ser sensible a la estructura del texto, pero hace falta seleccionar con cuidado el atributo de entre los muchos posibles (Manning y Schütze, 1999).

En este capítulo, hemos tratado la problemática de la información usada en la DSA. Hemos presentado los principales tipos de fuentes léxicas explotadas para la DSA, tanto estructuradas (DAO, tesauros, redes semánticas, etc.) como no estructuradas (corpus u otras colecciones de datos). En el caso de los datos etiquetados con sentidos, hemos descrito la metodología empleada para su obtención. A continuación, hemos sintetizado las clases de información útiles para la tarea de DSA. Por fin, hemos repasado la cuestión del contexto en el área de la DSA.

[104] Ver apartado 4.2.2.3.

4 Metodología de la DSA (II): métodos

En este capítulo, presentamos los enfoques de referencia a la tarea de DSA. Abrimos la presentación con unos preliminares metodológicos referentes a: la clasificación de los métodos de DSA, el aprendizaje automático y la distinción entre métodos de DSA supervisados y no supervisados, los problemas de los métodos basados en corpus y soluciones posibles, estrategias para la combinación de información en la DSA y medidas para la evaluación de los sistemas de DSA (apartado 4.1.). Esta relación nos proporciona una organización para la presentación de los sistemas de DSA a lo largo de este capítulo. Así, dedicamos una primera parte a los métodos basados en fuentes de conocimiento léxico estructuradas: métodos que utilizan representaciones del conocimiento léxico de dimensiones reducidas, procedentes de la Inteligencia Artificial, y representaciones del conocimiento léxico de grandes dimensiones (como lexicones y bases de datos léxicos) (apartado 4.2.). Seguimos con la exposición de los métodos basados en corpus, etiquetados o no con sentidos (apartado 4.3.). Finalmente, presentamos unos casos relevantes de métodos mixtos que combinan ambos tipos de fuentes léxicas y/o diferentes algoritmos (4.4.).

4.1 Preliminares metodológicos

4.1.1 Clasificación de los métodos de DSA

La gran variedad de métodos y la diversidad de clasificaciones propuestas dificultan el intento de ofrecer una imagen sintética de la labor desarrollada hasta la actualidad en la DSA. Los principales criterios que se suelen tener en cuenta a la hora de clasificar los sistemas de DSA son múltiples:

- el área de donde proviene la técnica (métodos estadísticos, lingüísticos, o de la Inteligencia Artificial);
- el nivel o tipo de conocimiento que se maneja (métodos lingüísticos *vs*. estadísticos, simbólicos *vs*. numéricos o también simbólicos *vs*. subsimbólicos);
- la fuente de conocimiento léxico usada (métodos basados en datos o corpus *vs.* métodos basados en conocimiento);
- la necesidad, o no, de datos etiquetados con sentidos;
- la cobertura del método (restringida o ilimitada), etc.

Sin embargo, todas estas clasificaciones son de carácter orientativo, ya que la realidad es bien distinta: las grandes categorías propuestas, DSA estadística *vs.* DSA basada en el conocimiento, simbólica *vs.* numérica, etc., no son dicotómicas, sino complementarias. Es más, interfieren, son difusas. En la práctica, los sistemas de DSA actuales suelen ser mixtos desde algún punto de vista y su inclusión en una de las dichas categorías se debe a la dominancia de un enfoque sobre el otro. A menudo la atribución de un método de DSA en una u otra clase varía según el aspecto que se enfoca.

Los intentos taxonómicos se complican aun más debido a controversias abiertas, como: los límites entre conocimiento simbólico y conocimiento subsimbólico o las diferencias entre conocimiento numérico y conocimiento estadístico, la dificultad de establecer una delimitación bien definida entre lingüística basada en datos y lingüística guiada por los datos, etc. Las distinciones sobrepasan en complejidad el marco del presente trabajo[105].

Nos detenemos, a continuación, a comentar algunas de las clasificaciones más usadas.

[105] Para una discusión sobre la relación entre estadística y lingüística en PLN, enviamos, entre otros, a (Klavans y Resnik, 1997) y a (Rodríguez, 2001). Para la diferencia entre lingüística basada en datos y lingüística guiada por datos, véase, p.e. (Sinclair, 1995, *apud* Ooi, 1998). Para más información sobre la controversia entre conocimiento simbólico y subsimbólico, véase Moisl (2000).

La división que con más frecuencia se establece es entre métodos basados en fuentes de conocimiento léxico pre-existentes y métodos que adquieren este conocimiento a partir de un corpus. Es una práctica casi general llamar a estas categorías *métodos basados en conocimiento* (en inglés, *knowledge-based*) y *métodos basados en corpus* (en inglés, *corpus-based*) respectivamente (Ide y Véronis, 1998, Manning y Schütze, 1999, Yarowsky, 2000b, Rigau, 2002, Wilks y Stevenson, 2003, etc.). La terminología alternativa para las dos clases es *sistemas de DSA de carácter racionalista* y *empírico* respectivamente. Creemos encontrar un equivalente de esta oposición en los dos enfoques principales que identifica Agirre (1998) en la DSA: los sistemas *knowledge-based*, con conocimiento intensivo, y los sistemas *knowledge-poor*, con un uso extensivo del conocimiento. Dentro de los sistemas de DSA basados en conocimiento, se suele distinguir entre los primeros sistemas, procedentes de la Inteligencia Artificial, *de cobertura limitada* (*sistemas de juguete*), y los que usan fuentes de conocimiento léxico estructuradas, *de amplia cobertura*. Los métodos basados en corpus se diferencian, ellos también, según el *corpus explotado* sea *etiquetado* o *no* a nivel de sentido.

Consideramos en parte impropia la formulación de la delimitación entre "métodos basados en corpus" y "métodos basados en conocimiento". Aunque la terminología se justifique desde una perspectiva técnica, para hacer referencia al uso de una fuente de conocimiento léxico explícito, y no textual, implícito, desde la perspectiva lingüística la delimitación no es sostenible en estos términos. Tanto en las fuentes léxicas estructuradas como en el corpus hay conocimiento lingüístico: en el primer caso, se ofrece información sobre el dominio que un hablante ideal tendría sobre su lengua a nivel léxico; en el segundo caso, se provee información sobre la producción escrita de los hablantes de una lengua. En términos chomskyanos, se trataría de conocimiento sobre la *competence* (léxica) y la *performance* respectivamente de los hablantes de una lengua. Nos parece, por lo tanto, más adecuado delimitar entre *métodos basados en fuentes estructuradas de conocimiento léxico* y *métodos basados en datos* o *en corpus*.

Un criterio muy importante en la caracterización de los sistemas de DSA es su dependencia de un corpus etiquetado con sentidos, lo que equivale a la dependencia de la intervención humana en el proceso de resolución léxica. Desde este punto de vista, se distingue entre *métodos supervisados* y *métodos no supervisados*. La delimitación, usada inicialmente para los sistemas de DSA basados en corpus, se ha extendido a los métodos basados en fuentes estructuradas de conocimiento léxico, a las que ha incorporado como métodos no supervisados[106].

Desde la perspectiva de la cobertura, los sistemas de desambiguación se separan entre los que sirven de modelo sólo para muestras de palabras, y los que pueden ser probados plenamente en textos de dimensiones reales. La distinción suele denominarse, en la bibliografía de lengua inglesa, *lexical-sample paradigm vs. all-word paradigm*. Los primeros sistemas de DSA enfocan el problema para un conjunto limitado de palabras, mientras que los segundos tratan un vocabulario amplio, no restringido. Obviamente, es un aspecto fundamental desde la perspectiva de la aplicabilidad de los sistemas a textos reales, que ha influido en la evolución misma del campo de la DSA. La categorización entre métodos que cubren una muestra de palabras y métodos de cobertura teóricamente ilimitada se ha utilizado en las tres ediciones del ejercicio Senseval (capítulo 5).

De acuerdo con el foco de interés de nuestra tesis –el conocimiento lingüístico en la DSA-, adoptamos la taxonomía inducida por la procedencia del conocimiento usado para la resolución léxica. Hablaremos de dos categorías principales:

1) *métodos basados en fuentes estructuradas de conocimiento léxico* y
2) *métodos de DSA basados en datos o en corpus*,

y además de las fórmulas de compromiso entre las dos:

3) *métodos mixtos*.

Nuestro objetivo es cubrir, con las primeras dos clases, la variedad tipológica de las fuentes, tipos de conocimiento y algoritmos usados para la DSA, y, con la tercera clase, inventariar las formulas

[106] Los términos de *supervisado* y *no supervisado* conocen cierta variabilidad dentro del área de la DSA. Detallamos la cuestión en el apartado 4.1.3.

complejas para la resolución léxica, que combinan diferentes fuentes léxicas o algoritmos[107]. La clasificación nos permite una presentación de los sistemas de DSA neutra respecto de las vacilaciones teóricas y terminológicas (como, por ejemplo, alrededor de los términos *de métodos supervisados* y *métodos no supervisados*) y respecto de nuestras reservas frente a la distinción *de métodos basados en conocimiento vs. métodos basados en datos*. Encontramos una confirmación de nuestra opción por las fuentes de conocimiento como eje de la taxonomía en (Manning y Schütze, 1999), según los cuales tiene más sentido identificar las fuentes de conocimiento que los sistemas necesitan que encuadrarlos en clases prefijadas.

4.1.2 Aprendizaje Automático

La DSA, como otras tareas relacionadas con la comprensión del lenguaje, requiere una gran cantidad de conocimiento lingüístico y conocimiento general sobre el mundo. La adquisición y la codificación de todo este conocimiento es uno de los mayores impedimentos en el desarrollo de sistemas robustos para la DSA, tal como lo es para toda tarea del procesamiento del lenguaje natural. Los métodos de aprendizaje automático permiten potencialmente la automatización de la adquisición de este conocimiento a partir de corpus anotados o no anotados.

El Aprendizaje Automático (AA; en inglés, *Machine Learning, ML*) se define como el estudio de los sistemas computacionales que mejoran sus resultados para alguna tarea en base a la experiencia (Mitchell, 1999, Langley, 1996, *apud* Màrquez, 2002). Los métodos de aprendizaje automático tratan generalmente la tarea de clasificación de ejemplos (descritos por un conjunto de atributos) en una de varias categorías disjuntas dadas. Por esta razón, estos métodos se llaman a veces *clasificadores*. Al sistema de aprendizaje se le da un conjunto de ejemplos de entrenamiento usualmente supervisados, a los cuales les está asignada la categoría correcta. A partir de estos ejemplos, se determinan los valores de los atributos relevantes para cada categoría, lo que equivale a establecer un modelo de las categorías, y a partir de esta información el sistema debe producir un procedimiento para la correcta categorización de futuros ejemplos (Mooney, 2003).

En otras palabras, los sistemas de aprendizaje automático asumen una representación del conocimiento y consisten en un algoritmo que induce esa representación a partir de ejemplos de entrenamiento y que a la vez usa el conocimiento adquirido para la clasificación de futuros casos. Las representaciones del conocimiento (modelos) pueden ser árboles de decisión, reglas, clasificadores basados en casos, etc. (Mooney, 2003). Las diversas técnicas de AA se pueden clasificar según distintos parámetros: el conocimiento que se adquiere (simbólico o subsimbólico), la forma de aprendizaje (supervisado o bien no supervisado), el modelo en que se basan (estocásticos o de razonamiento inductivo) (Màrquez, 2002). Sin embargo, es difícil establecer clases claramente definidas según estos parámetros, ya que las combinaciones de los mismos son innumerables. Màrquez (2002) propone clasificar los métodos de AA según la línea de investigación seguida.

Frecuentemente, el área del AA se restringe al aprendizaje simbólico, o sea a los métodos que representan el conocimiento adquirido de forma declarativa, simbólica. Estos métodos son opuestos a los métodos de entrenamiento estadísticos o con redes neuronales, de orientación más numérica. En particular, el término de AA hace referencia en este caso a métodos que representan el conocimiento adquirido en forma de modelos simbólicos como árboles de decisión, reglas lógicas o casos almacenados. En el presente trabajo usaremos el sentido amplio del AA, incluyendo todo tipo de técnica que extrae conocimiento o un modelo para DSA a partir de corpus textuales. Siguiendo a autores como Màrquez (2002), consideramos tanto las técnicas propias del área del AA como las prestadas de la Inteligencia Artificial o de la Estadística.

Los métodos de aprendizaje automático suelen provenir de varias áreas de conocimiento: se pueden basar en modelos estocásticos o de la Inteligencia Artificial. A la vez, dentro de la misma área del

[107] Resaltamos que, en esta división, hemos seguido la delimitación de la esfera de DSA operada en el capítulo 1: el análisis gramatical (morfológico y eventualmente sintáctico) es más bien un preprocesamiento de la desambiguación semántica, que se limita a la desambiguación exclusiva entre sentidos de la misma categoría morfosintáctica. Todos los sistemas de DSA suelen recurrir a un etiquetado previo con categorías morfosintácticas, en el texto por desambiguar y también en el corpus de entrenamiento, en el caso de los métodos basados en datos. Desde esta perspectiva, nos distanciamos de los trabajos que consideran los sistemas que usan información sintáctica añadida a los corpus como métodos mixtos (Rigau, 2002; Montoyo, 2002).

Aprendizaje Automático, se han desarrollado métodos de la Teoría del Aprendizaje Computacional, métodos para la combinación de clasificadores o métodos semi-supervisados. Tomando como criterio clasificador las líneas de investigación, se distinguen las siguientes familias de métodos de aprendizaje automático (Màrquez, 2002)[108]:

Métodos basados en aprendizaje estadístico. Su objetivo es aprender aquellos modelos estocásticos que más adecuadamente modelan los datos, es decir, que describen el proceso que con más probabilidad ha generado los datos o ejemplos observados. Un modelo estocástico describe un proceso bajo la forma de una red o grafo probabilístico que representa las dependencias probabilísticas entre las variables aleatorias implicadas en el proceso: cada nodo en el grafo constituye una variable aleatoria y tiene asociada una distribución de probabilidad. A partir de estas distribuciones individuales se puede calcular la distribución conjunta de los ejemplos observados. En otras palabras, los modelos estadísticos proporcionan la probabilidad de que un hecho dado sea de una determinada categoría según los valores que los atributos del hecho tienen. Las probabilidades del modelo se obtienen a partir de un conjunto de hechos observados y de los valores de sus atributos. El conocimiento probabilístico así adquirido permite asignar a un nuevo caso la categoría con mayor evidencia estadística. Las diferentes aproximaciones varían con respecto a la forma en que se construye o aprende la red probabilística y con respecto al método utilizado para combinar las distribuciones de probabilidad individuales (Dietterich, 1997; Rigau, 2002). Entre los modelos estocásticos más utilizados mencionamos: las redes bayesianas, las cadenas ocultas de Markov, el principio de Máxima Entropía o la búsqueda del atributo más informativo.

Métodos tradicionales de la Inteligencia Artificial. Por tradición, la Inteligencia Artificial suele distinguir entre conocimiento simbólico, que es interpretable y manejable por los humanos, y conocimiento subsimbólico, que no lo es. De igual manera, se diferencia entre representaciones de conocimiento de tipo simbólico y de tipo subsimbólico. Una ventaja potencial de los métodos de aprendizaje simbólico sobre los métodos subsimbólicos es que el conocimiento adquirido está representado en una forma más fácil de interpretar por los humanos y más similar a las representaciones usadas en los sistemas desarrollados manualmente. Este conocimiento interpretable permite potencialmente una mayor introspección en los fenómenos lingüísticos, la mejora del conocimiento adquirido a través de la intervención humana y una más fácil integración con los sistemas desarrollados a mano (Mooney, 2003). Los sistemas de DSA usan a menudo reglas organizadas en árboles o listas de decisión. Otros métodos de tipo subsimbólico son la inducción de programas lógicos, el aprendizaje guiado por el error basado en transformación y el aprendizaje basado en ejemplos (Márquez, 2002).

Métodos de la Teoría del Aprendizaje Computacional. La Teoría del Aprendizaje Computacional (en inglés, *Computational Learning Theory*) trata conceptos clave del aprendizaje visto como proceso computacional y las relaciones que se establecen entre ellos. Por ejemplo, se buscan métodos para determinar el tamaño óptimo de las muestras de aprendizaje, la complejidad del espacio de hipótesis, el grado de acierto con que se puede aproximar el concepto objetivo, la forma de presentación de los ejemplos de aprendizaje, etc. Los algoritmos desarrollados en esta área en los últimos años han tenido notable éxito y se han trasladado también al PLN. Algunos de los algoritmos representativos son: los separadores lineales, como *Winnow*, *SnoW* (Littlestone, 1988; Roth, 1998), los algoritmos de *boosting*, como *AdaBoost* (Freund y Schapire, 1996; Schapire y Singer, 1999), y los algoritmos llamados *Support Vector Machines* (Cristianini y Shawe-Taylor, 2000; Vapnik, 1995), entre otros.

Combinación de clasificadores. La combinación de diversos modelos de clasificación aprendidos sobre un mismo problema puede mejorar notablemente el rendimiento de cada uno de ellos por separado, especialmente si los modelos son independientes y complementarios (Ali *et al.*, 1996). Un área del aprendizaje automático se dedica especialmente al estudio de los conjuntos de clasificadores y sus diversas formas de combinación (Dietterich, 1997; Màrquez, 1999).

[108] Para las referencias bibliográficas de este apartado, consúltese (Màrquez, 2000).

Métodos semi-supervisados. Estos métodos constituyen un intento de superar el "cuello de botella" de los ejemplos supervisados. El progreso en el área de aprendizaje automático se ha visto a menudo supeditado al lento y costoso desarrollo de colecciones de ejemplos lo suficientemente amplias para poder aplicar estos métodos con suficiente fiabilidad. Los métodos semi-supervisados pretenden aprender modelos fiables a partir de colecciones relativamente pequeñas de ejemplos supervisados, en algunos casos combinadas con grandes colecciones de ejemplos no supervisados, mucho más fáciles de obtener que los primeros. Dentro de esta familia se encuentran los populares algoritmos de *bootstrapping*, y también algunos métodos más generales, como el algoritmo de *Expectación-Maximización* (McLatchlan y Krishnan, 1987) el *co-training* (Blum y Mitchell, 1998) o las *Transductive Support Vector Machine*s (Joachims, 1999).

4.1.3 Métodos supervisados *vs.* métodos no supervisados

Una cuestión terminológica con incidencia en la categorización de los sistemas de DSA es la distinción entre *métodos supervisados* y *métodos no supervisados*[109]. La distinción proviene de la Inteligencia Artificial, específicamente del área del Aprendizaje Automático, y está relacionada con la existencia de un conjunto de clasificaciones para los datos de entrenamiento. Según Duda y Hart (1973), en el aprendizaje supervisado se conoce el estado real (aquí, la etiqueta de sentido) para cada conjunto de datos sobre los que se hace el entrenamiento, mientras que en el no supervisado no se conoce la clasificación de los datos en la muestra de entrenamiento. En otras palabras, en el aprendizaje automático se establece un conjunto de clases potenciales *a priori* al proceso de aprendizaje. Cuando se aplica a la DSA, esto significa que el algoritmo de desambiguación está provisto de un conjunto de sentidos. En el caso del aprendizaje no supervisado, el conjunto de clases no está predeterminado sino que las clases se infieren a través del análisis de los datos y la identificación de agrupaciones (*clusters*) de casos similares. En DSA, esto significa que el algoritmo no está provisto de un conjunto de sentidos *a priori* al análisis de corpus sino que éstos se infieren del texto *a posteriori* (Quinlan, 1986, Mitchell, 1997). En consecuencia, el aprendizaje no supervisado se puede ver a menudo como una tarea de agrupación (en inglés, *clustering*) y el aprendizaje supervisado, como una tarea de clasificación (*function-fitting*)[110] (Manning y Schütze, 1999).

Si se adopta la acepción procedente de la AA, un sistema de DSA es supervisado si se debe entrenar sobre un corpus etiquetado con sentidos, y es no supervisado si hace uso de un corpus no etiquetado con sentidos. Sin embargo, los calificativos *supervisado* y *no supervisado* no se usan siempre de esta manera.

La variabilidad de estos términos tiene mucho que ver con el tipo de etiquetas con que puede estar anotado el corpus de entrenamiento. En la acepción clásica del término de "aprendizaje supervisado", las etiquetas deben ser sentidos. Pero si las etiquetas son de distinta naturaleza, entonces se trata sólo de discriminación de sentidos (Schütze, 1998). Muchos de los enfoques no supervisados usan una fuente de conocimiento secundaria (como la jerarquía semántica conceptual de *WordNet*) para ayudar a la obtención de la estructura a partir de los datos no etiquetados (*bootstrapping*)[111]. Estos métodos se pueden considerar justificadamente no supervisados, ya que se basan en fuentes de conocimiento independientes, sin directa supervisión del fenómeno que se está aprendiendo. Las necesidades de conocimiento semántico explícito son parcialmente cubiertas: o bien se posee un corpus reducido etiquetado o bien el tipo de anotación del corpus no está orientado expresamente hacia la DSA y las etiquetas no son sentidos. Yarowsky (2000b) propone el término relativamente conservador de "semi o mínimamente supervisado" para hacer referencia a esta clase de algoritmos. A pesar de cierta variabilidad en la bibliografía, generalmente se suele hablar de sistemas de DSA supervisados si se entrenan con un corpus etiquetado con sentidos y de sistemas de DSA no supervisados si el corpus de entrenamiento tiene otro tipo de anotación o no tiene ninguna (Resnik y Yarowsky, 2000; Yarowsky, 2000b, etc.).

[109] Seguimos en la presentación de esta controversia a Manning y Schütze (1999), Stevenson (2003), Yarowsky (2000), Rigau (2002). Para las referencias bibliográficas de otras fuentes aquí citadas, enviamos a estos trabajos.
[110] Este proceso consiste en la extrapolación de la forma de una función en base de unos puntos de datos.
[111] Ver el apartado próximo, 4.1.4.

Por otra parte, en la DSA la cuestión importante no es tanto la existencia o no de un conjunto de clases (en este caso, sentidos), sino la cantidad de datos de entrenamiento desambiguados requeridos por el algoritmo. La obtención de datos etiquetados con sentidos se considera actualmente el mayor obstáculo para la DSA, de manera que la dependencia o no de estos datos se ha convertido en el criterio principal para la categorización de los sistemas de DSA, extendiéndose también a los métodos que usan fuentes de conocimiento léxico estructuradas. En consecuencia, actualmente se tiende hacia la clasificación de los métodos de DSA en sistemas supervisados, que necesitan ejemplos de entrenamiento etiquetados con sentidos para cada palabra por desambiguar, y sistemas no supervisados, que no necesitan tales ejemplos.[112] A pesar de ciertas críticas, motivadas por la confusión que puede producir la redefinición de los términos de *supervisado* y *no supervisado* con respecto a su sentido originario en el Aprendizaje Automático (por ejemplo, en (Stevenson, 2003)), la nueva acepción para sistemas supervisados y no supervisados parece imponerse (Resnik y Yarowsky, 2000; Yarowsky, 2000b; Rigau, 2002; Montoyo, 2002, etc.). Además, los mismos organizadores de Senseval han adoptado la distinción en esta última acepción, ampliada a todos los sistemas de DSA (Kilgarriff y Rozenweig, 2000; Edmonds y Cotton, 2002).

4.1.4 Problemas de los métodos basados en corpus. Soluciones[113]

Dedicamos este apartado a las dificultades con las cuales se enfrentan los sistemas de DSA basados en corpus. Así, describimos con detalle los problemas de escasez de datos (apartado 4.1.4.1), dificultad en la adquisición de conocimiento léxico (apartado 4.1.4.2), transportabilidad y adaptación (apartado 4.1.4.3). Preparamos así la presentación de los sistemas de DSA en los apartados siguientes (4.2, 4.3, 4.4).

4.1.4.1 Escasez de datos[114]

Uno de los principales problemas relacionados con la adquisición de conocimiento a partir de los corpus es la escasez de datos (en inglés, *data sparseness*), es decir la falta de ocurrencias sobre determinadas palabras (las menos frecuentes) o de determinados sentidos de las palabras polisémicas. Las palabras y sus ocurrencias con sus diferentes sentidos presentan una variabilidad y una distribución muy dispersa, de modo que se necesitan cantidades enormes de texto para obtener información sobre todos los sentidos de las palabras polisémicas. Aún así, es probable que no se encuentren muchas de las posibles coocurrencias de una palabra polisémica. En consecuencia, el problema de la dispersión de datos es crucial para la DSA. En general, para cualquier sistema en que los parámetros deben estimarse a partir de una muestra, se plantean problemas relacionados con la falta de representatividad de algunos casos, la escasa fiabilidad de los casos poco frecuentes y la fiabilidad nula de los casos que no aparecen en el corpus utilizado.

Entre las diferentes soluciones al problema de la falta de datos cabe señalar los modelos de *suavizado* y los basados en clases o en similitud, la interpolación de modelos y las técnicas de *back-off*.

Los *modelos de suavizado* (en inglés, *smoothing*) intentan solucionar, dentro de los algoritmos de DSA que usan probabilidades, los casos poco frecuentes o no observados en los datos de entrenamiento. Estos modelos tienen como objetivo que los casos no vistos se consideren con probabilidad nula, descontando, de la probabilidad que se asigna a los casos estimados por el algoritmo, una cantidad que se distribuye entre los casos no representados. Por esta razón, este tipo de técnicas se denominan también *técnicas de descuento*. Una forma sumamente simple pero efectiva de implementar este tipo de técnica es suponer que los casos que no han aparecido en el corpus de

[112] En (Rigau, 2002), la distinción entre métodos supervisados y métodos no supervisados reside en la necesidad o no de la anotación manual de un corpus de entrenamiento. A nuestro entender, en este caso el criterio de delimitación entre las dos categorías es la intervención humana.

[113] Seguimos, en este apartado, principalmente a Ide y Véronis (1998), a Rodríguez (2001) y a Màrquez (2002). Para las referencias citadas, remitimos a estos trabajos.

[114] La terminología española vacila entre *dispersión de datos* y *escasez de datos*. En nuestra opinión, entre los dos conceptos hay una relación de causa-efecto. Sin embargo, la distinción es mínima y a menudo irrelevante. En este trabajo utilizaremos ambos términos indistintamente.

estimación han aparecido al menos una vez, es decir sumar 1 a cada contador. Algunas de las propuestas más conocidas son (Good, 1953), (Jelinek y Mercer, 1985), (Church y Gale, 1991).

En los *modelos basados en clases*, las observaciones sobre el corpus no se hacen sobre palabras individuales sino sobre clases de palabras que pertenecen a una misma categoría. Además de resolver parcialmente el problema de la escasez de datos, los métodos basados en clases eliminan la necesidad de datos etiquetados. Sin embargo, la hipótesis de que todas las palabras de una misma clase se comportan de manera similar es excesivamente fuerte y determina pérdida de información. Para la delimitación de clases, se han tomado como base: las propiedades distribucionales de las palabras en un corpus (Brown *et al.*, 1993; Pereria y Tishby, 1992; Pereira *et al.*, 1993), la taxonomía de *WordNet* (Resnik, 1992), las categorías del *Roget's International Thesaurus*, los códigos temáticos del *LDOCE* (Slator, 1992; Liddy y Paik, 1993), conjuntos conceptuales construidos a partir de las definiciones del *LDOCE* (Luk, 1995), los dominios de *WordNet Domains* (Magnini y Cavaglià, 2000; Vázquez *et al*., 2003).

Los *modelos basados en similitud* explotan la misma idea de juntar observaciones para palabras similares, pero sin agruparlas en clases. Cada palabra tiene potencialmente un conjunto distinto de palabras similares. Para el cálculo de la similitud, se usa una métrica, como se usa también en muchos modelos basados en clases. Algunas propuestas de este tipo son (Rada *et al*, 1989), (Dagan *et al.*, 1993), (Dagan *et al.*, 1994), (Lin, 1997), (Karov y Edelman, 1998), (Leacock y Chodorow, 1998), (Resnik, 1998).

La *interpolación de modelos* consiste en combinar varios modelos probabilísticos en uno mixto, de manera lineal. Es decir, se suman los modelos simples, con una contribución al modelo final que puede variar mediante un coeficiente. Así Charniak (1993) propone un modelo que combina linealmente 1-gramas, 2-gramas y 3-gramas[115]; el modelo resultante de la interpolación contiene los parámetros de los modelos simples y los pesos de cada uno de ellos:

$$P(w_i \mid w_{i-2}, w_{i-1}) = l_1 \cdot P(w_i) + l_2 \cdot P(w_i \mid w_{i-1}) + l_3 \cdot P(w_i \mid w_{i-2}, w_{i-1})$$[116].

Yarowsky (1994, 1995) propone el uso de estrategias de aprendizaje interpolado para superar problemas de escasez de datos en la construcción de listas de decisión.

En el caso de las técnicas denominadas de *back-off*[117], se utilizan varios modelos de diferente granularidad, es decir, se prueban sucesivamente modelos de diferente granularidad, cada vez con menos aciertos. Para efectuar una selección, se intenta aplicar el modelo más preciso y, en caso de no encontrar suficiente evidencia estadística para el nivel de fiabilidad deseado, se pasa a utilizar un modelo menos preciso, como, por ejemplo, pasar de un modelo de 3-gramas a uno de 2-gramas y así sucesivamente hasta llegar a 0-grama[118]. Resnik (1993) y Leacock *et al.* (1998) usan técnicas de *back-off* para la DSA.

---

[115] Los *n*-gramas o modelos de Markov de orden (n-1) modelizan secuencias de *n* palabras de un texto, a base de la hipótesis simplificadora de Markov de que la ocurrencia de una palabra está determinada por las (*n-1*) palabras previas. Una presentación general de los modelos de Markov y de su uso en el PLN ofrecen, entre otros, Manning y Schütze (1999),191-227.

[116] Hemos notado:
- por *P(a)*, la probabilidad de aparición de *a*; lo que significa la probabilidad de aparición de la palabra *a* o, lo que es lo mismo, del 1-grama *a*;
- por *P(a|b)*, la probabilidad de *a* condicionado por *b*; lo que significa la probabilidad que aparezca la palabra *a* cuando aparece la palabra *b*, o, lo que es lo mismo, que aparezca el bigrama *(a,b)*;
- por *P(a|b,c)*, la probablidad de *a* condicionado por *b* y *c*; lo que significa la probabilidad que aparezca la palabra *a* cuando aparecen las palabras *b* y *c*, o, lo que es lo mismo, que aparezca el trigrama *(a,b,c)*.

[117] *Back off* significa 'dar marcha para atrás'.

[118] 0-grama corresponde a la frecuencia simple para la aparición de cualquier palabra en un texto, sin tomar en cuenta la forma particular de la palabra.

4.1.4.2 Dificultad de la adquisición de conocimiento léxico

Otro problema es la falta de datos etiquetados a nivel de sentido para los métodos de aprendizaje supervisado, el llamado "cuello de botella" de la adquisición del conocimiento léxico (en inglés, *knowledge acquisition bottleneck*, Gale *et al*., 1992). El ritmo de progreso de la DSA supervisada está crucialmente subordinado al desarrollo de grandes baterías de ejemplos etiquetados. Si en el apartado 3.1.2.2. hemos presentado algunos métodos para la obtención automática de datos etiquetados, mencionamos aquí algunas estrategias posibles para solucionar el problema usando las fuentes de conocimiento léxico disponibles.

Así, una opción es dar a los algoritmos un buen punto de partida mediante *distintas fuentes de conocimiento*, como diccionarios y como textos bilingües alineados.

En otros métodos, como el *bootstrapping*, el sistema se nutre con unos pocos datos etiquetados, pero estos datos aumentan en cantidad, a través de aprendizaje posterior de datos no etiquetados. Estos métodos se pueden considerar justificadamente no supervisados, ya que se basan en fuentes de conocimiento independientes, sin una supervisión directa del hecho que se está aprendiendo. Yarowsky (2000b) llama esta clase de algoritmos *semi* o *mínimamente supervisados*. El algoritmo *Expectación-Maximización* (EM, propuesto por McLatchlan y Krishnan, 1987, *apud* Màrquez, 2002, y aplicado en DSA, entre otros, por Gale *et al.*, 1992) constituye un ejemplo de algoritmo de *bootstrapping*. Empieza con unas estimaciones iniciales de los parámetros del modelo (que se suelen

ieza, yaperes del mos sea genonsidlo estos dobse serviodecf.uelen
Màrquez, 2002, -0.96 0 TD 0 Tc 0 Tw ( )5Tj 205.96 -12.6 TD ( ) Tj 0 -12.72 TD086.0322 48-0.0486

tópico de los corpus serían dos parámetros a tener en cuenta en los modelos de la DSA. Estas conclusiones quedan reflejadas en los experimentos con sistemas supervisados de Escudero *et al.* (2000b), quienes afirman que la DSA es muy dependiente del dominio de aplicación y que, para asegurar la portabilidad de los sistemas es imprescindible la aplicación de algún procedimiento de adaptación al nuevo dominio.

4.1.5 Estrategias para la combinación de información en la DSA

A diferencia del etiquetado morfológico (en inglés, *POS-tagging*), en el caso de la DSA parece que se obtienen mejores resultados cuando se utilizan diferentes tipos de información, de modo que la desambiguación semántica se puede considerar como altamente especializada a nivel léxico, necesitando en realidad desambiguadores especializados para cada palabra polisémica (Resnik y Yarowsky, 1997). Los distintos que usan los sistemas de DSA se suelen llamar *atributos* y el conjunto de estos tipos de información, *espacio de atributos*.

La combinación de varios tipos de conocimiento se debe hacer según unos principios y una metodología. En Inteligencia Artificial, los tipos de información que no pueden resolver un problema en su totalidad se llaman *fuentes de información "débiles"* (Newell, 1973, *apud* Wilks *et al.*, 1990). En PLN, Wilks *et al.* (1990) proponen combinar estos tipos de información "débiles" para obtener *métodos "fuertes"* de procesamiento general de texto. Con fines de DSA, McRoy (1992) y más reciente Ng y Lee (1996) son los primeros en proponer una metodología explícita, que consiste en juntar un número determinado de tipos de información sobre un fenómeno y combinarlos de manera rigurosa. A continuación, sintetizamos algunos de los enfoques notables a este problema de la DSA.

En el sistema de McRoy (1992), la combinación de las diferentes clases de información se realiza mediante la suma ponderada de sus resultados; es decir, cada clase asigna en parte un valor numérico a cada uno de los sentidos de la palabra. Los valores numéricos de la votación de cada tipo de información se determinan previamente a mano y finalmente se elige el sentido con más puntos.

Otra estrategia sencilla es la *clasificación bayesiana*, que proviene de la estadística. Ésta usa toda la información disponible, asumiendo la independencia de los atributos (Gale *et al.*, 1992).

Ng y Lee (1996) enfocan el problema desde el *aprendizaje basado en ejemplos*, conocido también como *aprendizaje del vecino más próximo*. Así, para una palabra por desambiguar, se elige el sentido del ejemplo más similar a él entre los casos etiquetados previamente, que el sistema conoce.

Leacock *et al.* (1998) aplican la estrategia de *back-off* (Resnik, 1993), para la combinación de diferentes tipos de información disponibles. En este caso, las clases de información se prueban en orden descendiente de fiabilidad, hasta que se encuentra alguna correspodencia (en inglés, *matching*).

En los trabajos Wilks y Stevenson (1997), Stevenson y Wilks (2000, 2001) y Stevenson (2003), se desarrolla un marco para la combinación sistemática de diferentes tipos de información en DSA. Según la utilidad que se les da para la DSA, las clases de conocimiento se pueden separar entre filtros y etiquetadores parciales (en inglés, *partial taggers*). Los filtros eliminan sentidos que no son probables, reduciendo así la complejidad de la tarea de desambiguación. En cambio, los etiquetadores parciales hacen uso cada uno de una fuente distinta de conocimiento y sugieren sentidos que pueden ser correctos. Así, un tipo de conocimiento se considera filtro o etiquetador parcial según su capacidad de asignar sentidos correctos. Si es improbable que elimine el sentido correcto, se puede implementar como *filtro*; en caso contrario, es más adecuado que se use como *etiquetador parcial* (Wilks y Stevenson, 2000). En (Wilks y Stevenson, 1997), los sentidos reciben votos de parte de cada etiquetador parcial y finalmente se elige el sentido más votado. Alternativamente, en (Stevenson y Wilks, 2001) y (Stevenson, 2003), los resultados de los etiquetadores parciales se combinan con la ayuda de un algoritmo de aprendizaje, lo que permite identificar la combinación óptima entre varias clases de conocimiento en el proceso de desambiguación. Es más, el algoritmo mismo de combinación puede ser objeto de estudio para la selección del algoritmo que maximiza los resultados (Stevenson y Wilks, 2001). A cada ocurrencia anotada de entrenamiento se le asocia una serie de vectores, uno para cada sentido de la palabra. Es decir, los sentidos de la palabra se representan como vectores de

atributos. En un vector se registran la información de las fuentes léxicas usadas, la salida de los etiquetadores parciales, la forma flexiva de la palabra en el texto, las colocaciones y una indicación de si el sentido es el correcto. Los casos nuevos por etiquetar se representarán en el mismo formato de tipo vectorial. Los vectores de sentido de las ocurrencias nuevas se comparan con los vectores de sentido de las ocurrencias de entrenamiento y se elige el sentido correspondiente al vector más próximo encontrado.

Un enfoque distinto a la combinación de varios tipos de información para la DSA se inicia con el trabajo de Brown *et al.* (1990), inspirado en la Teoría de la Información. La importancia de diferentes clases de información en el proceso de DSA se establece para cada palabra de manera individual, de forma empírica, mediante el uso de un algoritmo "*Flip-Flop*" (Nadas *et al.*, 1991, *apud* Brown *et al.*, 1990).

Un paso más hacia el uso selectivo y matizado de la información en la DSA es la propuesta de Mihalcea y Moldovan (2001), en que se aprovecha el método de la *selección activa de atributos* (en inglés, *active feature selection*) procedente de la Inteligencia Artificial. La idea detrás de este enfoque es que diferentes conjuntos de atributos pueden tener diferentes efectos sobre la asignación de sentido, en función de la palabra por desambiguar y, por lo tanto, se deben identificar criterios eficientes para la selección de los atributos previamente a la fase de aprendizaje. Frente al enfoque clásico a la DSA de construir expertos de palabras a través de un proceso de aprendizaje que determina los valores para un conjunto predefinido de atributos, Mihalcea y Moldovan proponen aprender primero el conjunto de atributos que mejor modelizan las características de la palabra, explotando así al máximo la idiosincrasia de las palabras, y, en un segundo paso, determinar los valores para los atributos previamente seleccionados, creando así los expertos de palabras de manera efectiva[120].

4.1.6 Medidas para la evaluación de los sistemas de DSA

En este apartado, introducimos las medidas más usadas para la evaluación de los sistemas de DSA. Detallaremos la problemática de la evaluación en el área de la DSA a lo largo del capítulo 5.

La calidad de un sistema de DSA se suele cuantificar mediante dos medidas de precisión, la precisión absoluta y la precisión relativa, importadas desde el área de la Recuperación de Información (Van Rijsbergen, 1979, Salton y McGill, 1983, *apud* Stevenson, 2003).

La *precisión absoluta* (en inglés, *recall*, R) es la métrica básica para medir la calidad de la tarea de DSA porque muestra el número de casos correctamente desambiguados sobre todos los casos en la prueba.

La *precisión relativa* (en inglés, *precision*, P) se define como el porcentaje de respuestas correctas sobre todos los casos tratados por el sistema de DSA. Esta medida favorece a los sistemas que obtienen una alta fiabilidad en la asignación de sentido.

Otra medida para evaluar los resultados de los sistemas es la *cobertura*, que calcula el porcentaje de ocurrencias que reciben una asignación de sentido sobre el número total de casos por desambiguar. La precisión relativa, cuando tiene el valor máximo de 1, indica que todos los casos resueltos por el sistema están resueltos correctamente; mientras que la precisión absoluta, cuando tiene el valor máximo de 1, indica que el sistema ha tratado todos los casos en la prueba.

Las métricas son realmente informativas si se interpretan conjuntamente. Un sistema puede tener una precisión relativa muy buena, cerca de 1, pero una precisión absoluta muy baja, lo que significa que ha identificado la respuesta correctamente para muy pocos casos, pero con una alta fiabilidad. Inversamente, una precisión relativa baja y una precisión absoluta alta equivale a que el sistema ha tratado buena parte de las ocurrencias por desambiguar, pero que muchas respuestas son incorrectas. En general, hay un equilibrio entre las dos medidas y su combinación se puede expresar mediante la *métrica F* (en inglés, *F-measure*, Van Rijsbergen, 1979), según la fórmula:

[120] Remitimos al apartado 4.2.1.1. para una descripción más completa del enfoque.

$$F = \frac{(b+1)\cdot P \cdot R}{b \cdot P + R},$$

donde b es un parámetro variable. La importancia relativa de las dos medidas, P y R, en la métrica F se puede cambiar mediante la variación de b: si b se fija a 1, la contribución de ambas es igual; si b es igual a 0,5, la precisión relativa cuenta dos veces más que la precisión absoluta; inversamente, si b es igual a 2, la precisión absoluta cuenta dos veces más que la precisión relativa. Usualmente, el valor de b se preestablece a 1 (cf. Stevenson, 2003).

En este trabajo, usaremos las medidas de precisión absoluta, de precisión relativa y de cobertura.

4.2 Métodos basados en fuentes estructuradas de conocimiento léxico

Una de las líneas más difundidas en la labor de DSA es el uso de una caracterización general de los sentidos, como información explícita contenida en recursos estructurados de conocimiento léxico. Así, en los primeros enfoques procedentes de la Inteligencia Artificial, se usaron modelos del conocimiento léxico *ad-hoc*, construidos a propósito para el proceso de resolución léxica de un conjunto limitado de palabras. Más tarde se explotan lexicones y bases de datos léxicas preexistentes de amplia cobertura, construidos fuera del ámbito de la DSA.

4.2.1 DSA basada en sistemas reducidos de la Inteligencia Artificial

A principios de los años sesenta, la DSA empieza a ser estudiada dentro de Inteligencia Artificial, no como problema aislado sino en el contexto de unos sistemas más amplios orientados a conseguir un PLN completo. Después de unos breves años que siguen al informe ALPAC[121], la DSA vuelve a ser objeto de atención de la IA en los años setenta.

La desambiguación semántica sigue las dos orientaciones típicas de la IA, una basada en métodos simbólicos y otra que se basa en métodos subsimbólicos (o conexionistas). Nos respaldamos a esta delimitación en la presentación de los métodos procedentes de la IA que se suelen aplicar a la DSA.

4.2.1.1 Métodos simbólicos

Los primeros pasos en la DSA utilizando la IA basada en conocimiento simbólico se dan con sistemas cuya unidad lingüística superior es la oración, sin tener en cuenta fenómenos de carácter discursivo como información sobre el tópico o el dominio. Para el desarrollo de tales sistemas se empieza a investigar sobre el conocimiento necesario para su funcionamiento y el modo de representarlo. Así aparecen las redes semánticas, como representaciones del significado de las palabras (Masterman, 1962; Quillian, 1961), las plantillas o esquemas (en inglés, *frames*), con información sobre el papel de las palabras y su relación con el resto de palabras de la oración, y las combinaciones entre ambos (Hates, 1976-8; Hirst, 1987). Tres modelos de referencia para el tratamiento de la DSA desde la IA simbólica son la semántica de las preferencias, los expertos de palabras y las palabras polaroid.

Wilks (1972, 1973, 1975) fue uno de los primeros en tratar con profundidad el problema de la DSA. Trata la DSA dentro de la cuestión más general de la construcción de la representación semántica de un texto. Wilks enfoca la DSA a través de su modelo de la *semántica de las preferencias* (en inglés, *Preference Semantics*), que se basa en restricciones de selección relajadas.

Small (1980) junto con Rieger (1982) proponen el procesamiento del lenguaje desde una perspectiva fundamentalmente léxica, consistente en la interacción de *expertos de palabras* (en inglés, *word experts*) para el análisis sintáctico y semántico. Los "expertos" incluyen a la vez restricciones de selección y reglas procedimentales construidas manualmente, dirigidas principalmente hacia oraciones con ambigüedad múltiple.

[121] Publicado en noviembre de 1966 en Es\tados Unidos, el informe de ALPAC (Automatic Language Processing Advisory Committee) asevera que la Traducción Automática es infactible. El impacto del informe fue extremadamente duro, paralizando por veinte años la financiación para la investigación en el campo de la TA.

En la misma línea del uso de expertos de palabras, Hirst (1987) sigue un enfoque más general, con reglas basadas principalmente en restricciones de selección con *back-off*[122] a plantillas más genéricas para mejorar la cobertura. Además, modela la interacción dinámica entre estos expertos en un mecanismo de paso de marcas (en inglés, *marker-passing*)[123] llamado *polaroid words*. El mecanismo elimina progresivamente los sentidos no adecuados, basándose en indicios sintácticos provistos de un analizador y en restricciones semánticas representadas en una red semántica (cf. Rigau, 2002).

El problema de esta aproximación es el coste que implica la construcción de los expertos de palabras cuando el proceso es manual, ya que un experto de palabras construido manualmente (Small y Rieger, 1982) puede tener hasta diez páginas de reglas (cf. Wilks y Stevenson, 1996).

La dificultad para crear manualmente las fuentes de conocimiento necesarias para los expertos de palabras ha restringido el enfoque a implementaciones piloto, con una cobertura lingüística muy reducida. El hecho de que los métodos de DSA de este tipo se prueben sólo sobre conjuntos reducidos de palabras y en contextos muy limitados hace imposible determinar su eficiencia sobre textos reales. En conclusión, la labor de DSA de los años setenta y ochenta basada en expertos de palabras ha tenido poca continuación por su carácter limitado a dominios lingüísticos y conceptuales restringidos.

Sin embargo, estas líneas de investigación sobre la DSA abiertas desde la IA no han sido completamente abandonadas. En los últimos años ha habido un intento de volver a incorporar algunas de estas propuestas con el apoyo de nuevas fuentes léxicas y técnicas computacionales. Así, la idea de los expertos de palabras se encuentra, por ejemplo, en el sistema de aprendizaje automático basado en memoria (en inglés, *memory-based learning*) de Veenstra *et al.* (2000), que participó en Senseval-1. Es una de las alternativas posibles de construir expertos de palabras en la aproximación a la DSA mediante el aprendizaje supervisado[124].

4.2.1.2 Métodos subsimbólicos (conexionistas)[125]

El estudio de la DSA desde la perspectiva subsimbólica dentro de la IA halla su fundamento en la corriente de la psicolingüística llamada *prioridad semántica* (en inglés, *semantic priming*) que se considera como la base de la DSA en los humanos. La *activación* de un determinado concepto influye y facilita el procesamiento de nuevos conceptos con los que mantiene una relación semántica y que son introducidos posteriormente.

Varios experimentos han puesto de relieve la importancia de la activación en la desambiguación, aportando pruebas acerca de su papel en la desambiguación semántica realizada por humanos (Tanenhaus *et al*., 1979, Swiney, 1979, Kawamoto, 1988). Así, en la comprensión de las oraciones, los sentidos no son seleccionados de manera instantánea, sino que se activan todos los diferentes sentidos de las palabras de la oración. Los sentidos que encuentran apoyo contextual permanecen activados, mientras que los inadecuados se desactivan gradualmente, a lo largo del procesamiento de la oración. La activación de un sentido se determina comprobando en que medida los humanos tienen información sobre el mismo.

El concepto de activación es central en toda la labor de tipo conexionista sobre la DSA, y los modelos basados en la activación están a la base de investigaciones sobre la desambiguación semántica y la comprensión del lenguaje.

Una de las clases más relevantes de modelos centrados en el concepto de activación es la *activación por propagación* (en inglés, *spreading activation*). En los modelos de activación por propagación, los conceptos de una red semántica se activan con el uso, y la activación se propaga a los nodos relacionados. La activación disminuye a medida que se propaga, pero determinados nodos pueden recibir activación desde varios focos y, por lo tanto, se pueden ver progresivamente reforzados (Collins y Loftus 1975, Anderson 1976, 1983, *apud* Ide y Véronis, 1998). Posteriormente, se añade al

[122] V. el apartado 4.1.4.1.
[123] Que permite realizar los pasos necesarios para la comprensión de las palabras.
[124] Más detalles sobre el aprendizaje automático, en el apartado 4.1.2.; sobre el sistema de DSA mencionado, en (Veenstra *et al*., 2000).
[125] Una presentación amplia de los enfoques del PLN basados en redes neuronales artificiales (*Artificial Neural Networks*) se encuentra en (Dale *et al*., 2000).

modelo la noción de inhibición entre los nodos: la activación de un nodo puede suprimir, más bien que activar, algunos de sus vecinos (McClelland y Rumelhart 1981, *apud* Ide y Véronis, 1998).

Estos modelos se pueden dividir en dos clases: locales, que se pueden construir *a priori* y cuyos nodos corresponden a un sólo concepto, y distribuidos, que requieren una fase de aprendizaje a base de ejemplos desambiguados, lo que limita su aplicabilidad (Moisl, 2000).

Como ya hemos mencionado, muchas de las propuestas desde el campo de la IA son fecundas en otras líneas de investigación de DSA. Recordamos, en este contexto, la construcción de redes neuronales a partir de los diccionarios, como la de Ide y Véronis en 1990, o la de Kozima y Fugurori en 1993 (Dolan *et al.*, 2000)[126]. En la misma línea, *MindNet* (Dolan *et al.*, 2000) es una base de datos de amplia cobertura con un núcleo derivado de dos diccionarios, pero de carácter abierto: la red es dinámica, se alimenta continuamente, procesando cualquier fragmento de texto que se le provee e incorporando el resultado[127]. Las redes derivadas de diccionarios pueden contener información estadística de la coocurrencia (Ide y Véronis, 1990, y también Kozima y Fugurori, 1993) o bien tener los arcos etiquetados con las relaciones semánticas que las palabras contraen entre sí en los textos de partida, como en el caso de *MindNet*.

4.2.2 DSA basada en fuentes léxicas estructuradas amplias

La labor en la DSA supera las limitaciones de los sistemas de laboratorio y amplía su cobertura cuando se empieza a disponer de fuentes de conocimiento léxico accesibles por el ordenador. A lo largo de esta sección, nos centramos en los sistemas de DSA basados en las fuentes léxicas estructuradas (diccionarios, redes semánticas, etc.). En estas fuentes, los sentidos suelen tener asociada una información explícita, estructurada y variada, a diferencia de los corpus, en los cuales la información relacionada a los sentidos requiere un intenso proceso de búsqueda, explicitación y representación.

4.2.2.1 DSA basada en diccionarios accesibles por ordenador

El uso de los diccionarios accesibles por ordenador (DAO) como fuente estructurada de conocimiento léxico para la DSA se propone por primera vez por Lesk (1986). Los enfoques y los métodos de desambiguación basados en diccionarios son variados, aunque siempre se explota la información asociada a los sentidos:

(a) sentido más frecuente,
(b) definición,
(c) campo temático, o
(d) información múltiple.

Detallamos en este apartado algunos de los métodos que explotan los diccionarios accesibles por ordenador para la DSA.

(a) Sentido más frecuente
La solución más simple es asociar a las palabras polisémicas el primer sentido de su entrada en el diccionario, ya que la práctica lexicográfica suele situar en esta posición el sentido más frecuente o habitual. Es una de las heurísticas de referencia (en inglés, *baselines*) para los sistemas de DSA (cf. capítulo 5). Asignando el primer sentido del diccionario *LDOCE*, Stevenson y Wilks (2001) obtienen una precisión de 30,9%.

(b) Definición
Las definiciones de los diccionarios se suelen usar para la DSA como bolsas de palabras (en inglés, *bag of words*), es decir se obtiene el conjunto de palabras léxicas de la definición como caracterizador de un sentido. Damos a continuación algunos de los enfoques más destacados en la explotación de las definiciones para la desambiguación semántica que utilizan las definiciones como bolsas de palabras.

[126] Cf. el apartado 4.2.1.2.
[127] Más detalles, en el apartado 3.1.1.3.

*Solapamiento entre definición y contexto***.** El método de Lesk (1986) es la base de buena parte de la investigación posterior de la DSA basada en los diccionarios. La idea de Lesk consiste en que las definiciones de una palabra en los diccionarios son probablemente buenos indicadores de los sentidos que definen. Su algoritmo calcula la función de solapamiento (en términos de palabras comunes) entre la definición de cada sentido de la palabra que se quiere desambiguar y las palabras del contexto de la ocurrencia ambigua.

Se trata de un método muy dependiente del inventario de palabras contenido en la definición, ya que la presencia o ausencia de una palabra determinada puede cambiar radicalmente los resultados. A la vez, como hipótesis de trabajo, se parte de una simplificación, porque se ignoran las distinciones de sentido de las palabras que ocurren en el contexto de la palabra por desambiguar. Por sí misma, la información de este tipo derivada de un diccionario es insuficiente para la DSA de alta calidad. Es por ello que sugiere una serie de mejoras: primero, realizar varias iteraciones del algoritmo sobre un texto; y segundo, en vez de usar las definiciones de todas las palabras que ocurren en el contexto de la palabra ambigua, usar sólo las palabras que aparecen en las definiciones de los sentidos adecuados al contexto, determinados en la iteración previa del algoritmo. El objetivo es que el algoritmo ejecutado iterativamente acierte eventualmente el sentido correcto para cada palabra del texto.

El algoritmo de Lesk presenta otros problemas. Por un lado, la explosión combinatoria que supone considerar una por una, para cada una de las palabras del contexto, las palabras que constituyen las definiciones de sus sentidos. Por otro lado, las distinciones finas de sentidos o los sentidos definidos con sinónimos pueden afectar los resultados del método, al igual que las *stop lists*[128] o las formas flexivas, que requieren una ampliación del método (cf. Wilks *et al*, 1996).

La DSA palabra a palabra (en inglés, *one word at a time*) plantea el problema, según señaló ya Lesk, de usar o no en el proceso los resultados previos; es decir, incorporar o no, y cómo, el sentido escogido para una palabra cuando se intenta desambiguar la siguiente. Y, a la inversa, ¿tendría que modificarse la desambiguación previa según los nuevos sentidos que se seleccionan? En otras palabras, ¿admite la DSA un proceso de *feed-back* o sea se puede considerar la desambiguación de la unidades léxicas como un proceso de *feed-back*? (Wilks *et al.*, 1996). Existen dos propuestas para solucionar este problema: operar a gran escala y simultáneamente para varias palabras como en las redes neuronales (Véronis y Ide, 1990) o, desde un enfoque de optimización computacional, activar todos los sentidos a la vez, mediante la técnica del temple simulado (Cowie *et al*., 1992).

Enfoque conexionista[129]**.** Véronis y Ide (1990) proponen una extensión del método de Lesk que consiste en la creación de una red neuronal amplia de activación por propagación (en inglés, *spreading activation*), mediante el uso de las definiciones que los sentidos de las palabras tienen en el diccionario (*Collins English Dictionary*). La red tiene las palabras y los sentidos provistos por el diccionario como nodos, y se construye de la siguiente manera: 1) se establecen enlaces (en inglés, *links*) de una palabra a sus sentidos y de cada sentido a las palabras de su definición; 2) las palabras de las definiciones se enlazan también con sus sentidos; 3) se introducen enlaces de inhibición entre los sentidos de la misma palabra. Para desambiguar una oración, sus palabras se activan en la red, y su activación se permite que se extienda con *feed-back.* Este ciclo se repite un número preestablecido de veces, por ejemplo 100 veces. La desambiguación consiste en elegir, al final de este proceso, para cada palabra su sentido más activado. Según la evaluación de Sutcliffe y Slater (1994, *apud* Rigau, 1998), usando el diccionario *Merrian-Webster*, para un texto de cien oraciones de la novela *La Revuelta de los Animales* de George Orwell, etiquetado morfosintácticamente, el método de Ide y Véronis desambigua correctamente el 68% de las palabras, frente al 31% del método de Lesk.

Temple simulado (en inglés, *simulated annealing*). Se trata de otra extensión, destinada particularmente a superar la explosión combinatoria del método de Lesk. El temple simulado es una técnica de optimización computacional propuesta por Metropolis *et al.* (1953) y Kirkpatrick *et al.* (1983). En DSA, consiste en optimizar simultáneamente el solapamiento entre las definiciones de sentidos correspondientes a todas las palabras de una oración, permitiendo desambiguar todas las

128

palabras a la vez. La estrategia se ha aplicado a la DSA por Cowie *et al.* (1992) sobre *LDOCE*, en un método que explota también los códigos temáticos y, por lo tanto, será presentado a continuación, en el punto d) de esta sección.

Los modelos de solapamiento estricto tienen el problema de la escasez de los datos debido a que las definiciones suelen ser breves. Por lo tanto, sin una ampliación o una generalización basada en clases, estos modelos no logran captar suficientemente bien el rango de la información colocacional necesaria para una amplia cobertura (Yarowsky, 2000b).

Métodos vectoriales[130]. Wilks *et al.* (1990) proponen también una optimización del algoritmo de Lesk, específicamente el enriquecimiento del conocimiento asociado a cada sentido mediante técnicas vectoriales. Sobre *LDOCE*, los autores derivan más información asociada a los sentidos, de la manera que describimos a continuación. En primer lugar, aprovechando que las definiciones están construidas con un léxico limitado de 2781 palabras, construyen un vector para cada palabra usada en el diccionario, juntando las frecuencias de coaparición de las palabras en las definiciones. En segundo lugar, construyen un vector para cada sentido, mediante la suma vectorial de los vectores correspondientes a las palabras que aparecen en su definición. Para la desambiguación de una palabra, se construye el vector del contexto como suma de los vectores de las palabras que contiene, y se calcula la similitud entre los vectores de cada sentido de la palabra y el vector del contexto. En un experimento con una sola palabra (*bank*), el método obtiene una precisión de 45% en la identificación de sentidos y de 90% en la identificación del homónimo.

(c) Códigos temáticos

El uso de otra información, distinta a la definición en los diccionarios, también ha dado resultados interesantes en la DSA. Los códigos temáticos (en inglés, *subject codes*) etiquetan los sentidos con dominios especializados. Walker y Amsler (1986) estiman el código temático como lo más adecuado para palabras como *bank*, con múltiples dominios especializados asociados en un tesauro. La inferencia básica de este enfoque es que las categorías semánticas de las palabras de un contexto determinan la categoría semántica del contexto como un todo –el código temático dominante para las palabras del contexto-, y esta categoría a su vez determina qué sentidos de las palabras individuales se usan. El recurso más utilizado es nuevamente el *LDOCE* (*Longman Dictionary of Contemporary English*)[131], que posee información adicional a la definición, como *box codes* y *subject codes*, para cada sentido. Los *box codes* son primitivas semánticas (ABSTRACTO, ANIMADO, HUMANO, etc.) y codifican restricciones sobre los nombres y los adjetivos, así como sobre los argumentos de los verbos. Los *subject codes* usan un conjunto de primitivas para clasificar sentidos de las palabras según el tema (ECONOMÍA, INGENIERÍA, etc.). Como en el caso de las definiciones o de las clases semánticas de los tesauros, este tipo de información temática se explota con métodos variados, como las funciones de similitud de Wilks *et al.* (1993) o las técnicas de clasificación óptima (Guthrie *et al.*, 1991; Cowie *et al.*, 1992).

*Código temático dominante en el contexto***.** Un método simple de desambiguación se debe a Walker (1987), en el que se asume que si una palabra tiene asignados en el diccionario diferentes códigos temáticos, éstos corresponden a sentidos distintos de la palabra. El algoritmo tiene dos pasos: 1) para una palabra ambigua dada, se cuenta el número de veces que el tesauro enumera cada código temático asignado a sus sentidos como posible tema de las palabras del contexto; 2) se elige el sentido que corresponde al tema con el mejor resultado obtenido previamente. Black (1988) obtiene una calidad de alrededor de un 50% en la aplicación del algoritmo de Walker a una muestra de cinco palabras. Los problemas principales de este algoritmo son dos; el primero es su adecuación a dominios particulares, ya que la categorización de las palabras en términos de temas (en inglés, *topics*) es a menudo demasiado general para campos específicos; el segundo es la cobertura, debido a que el tesauro puede que no contenga buenos indicadores para una categoría determinada.

[130] Para detalles sobre el uso de métodos vectoriales en la DSA, v. Charniak (1993).
[131] Ver apartado 3.1.1.1.

(d) Múltiples tipos de información

La primera y una de las propuestas más conocidas que explotan información variada de los DAO para la resolución léxica, es la de McRoy (1992). Como esta información se combina con colocaciones extraídas de un corpus, la presentaremos el sistema de McRoy en el apartado de métodos mixtos, 4.4.

El diccionario *LDOCE*[132], con su variada información asociada a los sentidos, favorece el uso combinado de estos tipos de conocimiento. Es el caso de dos métodos que se basan en técnicas de *clasificación óptima*, (Guthrie *et al.*, 1991) y (Cowie *et al.*, 1992).

Guthrie *et al.* (1991), además de contar los solapamientos entre definiciones y contextos, establecen una correspondencia entre códigos temáticos mediante un proceso iterativo. De momento, no existe una estimación cuantitativa del método.

Cowie *et al.* (1992) enriquecen este modelo a través de la búsqueda de la clasificación óptima en el caso de ambigüedades múltiples[133], usando la técnica de *temple simulado* (en inglés, *simulated annealing*) con el objetivo de facilitar la búsqueda. Se usa una configuración de sentidos, en la cual cada palabra de la oración tiene asignado un sentido de entre todos los posibles y, a través de este sentido, tiene asociadas todas las palabras de la definición y el código temático del sentido. El objetivo es encontrar aquella configuración de sentido para la cual hay un mayor solapamiento entre la información asociada a todas las palabras de la oración. Es decir el propósito es maximizar el número de palabras que aparecen más de una vez en las definiciones de los potenciales sentidos asignados a las diferentes palabras de la oración. Para la optimización, se define el concepto de "redundancia", R, consistente en el número de palabras y de códigos temáticos que se repiten en una configuración de sentidos, y el concepto de "función de energía", E, que indica las combinaciones buenas entre sentidos (E = 1/(1+R)). Así, la maximización de R equivale a la minimización de E, proceso que sí se puede desarrollar con el algoritmo de temple simulado.

El algoritmo consiste en partir con una configuración de sentidos, sustituir aleatoriamente el sentido asociado a una palabra y comparar la energía de la nueva configuración con la energía de la configuración previa. Si hay descenso de energía, se adopta la nueva configuración de sentidos. El proceso es iterativo. Cowie *et al.* (1992) aplican el algoritmo usando *LDOCE*, eligiendo el primer sentido en *LDOCE* para cada palabra como configuración inicial de sentidos. Para un conjunto de prueba de 50 oraciones no etiquetadas morfosintácticamente, se obtiene un éxito del 47% para el nivel del sentido y del 72% al nivel del homónimo (ambos resultados para todos los tipos de palabras).

Otra propuesta para la explotación de la información contenida en los diccionarios corresponde a los trabajos de Wilks y Stevenson (Wilks y Stevenson, 1997, Stevenson y Wilks, 2000, 2001, Stevenson, 2003)[134]. La fuente léxica que se explota es *LDOCE*[135], en el formato de base de datos desarrollada en el Computing Research Lab of New Mexico State University (Wilks *et al.*, 1988, 1990). Los autores explotan los siguientes tipos de información de *LDOCE*: las categorías morfosintácticas, los códigos gramaticales, las definiciones, los códigos pragmáticos y las preferencias de selección. El sistema está articulado en tres partes principales: componente de preprocesamiento, módulos de desambiguación y módulo de combinación. En el preprocesamiento se incluyen las operaciones de identificación de los nombres propios y de análisis sintáctico superficial[136]. Los módulos de desambiguación son:

1) un filtro (categoría gramatical del diccionario),

[132] Ver el apartado 3.1.1.1.

[133] El enfoque consiste en la desambiguar a la vez todas las palabras de contenido léxico de una oración.

[134] En (Stevenson y Wilks, 2001), se incorpora al sistema también información de corpus, precisamente en el etiquetador parcial, o sea módulo del sistema de DSA, basado en códigos de dominio. Así, el algoritmo usado en este módulo explota unas probabilidades aprendidas sobre la parte escrita del British National Corpus (Burnard, 1995, *apud* Stevenson y Wilks, 2001). Por lo tanto, esta variante del sistema es de tipo mixto.

[135] Ver detalles en el apartado 3.1.1.1.

[136] En el apartado 4.1.5., hemos ofrecido una idea general de la estrategia combinatoria adoptada en las variantes de este sistema para la explotación de información múltiple en la DSA.

2) tres etiquetadores parciales (la definición explotada con la ayuda de un algoritmo de temple simulado[137], el código pragmático, adaptado mediante una variante del algoritmo de Yarowsky (1992), y las restricciones de selección, al estilo de Resnik (1995)) y 3) un extractor de atributos (para la identificación de las colocaciones).

Los resultados de los etiquetadores parciales y del extractor de atributos pasan al módulo de combinación, en que se decide la asignación de sentido. Si inicialmente los módulos se han combinado a mano, posteriormente se han probado diferentes algoritmos de aprendizaje automático, optándose finalmente por el algoritmo Timbl basado en memoria (Daelemans *et al*., 1999, *apud* Stevenson, 2003). Los autores obtienen una precisión del 90% al nivel de sentidos y del 94% al nivel de homónimos, sobre un corpus especialmente construido[138], y un 89,2% sobre el corpus "*interest*"[139].

En conclusión, los métodos basados en DAO presentan los siguientes problemas: inconsistencia de las definiciones, granularidad a menudo inadecuada para la DSA y falta de información pragmática requerida para la determinación del sentido. La DSA basada en las definiciones de los sentidos es insuficiente para una DSA de alta calidad, los resultados obtenidos son demasiado bajos. Sin embargo, la amplia cobertura y la variedad de información contenida en algunos DAO los convierten en una fuente de información valiosa para la DSA. Si la DSA que usa un único tipo de información de los DAO es limitada, la combinación de diferentes clases de información presente en los DAO mejora radicalmente la calidad del proceso.

4.2.2.2 DSA basada en tesauros

La desambiguación basada en tesauros explota la caracterización semántica proporcionada por un tesauro como el *Roget's International Thesaurus* (Roget, 1946) de manera similar a como se usa un diccionario con categorías temáticas como el *Longman's* (*LDOCE*, Procter, 1978). Los sistemas de DSA que explotan los tesauros parten, generalmente, de la hipótesis de que cada ocurrencia de la misma palabra que pertenece a diferentes categorías del tesauro representa un sentido diferente de la palabra. Por lo tanto, las categorías corresponden a los sentidos de las palabras. Así, la asignación de una palabra a una clase equivale a la identificación del sentido.

El primer uso del *Roget's* para la DSA se debe a Masterman, en 1957, dentro de un intento de traducción automática del latín al inglés, en combinación con un diccionario bilingüe. El algoritmo asocia, a cada palabra latina, las categorías superiores (en inglés, *heads*), de sus posibles equivalentes ingleses. A continuación, para las palabras de una oración en latín, el sistema elige, como traducciones correctas al inglés, cada una de las palabras equivalentes cuyas categorías superiores se repiten más entre las palabras de la oración (Ide y Véronis, 1998).

Mencionamos algunas propuestas posteriores para su explotación en la desambiguación. Briena (1973) distingue los homónimos aplicando una métrica basada en las relaciones definidas por sus cadenas. Otro trabajo similar es el de Sedelow y Mooney (1988). Patrick (1985) utiliza el tesauro para discriminar entre sentidos de los verbos examinando los *clusters* semánticos formados por cadenas derivadas del tesauro y los "vecinos próximos" de los niveles bajos de la jerarquía. Walker y Amsler (1986) extraen del tesauro las posibles categorías de la palabra por desambiguar y de las palabras en el contexto oracional. Posteriormente, calculan la frecuencia de aparición de las palabras de la oración en las categorías y así obtienen la categoría más frecuente entre las palabras de la oración. Finalmente seleccionan, para la palabra ambigua, el sentido correspondiente a esta categoría.

Un algoritmo de referencia en el ámbito de la desambiguación basada en un tesauro es el de Yarowsky (1992), que consiste en la adaptación de la clasificación temática (en inglés, *topic classification*) a un corpus, en su caso el texto de aproximadamente 10 millones de palabras de *Grolier's Encyclopedia*. Debido a que la enciclopedia se usa como un corpus y se aplican técnicas de tratamiento de datos, el método es más bien de tipo mixto y será presentado en el apartado 4.4.2.

[137] Ver el apartado 4.2.2.1.
[138] Ver el apartado 3.1.2.2.
[139] Ver el apartado 3.1.2.2.

En síntesis, la DSA basada en tesauros presenta el problema de la caracterización general de las palabras en temas (en inglés, *topics*), que es a menudo inadecuada para un dominio particular. El algoritmo de Yarowsky (1992)[140], que adapta un clasificador temático a un corpus, falla cuando un sentido se propaga a través de varios temas, o sea en el caso de distinciones de sentido independientes del tema. Por otra parte, según Stevenson y Wilks (2000b), los métodos basados en tesauros trabajan mejor para nombres que para verbos, ya que los primeros son predominantes en este tipo de obras.

4.2.2.3 DSA con bases de datos léxicas

En esta sección, presentamos tres enfoques a la DSA que usan las bases de conocimiento léxicas:

(a) usar el número de los sentidos, seleccionando el primer sentido,
(b) explotar la jerarquía de nodos y relaciones léxico-semánticas, a través de medidas de la similitud entre sentidos, y
(c) aprovechar las etiquetas de dominio asociadas a los nodos de la jerarquía.

(a) El número de los sentidos
La modalidad más directa de desambiguar las palabras de un texto es asignarles el primer sentido en la base de datos léxica. *WordNet*, por ejemplo, tiene los sentidos numerados en orden descendiente según la frecuencia de aparición en el corpus. Los resultados que se obtienen aplicando esta técnica difieren ligeramente de un experimento a otro: por ejemplo, Peh y Ng (1997) llegan al 74% de precisión, mientras Stetina *et al*. (1998) obtienen un 75,2% sobre unos ficheros de *SemCor*. Este método se usa para obtener la *baseline* sobre la que se comparan los otros métodos.

(b) Medidas de similitud
Las redes semánticas estructuradas permiten enfoques que miden la relación (en inglés, *relatedness*) entre palabras o entre los sentidos de diferentes palabras usando funciones de similitud. La hipótesis básica de estos métodos es que, para un conjunto dado de nombres coocurrentes en un texto, la elección de los sentidos correctos se traduce en la elección de los sentidos que minimizan la distancia entre ellos en la red semántica. En la mayoría de los estudios presentados aquí se consideran sólo nombres, con la excepción notable de Stetina *et al*. (1998) y de Mihalcea y Moldovan (1999, 2000). La principal ontología, la más conocida y la más utilizada para la DSA es *WordNet*. Pedersen *et al*. (2003) ofrecen una descripción detallada y estructurada de las medidas de similitud definidas sobre *WordNet*. Para una evaluación y comparación de varias medidas, es relevante también al estudio de Budanitsky y Hirst (2001). A continuación, enumeramos solamente algunas de estas medidas[141].

*Extensión del camino***.** Un método sencillo para medir la proximidad semántica entre dos palabras A y B en una red semántica es el de Rada *et al.* (1989), en que se mide la longitud del camino p más corto entre los posibles sentidos de las dos palabras a través de la relación ES-UN:

$$dist(A,B) = \min_{p \in ca\min o(A,B)} longitud(p).$$

Se asume que la proximidad semántica entre las palabras es inversamente proporcional con la longitud de este camino: cuanto más corto el camino entre dos palabras, más próximas semánticamente son las palabras.

*Distancia semántica***.** Sussna (1993) utiliza una noción compleja de la distancia conceptual entre nodos (sentidos de palabras) de la red. La distancia semántica se calcula de la siguiente manera: se asignan pesos a los enlaces (en inglés, *links*) del *WordNet* en base al tipo de relación (sinonimia, homonimia, etc.), y se define una métrica que tiene en cuenta el número de arcos del mismo tipo que parten de un nodo y la profundidad de la sección o límite (en inglés, *edge*) en el árbol general. Esta métrica se aplica luego sobre los arcos del camino más corto que une dos nodos (sentidos de palabras) para

[140] Ver el apartado 4.4.2.
[141] Otras síntesis sobre medidas de similitud son (Dagan, 2000) y (Rodríguez, 2002).

computar la distancia semántica. Los resultados obtenidos son de un 55.8% de precisión para nombres polisémicos y de un 71% para todos los nombres al nivel de sentido. El método de Sussna es particularmente importante porque es uno de los pocos hasta ahora que utiliza no sólo la jerarquía ES-UN sino también otras relaciones léxico-semánticas de *WordNet*.

*Distancia conceptual***.** Para calcular la distancia entre dos conceptos *a* y *b*, Agirre y Rigau (1996) toman en cuenta la profundidad, en la jerarquía de *WordNet*, de los conceptos c_i intermediarios del camino que une a los dos conceptos:

$$dist(a,b) = \min_{p \in ca \min o(a,b)} \sum_{c_i \in p} \frac{1}{profundidad(c_i)}.$$

Rigau *et al.* (1997) usan la distancia conceptual como una medida para la desambiguación semántica del término genérico en las definiciones de DGILE y LPPL[142]. Esta técnica llega al 49% de precisión en el DGILE para nombres polisémicos y al 57% para todos los nombres.

*Densidad conceptual***.** Definida en Agirre y Rigau (1996), la densidad conceptual es una medida semántica más compleja, que toma en cuenta más factores en el cálculo de la proximidad semántica entre los conceptos:

1) la longitud del camino más corto que une los conceptos;
2) la profundidad en la jerarquía, considerando que los conceptos profundos son más cercanos que los de las partes altas;
3) la densidad de conceptos en la jerarquía, de manera que en las áreas con muchos más conceptos, éstos deben considerarse más próximos que los hallados en áreas con menos conceptos;
4) la necesaria independencia del número de conceptos que intervienen en su cómputo.

La densidad conceptual permite desambiguar una palabra a través de la identificación de su hiperónimo cuyo subárbol tiene la densidad máxima de sentidos de las palabras del contexto. La densidad de un concepto *c* cuando su subjerarquía contiene *m* sentidos de las palabras a desambiguar se calcula según la fórmula siguiente:

$$CD(c,m) = \frac{\sum_{i=0,m-1} nhyp^{0,20}}{\sum_{i=0,h-1} nhyp^{i}},$$

donde *nhyp* es el número de hipónimos por nodo y *h* es la altura de la subjeraquía. De manera intuitiva, la densidad conceptual expresa la proporción entre el área de la subjerarquía que contendría los *m* sentidos de las palabras ambiguas y el área de la subjerarquía de *c* (ésta última equivalente al número de sentidos descendientes de *c*).

Aplicada a la DSA sobre el *Brown Corpus* y evaluada sobre algunos documentos de *SemCor*, la medida lleva a una precisión de 43% (Agirre y Rigau, 1996). En este experimento, se utiliza un parámetro que controla en qué medida se afecta a los resultados si se añaden las relaciones de meronimia a las relaciones de hipo/hiperonimia. Según los autores, en el caso de añadir la información de meronimia, la precisión del método no mejora, y la cobertura aumenta sólo un 3%. La contribución de las relaciones de meronimia a la DSA sería mínima.

Tomando como punto de partida el experimento anterior de Agirre y Rigau, Fernández-Amorós *et al.* (2001) exploran varias maneras de usar las medidas de densidad conceptual. La serie de experimentos que ellos desarrollan tiene el objetivo de estudiar en profundidad el papel que juegan las diferentes relaciones léxico-semánticas en la DSA, es decir su poder discriminatorio individual y la influencia de su interacción sobre el proceso de DSA. Así, se considera como transitiva la

[142] *Le Plus Petit Larousse*, G. Gougenheim (ed.), Librarie Larousse, 1980.

combinación entre la unión de relaciones semánticas como hiperonimia y meronimia. Por otra parte, se estudia el papel de otro factor que puede potenciar la contribución de dichas relaciones a la DSA: la dimensión del contexto necesario alrededor de la palabra a desambiguar. Respecto de este parámetro, se ha variado su valor hasta 500, obteniéndose los mejores resultados con una ventana de más de 150 palabras. Realizando una serie de 50 variantes experimentales, los autores proponen una evaluación exhaustiva de distintos algoritmos de DSA que se basan únicamente en relaciones conceptuales entre sentidos candidatos. La precisión más alta, obtenida en todos estos experimentos para las diferentes relaciones léxico-semánticas o sus combinaciones, es la siguiente: hiperonimia sola, 31,28%; meronimia sola, 26,62%; combinación de hiperonimia y meronimia, 31,28%; holonimia sola, 27%; combinación de hiperonimia y holonimia, 30,88%. La serie de experimentos llevan a la misma conclusión que la de Agirre y Rigau, la de que la meronimia y la holonimia no añaden ninguna información útil a la hipo/hiperonimia. Por otra parte, los autores afirman la necesidad de combinar las relaciones conceptuales con otros tipos de información: contextual, sintáctica, dominio, información, etc. para mejorar el nivel de precisión en la desambiguación.

*Marca de especificidad***.** Otra noción de proximidad semántica, derivada de la densidad conceptual, es la marca de especificidad, definida por Montoyo y Palomar (2000a, 2000b). La idea que hay detrás de esta propuesta es que cuanta más información comparten dos conceptos, más relacionados estarán. La información compartida corresponde al concepto que subsume a los dos conceptos de partida en la jerarquía ES-UN de *WordNet*. A esta noción común se llamará *marca de especificidad* (ME). Ponemos de relieve que si bien el enfoque es muy próximo al de Resnik (1995), en este caso se estudia la información compartida entre dos conceptos y no entre dos palabras. Además, la marca de especificidad es independiente del uso de un corpus. La noción de ME se ha definido para nombres. El método que la explota parte de la asunción de que las palabras que aparecen en un contexto están fuertemente relacionadas. En consecuencia, el método toma como entrada el conjunto de nombres en la oración de la palabra por desambiguar. Se recorre luego toda la jerarquía de *WordNet*, en búsqueda de aquella marca de especificidad en cuyo subárbol haya la mayor densidad de sentidos de las diferentes palabras de entrada. A las palabras de entrada se les asignan los sentidos que tienen en este subárbol. Evaluado sobre *SemCor* para dieciséis nombres, el algoritmo obtiene una precisión relativa de un 40% y una cobertura de 97,8%, con lo cual la precisión absoluta es de 39,5% (Suárez y Montoyo, 2001).

*Contenido de información compartido por las palabras***.** Los métodos que calculan la extensión del camino entre dos conceptos (por ejemplo, Sussna, 1993) encuentran un problema en la densidad variable de la taxonomía de *WordNet*. Como alternativa, Resnik (1995) propone el contenido de información compartido por las palabras para medir su similitud. En el cálculo de esta similitud semántica entre palabras intervienen probabilidades obtenidas a partir de un corpus, por lo tanto su propuesta es un método mixto que presentaremos en el apartado 4.4.

Lin (1997) utiliza una medida próxima de similitud entre dos conceptos igual a la ratio entre la cantidad de información necesaria para describir el concepto superior común a los conceptos y la cantidad de información necesaria para describir cada concepto en parte. El sistema de DSA que utiliza esta medida se presentará entre los métodos mixtos (apartado 4.4.).

*Similitud diferenciada por categoría sintáctica***.** Stetina *et al.* (1998), calcula la similitud en base a una distancia semántica definida de manera diferenciada para las distintas clases de palabras: nombres y verbos, adjetivos, adverbios respectivamente. Así, en el caso de los nombres y de los verbos, se tiene en cuenta la profundidad en la taxonomía de *WordNet*, mientras que para los adjetivos y los adverbios, se tiene en cuenta el tipo de relación léxico-semántica entre los *synsets* de los sentidos. Ofrecemos una descripción del sistema igualmente entre los métodos mixtos (apartado 4.4.).

*Densidad conceptual intercategorial***.** Mihalcea y Moldovan (1999a), proponen otra medida de similitud semántica, esta vez entre sentidos de nombres y de verbos. Esta medida utiliza las glosas de WorNet que definen los sentidos de los verbos. Llamada igualmente “densidad conceptual”, la medida se define como el número de nombres que comparten las glosas en *WordNet* un sentido del nombre y

un sentido del verbo. Presentamos el método que explota esta medida en el apartado de los métodos mixtos (4.4).

Similitud para una función sintáctica dada. Li *et al.* (1995) combinan la similitud con información sintáctica, proponiendo un método para la desambiguación de los nombres en función de objeto. Para un nombre dado, se toma como contexto el verbo cuyo argumento es como objeto. Se consideran cuatro niveles de similitud en *WordNet*, entre un concepto dado y diferentes conceptos relacionados: sinónimos del mismo *synset*, hiperónimos directos, hipónimos inmediatos, coordinados. Se explota la similitud tanto para los nombres como para los verbos. El sistema se compone de las siguientes heurísticas que se aplican a un par formado por un verbo y su objeto, para la desambiguación del objeto:

1) se busca en el contexto otro objeto para el mismo verbo, que sea similar con el objeto de partida;
2) se busca en el contexto otro verbo con el mismo objeto, con el cual el objeto tenga el sentido ya determinado; este verbo tiene que ser similar con el de partida;
3) se busca en el contexto otro par verbo-objeto que tenga el verbo y el objeto similares con los del par de partida;
4) se busca en el contexto, inmediatamente después del par verbo-objeto, el relator "*such as*" seguido por un nombre, que sea además similar con el objeto de partida;
5) se busca en el contexto una estructura coordinativa del verbo de partida con otro verbo[143], similar a éste y además con el mismo objeto.

El sentido propuesto en cada heurística será el que se encuentra a través de la similitud. Cada una de las heurísticas tiene asignado un coeficiente de fiabilidad diferente y las heurísticas se aplican en el orden de su fiabilidad. Se proponen cinco posibles respuestas: una sola solución correcta, varias soluciones correctas, soluciones parcialmente correctas, soluciones erróneas, ninguna solución. La evaluación estima la precisión en cada caso; para una sola respuesta correcta, se obtiene un 57%, mientras que si se consideran varias respuestas, se alcanza el 72%.

*Generalización a cualquier similitud***.** Pedersen *et al.* (2003) introducen una estrategia nueva en el uso de las medidas de similitud en el proceso de DSA. Su método de DSA es una generalización de la idea de Lesk (1986). Éste consiste en seleccionar el sentido de la palabra por desambiguar que tiene la relación máxima con las palabras de contenido léxico de la ventana contextual. El algoritmo se puede usar con cualquier medida que calcula una relación entre dos conceptos, lo que permite identificar la medida más idónea.

(c) Etiquetas de dominio

Otro tipo de información asociada a *WordNet* y que se ha utilizado para la DSA son las etiquetas de dominio asignadas a los *synsets* (Magnini y Cavaglià, 2000, cf. apartado 3.1.1.3). En este caso, para cada palabra a desambiguar, se elige una etiqueta de dominio en vez de un sentido. Se obtiene así una variante de la tarea de DSA, Desambiguación del Dominio de la Palabra (DDP, en inglés, *Word Domain Disambiguation*, WDD). El uso de las etiquetas de dominio tiene como consecuencia la reducción de la polisemia de las palabras, excesiva en *WordNet*, debido a que generalmente el número de dominios asociados a una palabra es más bajo que el número de sentidos que ésta pueda tener.

Magnini y Strapparava (2000) aplican las etiquetas de dominio a la DSA a través de dos algoritmos, correspondientes a dos modalidades de medir la frecuencia de un dominio. El primer algoritmo mide la frecuencia de los dominios en un texto, sumando las veces que cada dominio está asignado a los sentidos de las palabras en el texto, y elige el dominio más frecuente como resultado de la desambiguación. El segundo algoritmo mide la frecuencia de un dominio respecto de una palabra, como el cociente de los sentidos pertenecientes al dominio sobre el total de sus sentidos. Los resultados de los dos algoritmos fueron de 85% y 86% respectivamente en términos de precisión, lo que indica que la DDP es una alternativa fiable para la DSA.

[143] Se consideran sólo las estructuras [verbo *and* verbo] y [verbo *or* verbo].

La idea básica del trabajo de Magnini y Strapparava (2000) es que la desambiguación de una palabra consiste principalmente en un proceso de comparación entre el dominio del contexto y los dominios de los sentidos de la palabra. Un inconveniente de este enfoque es que no se toman en cuenta las variaciones del dominio en un texto más amplio. Magnini *et al.* (2002) proponen soluciones para superar estas limitaciones. En esta nueva propuesta, se consideran fragmentos de textos, dentro de los cuales se calcula el dominio relevante, y además se construye un marco para integrar la información de dominio adquirida de textos anotados. La estructura de datos que recoge la información de dominio es el vector de dominio, siendo su longitud igual al número de dominios considerados.

Se consideran dos tipos de vectores: 1) vectores de texto, que representan la relevancia de una porción de texto con respecto a cada uno del conjunto de dominios considerados, y 2) vectores de sentido, que representan la relevancia de un sentido para una palabra dada con respecto de cada dominio considerado. Los primeros se calculan a base de la anotación del texto con sentidos, mientras que los segundos se inducen a partir de ejemplos de entrenamiento o de información extraída de *WordNet Domains* u otras fuentes etiquetadas. Para la desambiguación de una ocurrencia de una palabra en una porción de texto, se calculan el vector de la porción de texto y los vectores para cada uno de los sentidos. El sistema elige el sentido cuyo vector maximiza la similitud con el vector del texto.

En Senseval-2 (capítulo 5), el sistema de Magnini obtuvo una precisión relativa muy buena (75% para la tarea *all-words*, respectivamente 66% en la tarea *lexical sample* para el inglés), pero una cobertura muy baja. La diferencia de precisión está relacionada con la dimensión del contexto disponible en las dos tareas. El sistema obtiene resultados mejores con una ventana contextual de 100 palabras, lo que se pudo usar en la tarea *all-words*, pero menos en la tarea *lexical sample*. Mencionamos algunas conclusiones del experimento. En la tarea *lexical sample*, la precisión desciende con el aumento de la polisemia. Sólo unas pocas palabras llevan efectivamente información de dominio relevante; sin embargo, la mayoría de las palabras se comportan como FACTOTUM y para éstas es necesaria la información local. Además, la comparación con otros sistemas ha puesto de manifiesto que la información de dominio es útil en la desambiguación de algunas palabras cuya ambigüedad no se ha resuelto por otros sistemas.

Vázquez *et al.* (2003) desarrollan a partir de *WordNet Domains* un nuevo recurso, denominado "Dominios Relevantes" (cf. apartado 3.1.1.3). El recurso consiste en la lista ordenada de las palabras de *WordNet*, cada una de las palabras teniendo asignado el conjunto de dominios más relevantes a ella. Esta organización se basa en la ratio de asociación (en inglés, *association ratio*)[144]. Para la desambiguación de una palabra ambigua en contexto dado, se construyen dos vectores: un vector de contexto y un vector de sentido, según se detalla a continuación. Para la obtención del vector de contexto, se suman los valores del ratio de asociación de cada una de las palabras de contenido léxico del texto (nombres, verbos, adjetivos y adverbios). La forma final del vector corresponde a una lista de los dominios representativos del texto con los valores correspondientes del ratio de asociación, ordenados de manera descendiente según estos valores. Los vectores de sentido se obtienen de manera similar, pero el cálculo se realiza sobre las palabras de clase abierta de la glosa correspondiente. Se construye un vector de sentido para cada uno de los sentidos de la palabra a desambiguar. El proceso de desambiguación consiste en identificar el vector de sentido más próximo al vector de contexto, mediante la medida del coseno entre vectores. Evaluado sobre la tarea *all-words* para el inglés de Senseval-2, el método ha obtenido los mejores resultados con un contexto de 100 palabras y 43 dominios más generales (los del segundo nivel en la jerarquía de *WordNet Domains*): 54% precisión relativa y 43% precisión absoluta.

A título de conclusión, resaltamos que los límites de las redes semánticas derivan, entre otros motivos, de la dependencia de la jerarquía de las personas que la construyen. Es decir, una distinta estructuración lleva a distintos resultados. Por otra parte, usando redes semánticas se logra capturar una sola noción de proximidad semántica, limitada a las relaciones léxico-semánticas representadas en

[144] La fórmula de la ratio de asociación entre dos palabras es la siguiente:

$$AssociationRatio(X,Y) = \frac{P(x)\cdot P(y)}{P(X,Y)} \cdot \log P([X - R - Y])$$

ellas actualmente. Para capturar la noción de manera más adecuada, se deberían codificar otros tipos de asociaciones entre los sentidos de las palabras.

4.2.2.4 Métodos basados en diferentes fuentes léxicas estructuradas

Ilustramos este tipo de métodos con el sistema de Rigau *et al.* (1997). Para la desambiguación del término genérico en las definiciones nominales del DGILE (1987)[145], se combinan ocho heurísticas basadas en lexicones derivados del mismo diccionario[146]. Los criterios en los que se basa cada una de estas heurísticas son los siguientes:

1) el término genérico monosémico;
2) el orden de los sentidos en la entrada para los hiperónimos;
3) el dominio semántico del hipónimo y de hiperónimo;
4) el solapamiento de palabras en las definiciones;
5) la máxima ratio de asociación entre las definiciones de los posibles sentidos del hiperónimo y la definición del hipónimo;
6) máxima similitud entre los vectores de las definiciones de los hiperónimos y el vector de la definición del hipónimo;
7) máxima similitud entre los vectores de los hiperónimos y el vector semántico del hipónimo, donde los vectores semánticos se han asignado previamente en términos de pesos para las etiquetas de los 24 ficheros semánticos de *WordNet*;
8) distancia conceptual mínima entre las definiciones de los hiperónimos y la definición del hipónimo.

Los votos de una heurística para cada sentido son normalizados[147], dividiéndolos por el peso máximo asignado a un sentido. La combinación de las heurísticas se hace sumando los votos que cada una de las heurísticas da a los sentidos. La desambiguación obtenida permitió la construcción de la taxonomía del diccionario, con 111.264 nodos correspondientes a sentidos. La cobertura es del 100% y la precisión del 83%, por lo tanto el experimento es uno de los más amplios y precisos en DSA.

4.3 Métodos de DSA basados en corpus

Una alternativa al uso de fuentes léxicas estructuradas, que aportan conocimiento explícito sobre los sentidos de las palabras, consiste en derivar este conocimiento directamente de los textos. En este caso, el conocimiento sobre los sentidos se induce a partir de ejemplos en corpus, generalmente con la ayuda de un algoritmo de aprendizaje automático. Los algoritmos se entrenan sobre las ocurrencias de las palabras, de manera que pueden tratar casos nuevos en base al conocimiento adquirido.

La información se puede obtener de corpus etiquetados, es decir previamente desambiguados semánticamente, o bien de corpus no etiquetados. El etiquetado no consiste necesariamente en etiquetas de sentido explícitas, las etiquetas pueden ser categorías semánticas (HUMANO, ANIMAL, etc.), categorías de tesauro, equivalencias de traducción a otra lengua, etc.

Algunos métodos se basan en ejemplos no etiquetados, pero utilizan fuentes de conocimiento, como diccionarios en formato electrónico u bases de datos léxicas. Estos métodos, junto con los que se basan en *bootstrapping*[148], se consideran, en Yarowsky (2000b), un subconjunto diferenciado de los no supervisados, y reciben el nombre de *mínimamente supervisados* o *semi-supervisados*. Debido a que

texto determinados atributos y, con la ayuda de los atributos, se elabora una representación de cada

ignorancia sobre el proceso que se debe modelizar; no se considera ningún otro dato fuera de los datos de entrenamiento, representados por un corpus etiquetado con sentidos y con categorías morfosintácticas.

El aprendizaje para la construcción de un clasificador de sentido consiste en dos fases: 1) la construcción de una clase de modelos para representar los datos de entrenamiento, con las características adecuadas, y 2) la obtención del modelo de probabilidad óptimo que maximiza la entropía. Para representar la evidencia, los modelos MXE utilizan funciones binarias que indican la presencia o no de cierta característica dentro de cada ejemplo de aprendizaje. Cada contexto será representado por un vector de *n* características. Las funciones binarias informan sobre las palabras y combinaciones de palabras o sobre las categorías sintácticas que se hallan en la cercanía de la palabra objetivo, así como de su posición relativa respecto a la palabra objetiva. Por lo tanto, en la primera fase, se procesa el corpus etiquetado para definir las funciones. Luego se obtiene el valor de las funciones para todos los ejemplos de entrenamiento. Se llenan los vectores de características con los valores resultados de la evaluación de cada función. En la segunda fase, se estiman los parámetros que nos indican el peso o la importancia de cada función en el proceso de clasificación. La estimación se realiza mediante el procedimiento denominado *Generalized Iterative Scaling*. Al final, el modelo obtenido tiene la misma distribución de probabilidad que la observada en el conjunto de entrenamiento. Esta distribución óptima de probabilidad será, de todas las distribuciones posibles, aquella que maximice la entropía partiendo de la distribución observada. Tras la estimación de parámetros se obtiene una función que permite asociar el sentido correcto a un contexto lingüístico mediante la elección del valor de probabilidad máximo. El clasificador así obtenido dirá cual es el sentido de la palabra objetivo para cualquier contexto lingüístico suyo (Suárez y Palomar, 2002; Suárez y Montoyo, 2001).

La Máxima Entropía permite representar fuentes de información de contexto heterogéneas, obtener buenos resultados incluso con parámetros pobres y un tratamiento diferenciado de los problemas por resolver, mediante parámetros específicos (Ratnaparkhi, 1998, *apud* Suárez y Palomar, 2002).

Métodos basados en modelos ocultos de Markov (en inglés, *Hidden Markov Models*, *HMM*). Este modelo proviene de la estadística y previamente ha sido aplicado con éxito al etiquetado morfológico (en inglés, *POS-tagging*). Un modelo oculto de Markov se puede ver como un autómata finito, donde la función de transición es probabilística, es decir, los arcos entre los estados están etiquetados con probabilidades y los estados, representando las variables del modelo, están asociados con una función de probabilidad. Los modelos se llaman ocultos porque no se sabe cuál es la secuencia de estados por los que transita el autómata, sino sólo una función probabilística de la misma. La aplicación de un modelo oculto de Markov a un ejemplo por etiquetar (que representaría una secuencia de símbolos observados) consiste en encontrar la secuencia de estados (etiquetas) que maximiza la probabilidad de haber producido el ejemplo (cf. Màrquez, 2002). El algoritmo fue aplicado a la DSA, por ejemplo, por Segond *et al.* (2000).

4.3.1.2 Métodos basados en reglas

Las reglas son implicaciones en forma de lógica proposicional o de predicados que se usan para diseñar inferencias deductivas a partir de los datos. Hay una variedad de algoritmos para la inducción del conocimiento a partir de los ejemplos de entrenamiento.

*Métodos basados en árboles de decisión***.** Este tipo de métodos proviene de la IA. Los árboles de decisión son modelos simbólicos, es decir, funciones de clasificación representadas como árboles en que los nodos son tests de atributos, las ramas son valores de atributos, y las hojas son etiquetas de clases. Un árbol de decisión consiste en una secuencia jerárquica de preguntas, normalmente binarias, que parte progresivamente del conjunto de ejemplos observados. Para la clasificación de un ejemplo, se parte desde la raíz y se atraviesa recursivamente el árbol por un camino unívoco hasta alcanzar una hoja siguiendo el camino dictado por los valores que tienen los atributos para el ejemplo. El camino constituye una regla conjuntiva de clasificación. Cada nodo en el camino contiene una pregunta sobre un determinado atributo y el arco hacia el nodo siguiente contiene un valor posible para este atributo. La hoja final a la cual lleva el camino es una clase, o sentido, como resultado más probable para ese

ejemplo (cf. Màrquez, 2002; Mooney, 2003). Este método tiene dificultades para tratar la fragmentación de datos típica de la DSA, debido a la gran cantidad de datos que se tratan.

El punto de partida en el uso de estos algoritmos para la DSA se halla en el sistema de Brown *et al.* (1991) que usa el algoritmo Flip-Flop. En este algoritmo se adopta una estrategia contraria a la de la clasificación bayesiana, es decir, se intenta hallar un solo rasgo contextual que indique de manera fiable qué sentido de la palabra ambigua se usa. Además, para hacer un buen uso de un rasgo conceptual, sus valores tienen que estar categorizados en relación con el sentido que indican. Siguiendo a Manning y Schütze (1999), presentamos una variante simplificada, consistente en la desambiguación para dos únicos sentidos. En este sistema, se parte de un rasgo predeterminado, del conjunto de sus valores y del conjunto de las ocurrencias de la palabra ambigua. El algoritmo intenta iterativamente encontrar la partición adecuada dentro de cada uno de estos dos conjuntos, de tal manera que la información mutua[152] entre las dos particiones sea máxima. En cada paso se fija primero por separado una de las dos particiones y se busca una variante de la otra que maximice la información mutua. Posteriormente, el proceso se repite fijando esta vez la otra partición. Se ha demostrado que, con cada paso, la información mutua entre las dos particiones encontradas provisoriamente aumenta monótonicamente, con lo cual, un criterio natural para detener el proceso es el cese del incremento o bien el incremento insignificante. La división de los valores del indicador dará la mayor información para distinguir entre los sentidos de la palabra ambigua. Posteriormente se aplica el algoritmo para todos los posibles indicadores y finalmente se elige al indicador con la más alta información mutua.

Al final de la aplicación del algoritmo, hay una partición del conjunto de valores del indicador en correspondencia con la lista de sentidos posibles de la palabra ambigua; con lo cual, un valor del indicador remite a un sentido determinado. Las dos fases del proceso de DSA son por tanto:

> 1) el aprendizaje del rasgo y de la partición de sus valores más relevantes para la DSA desde la perspectiva de la información (es decir con la información que más aportación pueda tener a la desambiguación);
>
> 2) la desambiguación propiamente dicha, que supone: 2a) para cada ocurrencia de la palabra ambigua, se determina el valor del indicador, y 2b) el subconjunto dentro de la partición de los valores del indicador al cual pertenece el valor previamente determinado lleva al sentido adecuado.

Brown *et al.* (1991) usan el algoritmo *Flip-Flop* en la búsqueda de indicadores para la desambiguación. Cada ocurrencia de la palabra francesa *cent* está "etiquetada" no con su sentido sino con sus traducciones inglesas correspondientes; estas etiquetas de clase no son sentidos, por lo tanto el método es no supervisado en el sentido tradicional de la IA. Los autores afirman que se produce una mejora del 20% en el sistema de traducción automática cuando se incorpora el algoritmo.

*Métodos basados en listas de decisión***.** Estos métodos también provienen de la IA (Rivest, 1987). En el campo del aprendizaje de conceptos (en inglés, *concept learning*), los árboles se pueden traducir a reglas de clasificación para representar el concepto objetivo. El formalismo consiste en una lista ordenada de reglas conjuntivas, cada una en forma de tupla (condición, valor). Para la clasificación de un nuevo ejemplo, las reglas, ordenadas de más segura a menos segura, se evalúan de manera ordenada sobre el ejemplo, de forma que se aplica la primera regla cuya condición se cumple, ya que es la más segura. Las condiciones excepcionales suelen aparecer al principio de la lista, mientras que las condiciones generales, al final. Cuando se encuentra la condición que satisface la pregunta, se selecciona el valor correspondiente a esta condición como respuesta. El modelo se fundamenta en la hipótesis de que es posible discriminar los ejemplos de un dominio a base de reglas individuales sencillas. Las listas de decisión evitan en cierta medida el problema de la excesiva fragmentación de los datos. Por lo tanto, el método es altamente eficiente para conjuntos con un gran número de atributos no independientes y con baja entropía, como es el caso de la DSA (Yarowsky, 2000a). Sin embargo, la construcción de listas de decisión se ve perjudicada por problemas de escasez de datos.

Las listas de decisión fueron aplicadas a la DSA por Yarowsky (1994, 1995), Mooney (1996), Wilks y Stevenson (1998). En (Yarowsky, 1994) e (Yarowsky, 1995), las condiciones de las listas de decisión son colocaciones de la palabra por desambiguar y los valores correspondientes, los sentidos

[152] La información mutua entre dos entidades (como dos palabras) representa la cantidad de información que cada una de ellas ofrece sobre la otra. Para una definición formal, véase, por ejemplo, (Charniak, 1993:137).

correctos de la palabra y su probabilidad en las colocaciones. En la lista de decisión, las reglas se colocan en orden descendiente de las probabilidades. Yarowsky (1994) aplica el método para la restauración del acento en español y francés. En (Yarowsky, 1994, 1997), se usan listas de decisión interpoladas, lo que evita la fragmentación de los datos de entrenamiento observados en los árboles de decisión tradicionales, no interpoladas, de Rivest (1987). El mismo autor (Yarowsky, 2000a) explota además las listas de decisión jerárquicas, añadiendo, a las listas de decisión planas, un grado de ramificación condicional. De esta manera, se puede orientar el recorrido del procedimiento de decisión a lo largo de caminos relativamente independientes, especializados para las necesidades de la modelización de cada parte de la partición. Sin embargo, se mantiene la ventaja de la no fragmentación de los datos que presentan las listas de decisión interpoladas. Las claves contextuales que dirigen el algoritmo con listas de decisión son un conjunto rico que combina dos clases de atributos: a) tipos de *tokens* (palabra literal, lema, categoría morfosintáctica, clase de palabra, etc.) y b) las posiciones relativas respecto de la palabra focalizada (posiciones fijas, la palabra misma, coocurrencia dentro de una ventana de dimensión variable, n-gramas, relaciones sintácticas).

El sistema JHU, de la John Hopkins University, ha obtenido la mejor precisión (78, 4%) sobre las palabras con datos de entrenamiento en Senseval-1 para el inglés. Además, la comparación con las listas de decisión planas ha demostrado que la partición del flujo de datos para ciertos atributos (inflexiones de palabras clave, características sintácticas básicas, colocaciones idiomáticas y subsentidos) mejora los resultados, manteniendo las ventajas de las listas de decisión en el flujo de información.

En un experimento de carácter comparativo, Agirre y Martínez (2000) analizan el funcionamiento de las listas de decisión sobre tres corpus, dos disponibles (*SemCor* y *DSO*) y uno adquirido a partir de la web de manera automática.

4.3.1.3 Métodos basados en memoria

Estos métodos son de tipo simbólico, procedentes de la IA. El *aprendizaje basado en memoria* (en inglés, *memory based learning*) se llama igualmente *aprendizaje basado en ejemplos*, *en casos* o *en los vecinos más próximos*. En este caso, no se construye la definición de una función abstracta, sino se categorizan los nuevos ejemplos en base a su similitud con uno o más ejemplos de entrenamiento. El aprendizaje consiste en memorizar los ejemplos sobre los que se aprende, es decir, en almacenarlos sin procesarlos. Para clasificar un nuevo ejemplo, se procura obtener de la base de ejemplos el conjunto de los *k* ejemplos más parecidos al ejemplo que se quiere clasificar, los *k* vecinos más próximos (en inglés, *k-NearestNeighbours, kNN*), y asignar la clase mayoritaria entre ellos. Esta fase de clasificación implica la comparación del ejemplo que se quiere clasificar con cada uno de los ejemplos guardados y el cálculo de la distancia entre ellos, en conformidad con alguna métrica. La distancia más simple para tratar con atributos simbólicos es la distancia de Hamming, llamada también métrica de solapamiento. Ésta consiste en calcular la distancia entre dos ejemplos como la suma de las distancias entre los valores de los ejemplos obtenidos para cada atributo (simplificada a 1 en caso de igualdad y a 0 en caso de diferencia entre los valores), estando cada distancia asociada con un determinado peso (cf. Màrquez, 2002; Mooney, 2003).

La propiedad distintiva de este enfoque, como método de aprendizaje supervisado basado en clasificación, es que no abstrae a partir de los datos de entrenamiento, tal como hacen otros métodos de aprendizaje, denominados *"ansioso"* (en inglés, *eager*), como los basados en árboles de decisión, inducción de reglas o redes neuronales. El aprendizaje basado en memoria guarda en la memoria todos los datos de entrenamiento y sólo abstrae en el momento de la clasificación, extrapolando la clase de los casos más similares en la memoria. Por eso, a estos se les asocia el calificativo de *aprendizaje "perezoso"* (en inglés, *lazy*). A diferencia de los métodos "ansiosos", que "olvidan" información porque acortan y abstraen en base de la frecuencia, los métodos basados en la memoria "recuerdan" los casos excepcionales, de baja frecuencia. Además, la asignación automática de pesos en la matriz de similitud usada en el aprendizaje basado en memoria hace que el enfoque sea adecuado para dominios con un amplio número de atributos provenientes de fuentes heterogéneas, porque incorpora un método de suavizado[153] por similitud cuando los datos son escasos (Veenstra *et al.*, 2000).

[153] Para métodos de suavizado, ver el apartado 4.1.4.1.

El método ha sido aplicado a la DSA por Dagan *et al.* (1994), Ng y Lee (1996) y Zaorel y Daelemans (1997). Dagan *et al.* (1994) destacan los efectos benéficos de los enfoques *kNN* para los datos escasos (cf. Manning y Schütze, 1999). Daelemans (1999) confirma además la superioridad de los métodos de aprendizaje basados en memorización de ejemplos ya que, al omitir la generalización a partir de los ejemplos encontrados, no prescinden de los casos que representan excepciones y que pueden ser numerosos.

Una experimentación relevante para este enfoque es (Ng y Lee, 1996), usando el sistema supervisado de DSA llamado LEXAS (de LEXical Ambiguity-resolving System). El conjunto de atributos contiene: la categoría morfosintáctica, la forma morfológica, las palabras vecinas, las colocaciones locales y las palabras en relación sintáctica verbo-objeto. La distancia entre dos ejemplos se define como la suma de las distancias entre los valores característicos asociados a los ejemplos. O sea, entre la distribución estadística obtenida a partir del corpus de entrenamiento para los valores con cada uno de los sentidos de la palabra. La idea detrás de esta noción de similitud es que dos ejemplos son similares si tienen valores característicos con una distribución parecida en los datos de entrenamiento. El sistema permitió a Ng y Lee obtener un buen nivel de desambiguación. Así, sobre el corpus "*interest*"[154], se alcanza una precisión media de 87,4%, superior al 78% de Bruce y Wiebe (1994). Sobre fragmentos de *Brown Corpus* y de *Wall Street Journal*, la precisión absoluta obtenida (54%, respecto a 68,6%) está por encima de la selección del primer sentido o del sentido más frecuente. La investigación de Ng y Lee pone de manifiesto que los métodos basados en similitud son altamente dependientes de la selección de los atributos.

4.3.1.4 Métodos basados en corpus bilingües

Estos métodos se fundamentan en la idea de que la traducción distinta de una palabra a otra lengua corresponde a sentidos diferentes de la palabra. Hemos comentado en el apartado 3.1.2.2.3. las limitaciones de este enfoque. Un método de referencia en la explotación de los corpus bilingües en la DSA es el de Gale *et al.* (1992), como mejora del trabajo previo, (Gale y Church, 1991). Sobre el corpus Hansard[155] alineado automáticamente, Gale y Church (1991) obtienen para la palabra *bank* una precisión de 92%. Otros métodos que explotan corpus bilingües son: Brown *et al.* (1991), Hawkins y Nettleton (1998).

Como conclusión a la presentación de los sistemas de DSA basados en corpus etiquetados, resaltamos que obtienen actualmente los mejores resultados en tareas de DSA. En cambio, los sistemas supervisados presentan el problema de estar condicionados a la existencia, en cantidades lo más grandes posible, de datos etiquetados a nivel de sentido y esta tarea es muy costosa y difícil de realizar.

4.3.2 Métodos basados en corpus no etiquetados con sentidos

Una serie de algoritmos tratan de superar el problema del "cuello de botella" en la adquisición de conocimiento léxico, principal desventaja de los sistemas supervisados de DSA. El objetivo de estos métodos es inducir el conjunto de sentidos de una palabra ambigua a partir de un conjunto de ejemplos (no etiquetados) de esa palabra, con lo cual en este caso la DSA se limita a la discriminación de sentidos[156]. De esta manera, se evita la dependencia de datos etiquetados a nivel de sentido.

*Modelo del espacio vectorial***.** Los métodos más utilizados son los que usan *vectores de contenido* (en inglés, *content vectors*) (Schütze, 1992). Estos métodos tratan el contexto de una palabra como se tratan los documentos en recuperación de información. Es decir, al contexto dado se le asocia un vector en un espacio n-dimensional (siendo *n* el número de palabras del contexto) en el que cada fila del vector contiene una función de la frecuencia de esa palabra en el contexto. Sin embargo, para la DSA se han probado medidas de similitud de vectores distintas de las típicas de recuperación de información. Las medidas de similitud entre vectores se aplican para construir conjuntos con los vectores más similares entre sí, mediante técnicas de agrupamiento (en inglés, *clustering*) (Zernik,

[154] De Bruce y Wiebe (1994), ver el apartado 3.1.2.2.
[155] Ver 3.1.2.2.3.
[156] Cf. el apartado 1.1.

1991, Schütze, 1992). Estos conjuntos pueden considerarse como sentidos inducidos a partir de los ejemplos.

Es característico de esta aproximación a la DSA el algoritmo de Schütze (1998), denominado *discriminación basada en grupos de contextos* (en inglés, *context-group discrimination*), similar al algoritmo de Brown *et al.* (1991) dentro del enfoque de desambiguación supervisada. A diferencia del algoritmo bayesiano simple de Gale *et al.* para la DSA supervisada, en este caso se parte de una inicialización aleatoria (en inglés, *random*) de los parámetros, que luego son reestimados por el algoritmo *Expectación-Maximización* (EM) (en inglés, *Expectation-Maximization*). A partir de la inicialización aleatoria de los parámetros, se calcula, para cada contexto de la palabra ambigua, la probabilidad condicionada de que esta palabra se use con un sentido particular. Se usa esta categorización preliminar de los contextos como datos de entrenamiento. Luego se reestiman los parámetros de manera que se maximiza la similitud de los datos para el modelo dado. El algoritmo EM garantiza el aumento, en cada paso, de la similitud del modelo con los datos dados. El criterio para detener la aplicación del algoritmo es que el incremento de la similitud ya no sea significativo.

Una vez que los parámetros del modelo han sido estimados, se pueden clasificar los contextos de la palabra a desambiguar calculando la probabilidad de cada uno de los sentidos a base de las palabras que ocurren en cada contexto, es decir, las palabras que se usan como atributos para la desambiguación. Así, se parte con una probabilidad aleatoria de los sentidos, calculada sin información sobre el contexto, *a priori*, a base de la mera probabilidad, y se calcula a cada paso la probabilidad *a posteriori*, con información ofrecida por el modelo sobre el contexto. Se hace aquí también la asunción bayesiana simple de la independencia de los atributos informativos y se usa la regla de decisión bayesiana. Los resultados son variables, ya que dependen de las distintas inicializaciones, pero en general se sitúan entre un 5% y un 10% por debajo de algunos algoritmos basados en diccionarios. Para los sentidos con una clara correspondencia con un tema particular, el algoritmo funciona bien y la variabilidad es baja, pero falla para las palabras cuyos sentidos son temáticamente independientes.

Una ventaja de este método es que la granularidad es variable, y se puede adaptar con facilidad para producir distinciones entre tipos de usos de granularidad más fina de la que se encuentra en los diccionarios. Esto es útil en aplicaciones como la RI. La granularidad de los sentidos se puede elegir de una de las dos maneras que describimos a continuación. En la primera variante, se aplica el algoritmo con una gama de valores para el número de sentidos. Cuantos más sentidos hay, más estructurado estará el modelo, por lo tanto será capaz de explicar mejor los datos. Como resultado, la mejor similitud posible del modelo será más alta con cada sentido nuevo añadido. Se puede elegir el número óptimo de sentidos de manera automática, comprobando sobre los datos de validación. La segunda variante, más simple, consiste en hacer depender el número de sentidos de la cantidad disponible de material de entrenamiento.

Entre otros métodos no supervisados, destacamos una serie de trabajos que usan la agrupación (*clustering*) en diferentes formalizaciones específicas, cuando la desambiguación corresponde a la asignación de clases (*clusters*). Pereira *et al.* (1993) utiliza clases (*clusters*) de contextos estructurados de palabras, de manera similar al algoritmo de Schütze de 1998, pero basándose en otra formalización de la agrupación (*clustering*). En los trabajos de Zernik (1991) y de Dolan (1994), las representaciones de los sentidos se construyen a través de la combinación de varios subsentidos en un “supersentido”. Estos algoritmos encuentran su utilidad en obtener sentidos más bastos que los del diccionario o en relacionar definiciones de los sentidos entre dos diccionarios. Finalmente, Pedersen *et al.* (1997) conectan la salida del sistema con sentidos de fuentes léxicas estándar, obteniendo un 66% de desambiguación correcta, o sea un resultado bajo, por debajo de la selección del sentido más frecuente, con 77% de precisión.

La gran ventaja de estos métodos es que no están subordinados a los recursos etiquetados manualmente, por lo cual no están limitados en el número de palabras a tratar ni en las distinciones de sentido. Además, son independientes de la lengua, siempre que se disponga de un corpus lo suficientemente grande. Sin embargo, los sentidos inducidos de esta forma no son útiles para aplicaciones concretas. Para adecuarlos a necesidades concretas, a menudo se realiza una alineación posterior con sentidos predeterminados para una aplicación. Esta alineación puede realizarse manualmente o de forma semiautomática: si los sentidos finales se modelan como un vector de

contenido, se pueden aplicar medidas de similitud entre vectores y determinar qué sentido es más similar a cada uno de los conjuntos inducidos automáticamente.

4.4 Métodos mixtos

Hemos ofrecido en los apartados precedentes una imagen sintética de sistemas de DSA relevantes para los diferentes enfoques, fuentes léxicas, clases de conocimiento o algoritmos que se han usado en la DSA. Son, en su gran mayoría, sistemas "simples". Tomando los sistemas enumerados como puntos de referencia, en este apartado presentamos métodos híbridos para la resolución de la ambigüedad léxica. Se trata de combinaciones entre fuentes léxicas estructuradas y corpus: diccionarios y corpus (apartado 4.4.1.), tesauros y corpus (apartado 4.4.2.), bases de datos léxicos y corpus (apartado 4.4.3.), diferentes fuentes léxicas estructuradas y corpus (apartado 4.4.4.) o bien de fórmulas complejas, en que se usan técnicas de combinación de clasificadores o de aprendizaje computacional para juntar diferentes algoritmos (apartado 4.4.5.).

4.4.1 Métodos que combinan diccionarios y corpus

*Diccionario monolingüe y corpus***.** McRoy (1992) combina un lexicón de raíces únicas (aproximadamente 8.800), una jerarquía de conceptos (1.000) y una biblioteca de patrones colocacionales etiquetados con sentidos (1400). Todas las fuentes están construidas especialmente para este sistema de DSA. Las primeras dos se han desarrollado manualmente. En cambio, la última se ha obtenido extrayendo las colocaciones automáticamente, a partir de un corpus del mismo dominio que el texto por desambiguar (de tipo periodístico), y anotando manualmente los sentidos. De estas fuentes, se explotan las siguientes clases de información: etiquetas sintácticas, información morfológica, contexto semántico, colocaciones y asociaciones de palabras, el papel semántico y restricciones de selección.

Cada tipo de información asigna a los sentidos un número de puntos entre –10 y +10, en función del poder discriminativo o la especificidad de la fuente. La combinación de los diferentes tipos de información corresponde a la suma de los votos individuales. Sobre un texto de 25.000 palabras de *Wall Street Journal*, el sistema de McRoy obtiene una cobertura del 98% para los nombres comunes y no abreviados. Sobre la precisión no se dispone de datos.

*Diccionario bilingüe y corpus***.** La hipótesis que está en la base de estos métodos es que los diferentes sentidos de una palabra en una lengua (L1) tienen traducciones distintas a otra lengua (L2). Como tal, el análisis de las diferentes traducciones puede ofrecer información sobre los sentidos de la palabra en la lengua de partida. Vamos a presentar el enfoque a través de dos algoritmos, el de Dagan *et al.* (1991) y el de Dagan e Itai (1994), siguiendo a Manning y Schütze (1999).

El algoritmo de Dagan *et al.* (1991) se propone identificar, para un nombre dado, los sentidos de sus ocurrencias en un texto a base de las diferentes traducciones a otra lengua, en frases que contienen la relación verbo-objeto. A este propósito, usa un diccionario bilingüe para la conexión entre ambas lenguas (L1 y L2) y un corpus de la segunda lengua (L2). La técnica consiste en buscar en el diccionario las traducciones del nombre a la L2 y lo mismo para el verbo que le acompaña. Posteriormente, en el corpus de la L2 se cuenta el número de coocurrencias de las diferentes traducciones del nombre con las traducciones del verbo si se encuentra en la misma relación de verbo-objeto. La frecuencia más alta indica la traducción adecuada a la L2 de la coocurrencia verbo-nombre de la L1 e, implícitamente, el sentido adecuado del nombre en la L1. Dagan *et al.* aplican su algoritmo al caso particular del inglés como L1 y del alemán como L2, con lo cual usan un diccionario bilingüe inglés-alemán y un corpus alemán. Por ejemplo, si se quiere desambiguar la ocurrencia del nombre *interest* en un texto donde es objeto de *show*, el diccionario bilingüe ofrece, entre otras variantes, para los dos sentidos de *interest*, '*legal share*' y '*attention, concern*', las traducciones *Beteiligung* y *Interesse* respectivamente; la traducción alemana de *show* es *zeigen*. El algoritmo calcula, sobre el corpus alemán, la frecuencia de coaparición de *Beteiligung* y *zeigen* frente la frecuencia de coaparición de *Interesse* y *zeigen.* El resultado es netamente favorable al segundo par de traducciones, correspondientes al sentido de *Interesse* ('atención, preocupación'), para la coocurrencia *show interest*. El algoritmo de Dagan e Itai (1994) es más complejo y refina lo que acabamos de explicar, ya que sólo desambigua si se puede tomar una decisión fiable. En el caso en que las frecuencias de coocurrencias

verbo-nombre de las traducciones en el corpus sean equilibradas, la elección de la variante más frecuente implica una gran probabilidad de error[157]. Por ello se propone optar por tomar esta decisión solamente si se supera un umbral muy alto, de un 90%.

4.4.2 Métodos que combinan tesauros y corpus

El ejemplo por excelencia para los métodos que combinan tesauros y corpus en la DSA es el de Yarowsky (1992). El método aplica una técnica de *bootstrapping*[158], es decir se parte con palabras de las categorías de *Roget's*, palabras que se consideran como etiquetadas semánticamente, y se aumenta esta información usando el corpus no anotado. Su enfoque se puede considerar como desambiguación adaptativa basada en un tesauro, con dos operaciones básicas: adaptación de la categoría semántica de las palabras al corpus y desambiguación basada en el tesauro (Manning y Schütze, 1999). La adaptación del tesauro al corpus significa que el algoritmo añade palabras a una categoría del tesauro si aparecen en el corpus, en contextos de la categoría respectiva, un número de veces significativo que no puede deberse a la simple casualidad. Se usa un clasificador bayesiano a la vez para la adaptación y para la desambiguación, es decir se usa la hipótesis bayesiana débil (en inglés, *naive*) para el cálculo de las probabilidades. Yarowsky deriva clases de palabras del corpus tomando como punto de partida las palabras en las categorías usuales del *Roget's*. Para cada palabra de una categoría se extrae un contexto de 100 palabras y se usa una estadística del tipo "información mutua" para identificar las palabras más probables que coocurren con los miembros de una categoría. Las clases así resultantes se usan para desambiguar ocurrencias nuevas de la palabra polisémica. Para ello, se examina el contexto de 100 palabras de la ocurrencia polisémica y se busca el número de coincidencias con las palabras de las diferentes categorías del tesauro. Se aplica la regla bayesiana débil para determinar la clase más probable que corresponde a la palabra polisémica. Tal asignación de la ocurrencia a una clase equivale a la desambiguación, debido a la asunción de Yarowsky de que una clase representa un sentido particular de una palabra. La calidad es del 92% sobre una media de tres sentidos por palabra. Los resultados indican, según Yarowsky, una calidad alta para los sentidos que se alinean bien con los temas, pero baja en el caso de que los sentidos que pertenezcan a diversas categorías temáticas, es decir que no estén relacionados con un tema específico. El método es adecuado para extraer información sobre el tema, la cual, a su vez, es el conocimiento más útil para desambiguar nombres.

En (Yarowsky, 1995), para evitar un error de progresión geométrica en las sucesivas iteraciones propias de los métodos de *bootstrapping*, se aplican los dos supuestos presentados ya en nuestro trabajo: a) que una palabra mantiene el mismo sentido a lo largo de un mismo texto (en inglés, *one sense per discourse*), y b) que una palabra suele tener el mismo sentido en contextos similares (en inglés, *one sense per collocation*). Los vectores de partida para estos métodos se pueden obtener a partir de ejemplos etiquetados a mano, diccionarios o recursos temáticos como el *Roget Thesaurus*. Es una técnica no supervisada, no necesita textos etiquetados para aprender claves. Aunque el método hace uso de definiciones "semilla", sigue siendo no supervisado en el verdadero sentido del Aprendizaje Automático. Las definiciones "semilla" son sólo claves que sugieren sentidos, sin que se provea un esquema de clasificación completa antes de que el algoritmo opere.

4.4.3 Métodos que combinan *WordNet* y corpus

Una de las líneas más seguidas en la DSA de los últimos años es la combinación de *WordNet* con un corpus, generalmente etiquetado con sentidos y a veces también a nivel sintáctico. Las relaciones léxico-semánticas de *WordNet* permiten identificar en el corpus de entrenamiento casos etiquetados similares con el ejemplo por etiquetar. En otros métodos, la combinación permite ampliar el corpus de entrenamiento. A continuación, detallamos algunas de las propuestas de referencia en esta línea de investigación.

Resnik (1995) combina *WordNet* y un corpus para el cálculo de la similitud semántica de las palabras como contenido de información compartido por las palabras. En términos de la jerarquía de *WordNet* basada en la relación ES-UN, Resnik define este contenido de información compartido como la especificidad del concepto que subsume a ambas palabras. Se asume que cuanto más específico es

[157] Fácilmente calculable mediante el porcentaje de aparición de cada uno de los pares de traducciones sobre el número total de apariciones de las posibles traducciones a la L2 del par inicial verbo-nombre de L1.
[158] Ver el apartado 4.1.4.2.

este concepto antecesor, más relacionadas semánticamente serán las dos palabras. En este enfoque, la similitud semántica entre dos palabras equivale al contenido de información máximo entre los contenidos de información correspondientes a los diferentes conceptos antecesores comunes a cualquiera dos sentidos de las palabras:

$$Sim(w_1, w_2) = \max_{C \in subsumer(w_1, w_2)} [-\log P(c)],$$

donde *subsumer(w_1, w_2)* es el conjunto de *synsets* antecesores igualmente a w_1 y a w_2 para todos los sentidos de ambas palabras. El contenido de información es una medida de la especificidad de un concepto dentro de una jerarquía. Así, un concepto con un alto contenido de información es muy específico para un tema, mientras un concepto con un bajo contenido de información es más general. Para estimar el contenido de la información, se cuenta primero la frecuencia del concepto en un corpus amplio, considerando también la frecuencia de todos los conceptos a él subsumidos:

$$P(c) = \frac{Freq(word(c))}{N},$$

donde word(c) es el conjunto de nombres que tienen un sentido por debajo del concepto c en la taxonomía ES-UN de *WordNet* y N es el número de nombres observados en el corpus. De aquí se determina la probabilidad del concepto a través de una estimación de probabilidad máxima y luego su contenido de información se calcula usando la fórmula:

$$IC(concepto) = -\log(P(concepto)).$$

*Información sintáctica***.** En una serie de métodos mixtos, se usa la información que un analizador sintáctico añade al corpus y al contexto de la ocurrencia ambigua. Presentamos algunos sistemas de referencia para este enfoque.

Stetina *et al.* (1998) proponen el que posiblemente es el primer sistema que desambigua todas las clases abiertas de palabras (nombres, adjetivos, verbos, adverbios). Se combina *WordNet*, usado para el cálculo de similitud entre sentidos de palabras, y un corpus etiquetado semántica y sintácticamente, que permite el cálculo de probabilidades de combinación de sentidos de dos palabras en una relación sintáctica dada. La estrategia de Stetina *et al.* consiste en desambiguar simultáneamente todas las palabras de una oración analizada al nivel sintáctico. La hipótesis de partida es que los sentidos de las palabras de la oración están determinados por la combinación más probable entre los sentidos posibles de las palabras en todas las relaciones sintácticas derivadas de la estructura arborescente del análisis sintáctico. Los sentidos de cada palabra tienen probabilidades dadas por el conjunto de relaciones en que la palabra participa en la oración. Cada relación sintáctica extraída tiene asignadas probabilidades, aprendidas sobre el corpus, para las combinaciones de sentido de sus elementos, el núcleo y el modificador. Dentro de una relación, las probabilidades de los sentidos del núcleo dependen de las probabilidades de los sentidos del modificador y a la inversa.

En este enfoque, la tarea de desambiguación corresponde a la selección, para todas las palabras de contenido léxico de la oración, de la combinación de sentidos con la probabilidad relacional general máxima. Esta probabilidad se define como el producto de las probabilidades de las relaciones identificadas en la oración para una combinación dada de sentidos. Se opta por un enfoque de la DSA basado en medidas de similitud entre las relaciones de entrenamiento y las de prueba. Para cada relación, se define una matriz relacional con las probabilidades de la combinación de cada sentido del modificador con cada sentido del núcleo. La similitud se calcula en base a una distancia semántica en *WordNet*, definida de manera diferenciada para nombres, verbos, adjetivos y adverbios, respectivamente. El modelo probabilístico asigna valores de probabilidades a las combinaciones individuales de sentido basadas en las relaciones adquiridas en la fase de aprendizaje. En la combinación de las relaciones de una oración, debido al alto coste computacional, no es factible evaluar las probabilidades generales para todas las combinaciones de sentidos. En cambio, se explota

la estructura jerárquica de las oraciones y se obtiene la combinación óptima en dos fases: 1) la propagación ascendente de las probabilidades de los sentidos de los núcleos; 2) la desambiguación descendiente. En este desplazamiento a lo largo del árbol sintáctico de la oración, se supone que sólo hay relaciones entre el núcleo y los modificadores, con lo cual los sentidos de unos dependen exclusivamente de los sentidos de los otros. Se hace uso de vectores de sentido, en que se registran las probabilidades de los sentidos en el contexto. Las probabilidades iniciales de los sentidos de las palabras se basan en su distribución en todo el corpus de entrenamiento, independientemente del contexto; el número de apariciones se ajusta de manera que no haya probabilidades cero debido a las dimensiones reducidas del corpus. En la primera fase, a cada nivel de la jerarquía, las probabilidades de los sentidos del núcleo se modifican a través de las relaciones que establecen con sus modificadores y, recíprocamente, las probabilidades de los sentidos del núcleo contribuyen al cálculo de las probabilidades de los sentidos de los modificadores. El proceso se repite hasta que se alcanza la raíz del árbol sintáctico. Las probabilidades de los sentidos obtenidos así para el nodo raíz permitirá la selección del sentido para este nodo, el de la mayor probabilidad. En la segunda fase, se parte con el nodo raíz previamente desambiguado y se baja en la estructura del árbol, a través de los núcleos de los sintagmas. En cada nivel se desambiguan los modificadores del nodo núcleo respectivo, usando las probabilidades, en las matrices relacionales, de sus sentidos en combinación con el sentido identificado para el núcleo.

Se usa también una variante del algoritmo de desambiguación, en que se intenta aprovechar el contexto discursivo en una adaptación de la hipótesis de "un sentido por texto". Usando los ficheros semánticos (en inglés, *semantic files*) de *SemCor*, se cuentan las ocurrencias de las palabras de clase abierta que aparecen previamente en el mismo fichero visto como discurso. La conclusión de este experimento es que la hipótesis "un sentido por texto" se verifica bastante para los nombres (en un 88%), pero menos para adjetivos y adverbios y sobre todo para verbos (sólo en un 57%); además, la proporción entre las palabras con el mismo sentido y las palabras con sentido diferente depende de la distancia, en término de oraciones, a la cual al palabra aparece. El algoritmo inicial de DSA se modifica en el siguiente sentido: las palabras desambiguadas y sus sentidos se almacenan. Las palabras de las oraciones de entrada se comparan con estos pares (palabra, sentido); si la palabra se encuentra en este conjunto, entonces la probabilidad inicial para los sentidos asignada en el algoritmo anterior se ajusta con un factor igual a la probabilidad de que la palabra con la misma categoría morfosintáctica que ha ocurrido previamente en un determinado número de oraciones tenga el mismo sentido que su ocurrencia anterior.

La evaluación del método en las dos variantes, sobre quince ficheros aleatorios de *SemCor*, indica una precisión media en la desambiguación de 79,4% y de 80,3%, respectivamente, por encima de la elección del primer sentido en *WordNet*, con 75,2% de precisión.

*Solapamiento relajado con similitud***.** Leacock *et al.* (1998a) amplían el corpus de entrenamiento usado por unos clasificadores estadísticos a través de medidas de similitud basadas en *WordNet*, con el objetivo de disminuir el problema de la dispersión de datos. La hipótesis de partida es que las palabras semánticamente similares deben proporcionar claves contextuales similares. En otras palabras, se usa *WordNet* para generalizar la correspondencia entre la oración de entrada y los ejemplos del corpus de entrenamiento. Cuando no hay solapamiento preciso, o sea cuando en la oración de prueba hay palabras que no se encuentran en el corpus de entrenamiento, se usan las medidas de similitud. El método para combinar el clasificador contextual con la similitud basada en *WordNet* es el siguiente: 1) en la oración de prueba, se sustituyen las palabras que no aparecen en el corpus de prueba con las palabras más similares existentes en el corpus de entrenamiento, mediante el uso de dos métricas; 2) se aplica el clasificador sobre la nueva oración. En la fase 1), las dos métricas de similitud usadas son: la distancia entre dos palabras en *WordNet*, vista como la longitud del camino más breve que une algún sentido de una palabra con algún sentido de la otra, y la medida propuesta por Resnik (1993), como la clase más informativa (o sea la menos probable) a la cual pertenecen los dos conceptos[159]. Sin embargo, las medidas de similitud se usan sólo cuando hay acuerdo entre las dos métricas. En la fase 2), el contexto local está representado por tres tipos de distribuciones: a) categorías morfosintácticas (en una ventana de dos palabras de ambas partes de la ocurrencia ambigua), b) palabras de clase

[159] Hemos presentados la medida de Resnik (1993) en este mismo apartado.

cerrada (preposiciones, conjunciones, determinantes, pronombres) en la misma ventana y c) palabras de clase abierta (nombres, adjetivos, verbos, adverbios), a la izquierda o a la derecha de la ocurrencia focalizada, en una ventana variable. El clasificador sólo obtiene los mejores resultados para una ventana de seis palabras de ambas partes de la ocurrencia ambigua. Sobre el verbo *serve*, se obtiene una precisión variable según la dimensión del corpus de entrenamiento: 75% cuando se entrena sobre 50 oraciones, 79% sobre 100 oraciones y 83% para 200 oraciones. El uso de la similitud permite reducir a más de la mitad el corpus de entrenamiento necesario para obtener la misma precisión en la desambigación. En cambio, manteniendo el mismo corpus de entrenamiento, la combinación del clasificador local con la similitud determina una discreta mejora de los resultados. En concreto, en el caso que este corpus sea de pequeñas dimensiones, se obtiene un 3,5% si se dispone de diez oraciones para cada sentido. Relativamente al número de oraciones de entrenamiento, las precisiones obtenidas por la combinación son las siguientes: 77,7% (50 oraciones), 80,1% (100 oraciones), 83,1% (200 oraciones).

WordNet e Internet como corpus. Mihalcea y Moldovan (1999) desarrollan el método de Leacock *et al.* (1998)[160], explotando también las glosas de los *synsets* en *WordNet* y sustituyendo el corpus con Internet. Su método es diseñado para la desambiguación, en términos de *synsets* de *WordNet*, de todas las clases abiertas de palabras (nombres, adjetivos, verbos, adverbios). El sistema de Mihalcea y Moldovan y el de Stetina *et al.* (1998) se encuentran entre las pocas propuestas de este tipo. El método usa Internet para adquirir información estadística sobre la coocurrencias de palabras y clasificar así los sentidos de las palabras, mientras que *WordNet* ayuda a medir la similitud entre pares de palabras (verbo-nombre, adjetivo-nombre o adverbio-verbo) a través de una métrica llamada *densidad semántica*. La idea de partida es formar pares de palabras a partir del texto y desambiguar las dos palabras considerando a la otra como contexto. Los autores llaman a esta estrategia un enfoque de *dependencia entre palabras* (en inglés, *word-word dependency*). La arquitectura del sistema se compone de dos algoritmos: el primero filtra los sentidos de los nombres, guardando los primeros sentidos más probables. El segundo trabaja sobre los resultados del primero y anota con sentidos cada par de palabras. O sea, la segunda heurística actúa sobre la salida de la primera. Así, en el primer paso, sobre un par de palabras (w_1,w_2), se fija una de las dos palabras (w_1) y se forman listas de similitud para cada sentido de w_2 con los sinónimos y los hiperónimos del sentido en *WordNet*. A partir de estas listas, se forman preguntas en Internet para cada uno de los sentidos de la palabra w_2 como una disyunción de pares formados por w_1 y respectivamente un sinónimo o hiperónimo del sentido, contiguos o a poca distancia[161]. Los sentidos de w_2 se ordenan con respecto a w_1 según el número de respuestas obtenidas en Internet y se guardan sólo los primeros. El mismo procedimiento se hace, fijando w_2, para los sentidos de w_1. De esta manera, se eliminan los sentidos no adecuados y se reduce el número de sentidos entre los que se tiene que desambiguar. Una evaluación sobre casi 400 pares de palabras de *SemCor* indica que la elección de los primeros cuatro (para verbos y nombres) o dos (para adjetivos y adverbios) sentidos de la clasificación obtenida se cubren en buen grado los sentidos relevantes de las palabras, en un promedio por encima del 90%. Si se considera exclusivamente un sentido, la precisión de esta heurística es del 76% para nombres, del 60% para verbos, del 79,8% para adjetivos y del 87% para adverbios. En el segundo paso, se refina el orden previamente obtenido para los sentidos con la ayuda de la densidad semántica. La densidad semántica se calcula como el número de palabras comunes dentro de cierta distancia semántica de dos o más palabras. Así, cuanto más próximas sean dos palabras más alta será la densidad semántica entre ellas. Esta métrica ordena las posibles combinaciones de sentidos de las dos palabras del par. El algoritmo se aplica sólo para pares de tipo verbo-nombre. Por ejemplo, en el caso de un par verbo-nombre dado, para cada combinación de sentidos (v_i,n_j) de las combinaciones guardadas como salida del primer algoritmo, se procede de la siguiente manera:

a) Por una parte, se buscan las glosas de los conceptos verbales dentro de la subjerarquía que incluye el *synset* de v_i (o sea la jerarquía correspondiente al más alto nivel de sus hiperónimos

[160] El método se ha presentado en el apartado anterior.
[161] En el motor de búsqueda AltaVista, hay la opción NEAR para la ocurrencia de un grupo de palabras.

en la jerarquía), y el peso que indica el nivel en esta subjerarquía del concepto verbal. A partir de esto, se extraen todos los nombres que tengan el peso del concepto verbal correspondiente.

b) Por otra parte, se extraen los nombres en la subjerarquía de n_j (o sea la jerarquía que lo tiene como nodo superior).
c) Se calcula luego la densidad conceptual de conceptos previamente obtenidos, comunes entre las dos subjerarquías de v_i, respectivamente de n_j.
d) Se ordenan las combinaciones de sentido.

La combinación de los dos algoritmos se ha comprobado sobre 400 pares de palabras de *SemCor*, con una precisión muy elevada, del 98% para nombres, del 87% para verbos, del 93% para adjetivos y del 97% para adverbios si se guardan dos sentidos para nombres y verbos y dos para adjetivos y adverbios. Eligiendo sólo el primer sentido, la precisión es del 86,5% para nombres, del 67% para verbos, del 79,8% para adjetivos y del 87% para adverbios. Los autores indican como fuente de error las glosas, que no están anotadas con categorías morfosintácticas y proponen un etiquetado de este tipo para *WordNet*[162]. El método es dependiente del motor de búsqueda usado.

En Mihalcea y Moldovan (2000), se propone el entrenamiento con *SemCor* para eliminar esta dependencia. Además, el sistema de DSA se amplía, combinando en cascada ocho heurísticas, dos de ellas supervisadas:

1. Se reconocen las entidades (personas, empresas, etc.), aprovechando la anotación de *SemCor*.
2. Se anotan las palabras monosémicas.
3. Se forman pares de la palabra a desambiguar con la palabra previa y con la palabra sucesiva (no se forman pares con determinantes o con conjunciones). Si estos pares tienen un número de apariciones en *SemCor* por encima de un umbral fijado y en todas la palabra tiene el mismo sentido, entonces se le asigna este sentido.
4. Se construye el "contexto nominal", para cada sentido de la ocurrencia ambigua, con los nombres que pertenecen a los *synsets* hiperónimos del sentido en *WordNet* y con los nombres que aparecen en una ventana de diez palabras alrededor de las ocurrencias de este sentido en *SemCor*. A continuación, se cuentan los nombres comunes entre el contexto nominal y el contexto de la ocurrencia y se elige el sentido con más nombres así encontrados.
5. Se buscan palabras no etiquetadas que se encuentran, en *WordNet*, en el mismo *synset* de alguna palabra previamente desambiguada y se les asigna el sentido correspondiente a aquel *synset*.
6. Se buscan en el contexto dos palabras no desambiguadas que tengan sentidos en un mismo *synset* de *WordNet* y se les asigna aquellos sentidos.
7. Para una palabra no desambiguada, se busca en el contexto alguna palabra previamente desambiguada con el mismo *synset* o con el *synset* hipónimo o hiperónimo de alguno de los sentidos de la palabra ambigua y se le asigna el sentido correspondiente.
8. Se buscan pares de palabras no etiquetadas que tengan alguna combinación de sentidos, en *WordNet*, en relación de hipo/hiperonimia y se les asignan aquellos sentidos.

Evaluado sobre seis artículos de *SemCor*, el sistema obtiene una precisión relativa del 92% y precisión absoluta del 50%.

Los mismos autores (Mihalcea y Moldovan, 2001) proponen en Senseval-2 un sistema de resolución léxica que supone dos nuevos enfoques a la DSA: aprendizaje de patrones y selección activa de atributos[163]. Estos dos algoritmos constituyen los módulos del sistema de desambiguación. En el primer módulo, se aprenden patrones a partir de corpus etiquetados con sentidos (*SemCor* y *GenCor*[164]) y de definiciones de diccionarios (*WordNet*). Para cada palabra etiquetada con sentidos encontrada en el corpus, se construyen patrones que incluyen la palabra y su contexto local, o sea una ventana de máximo M palabras a la izquierda y de máximo N palabras a la derecha (donde M y N tienen el valor preestablecido en 2). Cuando se quiere desambiguar una palabra, se buscan primero los patrones disponibles que corresponden con el contexto de la palabra. Si se encuentran varios patrones, la selección se hace según la fuerza de los patrones. Esta fuerza se mide en términos del número de

[162] Ver, en el apartado 3.1.1.3., *Extended WordNet*.
[163] Los dos enfoques y el sistema de DSA se describen más en detalle en (Mihalcea, 2002b). Aprovechamos aquí ambos trabajos.
[164] Ver el apartado 3.1.2.2.

componentes especificados, número de ocurrencias y longitud. Otro paso importante en el proceso de desambiguación de todas las palabras es la propagación del sentido. Los patrones no garantizan una cobertura completa de las palabras contenidas en el texto de entrada, por lo tanto se necesitan métodos suplementarios. En base a la hipótesis "un sentido por texto", se asigna a cada palabra no desambiguada previamente el sentido de la ocurrencia más cercana, si existe alguna. A las palabras que quedan todavía ambiguas se les asigna el primer sentido en *WordNet*. El segundo módulo se activa sólo para palabras con un amplio número de ejemplos etiquetados de entrenamiento, como es el caso en la tarea *lexical sample* de Senseval.

Para el aprendizaje activo de los atributos[165] de cada palabra por desambiguar, se utiliza un algoritmo de aprendizaje basado en ejemplos[166], Timbl (Daelemans *et al.*, 2001, *apud* Mihalcea, 2002). A partir de un conjunto común de atributos, el algoritmo selecciona los atributos que supuestamente permiten la mejor precisión en la desambiguación. Se obtienen así, en efecto, meta expertos de palabras, que se volverán expertos efectivos de palabras sólo cuando se determinen los valores para el conjunto de atributos previamente seleccionados. Los atributos que se consideran en el conjunto común son características del contexto local que han demostrado su utilidad para el proceso de DS, como la palabra misma y su categoría mofosintáctica, palabras vecinas y su categoría mofosintáctica, colocaciones, eventual núcleo del sintagma nominal al que pertenece la palabra, palabras claves para determinados sentidos, bigramas, entidades nombradas, etc. El sistema ha obtenido los mejores resultados en el marco de Senseval-2, con un 69% para la granularidad fina y un 69, 8% para la granularidad basta en la tarea *all words*, y un 63,8% y un 71,2% para la tarea *lexical sample*.

4.4.4 Métodos que combinan diferentes fuentes léxicas estructuradas y corpus

Lin (1997) combina *WordNet* y una medida de similitud sobre una jerarquía conceptual (como *WordNet* o un tesauro) con un corpus no etiquetado, previamente analizado al nivel sintáctico. El autor parte de la hipótesis de que dos palabras que aparecen en contextos locales idénticos tienen sentidos similares. El contexto local se entiende en términos de dependencias sintácticas entre la palabra por desambiguar y las demás palabras de la oración. Para la representación y el tratamiento del contexto local, se define, para cada dependencia en que participa la palabra ambigua, una tripleta formada por el tipo de la dependencia (sujeto, adjunto, primer complemento, etc.), la palabra con la cual se contrae la relación de dependencia y el papel de la palabra ambigua dentro de la dependencia (núcleo o modificador). Previamente se construye, a partir de un corpus no etiquetado con sentidos pero sí analizado a nivel sintáctico, una base de datos de contextos locales. La base de datos contiene, como entradas, contextos locales, a los cuales se les asocia la siguiente información: la palabra que aparece en el contexto local, su frecuencia de aparición en este contexto y una ratio de probabilidad (en inglés, *likelihood ratio*) entre la palabra y el contexto.

El algoritmo funciona en los siguientes pasos:

1. Se analiza el texto de entrada a nivel sintáctico y se extraen las dependencias sintácticas de cada palabra por desambiguar.
2. Para cada palabra por desambiguar, se buscan en la base de datos sus contextos locales y se extraen todas las palabras que aparecen en el mismo contexto local (serán los "selectores" de la palabra ambigua).
3. Se elige el sentido que maximiza la similitud entre la palabra y el conjunto de sus selectores.
4. Se asigna este sentido a todas las ocurrencias de la palabra en el texto, usando la hipótesis de "un sentido por texto".

En el paso 3, se explota una matriz de similitudes entre los sentidos de la palabra por desambiguar y los sentidos de sus selectores. La medida de similitud usada se basa en la Teoría de la Información (Cover y Thomas, 1991, *apud* Lin, 1997) y se define, para dos conceptos, como la ratio entre la cantidad de información necesaria para establecer lo común entre los conceptos y la cantidad de información necesaria para la descripción completa de los conceptos. Lin aplica el método con un corpus de 25 millones de palabras del *Wall Street Journal* para la extracción de la base de datos de contextos locales y con *WordNet* para el cálculo de la similitud, sobre los nombres de un subcorpus periodístico de *SemCor*, de los cuales unos 2800 son polisémicos. Partiendo de la idea de que algunas

[165] Hemos introducido la idea básica de este enfoque en el apartado 4.1.5.
[166] Para el aprendizaje basado en ejemplo o en memoria, ver el apartado 4.3.1.3.

distinciones de sentidos no son necesarias, para minimizar la excesiva granularidad en la discriminación de sentidos, se usan tres criterios en la evaluación. Los criterios corresponden a tres valores para la similitud entre el sentido correcto de *SemCor* y la respuesta del sistema. Tendrá el valor 1 en caso de coincidencia, superior a 0, si los dos conceptos están debajo del mismo concepto raíz (en inglés, *top-level*) de *WordNet*, y superior a 0,27, como compromiso entre los dos extremos. El valor de 0,27 es la similitud media de 50000 pares de palabras aleatoriamente generados con palabras de la misma categoría en el *Roget's Thesaurus*. La precisión alcanzada es de 73, 6% para similitud superior a 0, de 68, 5% para similitud superior a 0,27 y de 56,1% para similitud igual a 1.

En (Agirre *et al.*, 2000), los autores desarrollan el sistema con que habían participado en la primera edición de Senseval. Partiendo de la idea que el uso exclusivo de métodos supervisados no es realista debido al alto coste de la intervención humana requerida, los autores proponen un sistema de DSA que combina ocho heurísticas, sólo una de ellas supervisada. Las heurísticas no supervisadas son las usadas en el trabajo de Rigau *et al.* (1997)[167], pero implementadas para el inglés y sobre fuentes léxicas distintas: el diccionario *Hector* de Senseval-1[168], *WordNet* y el *Collins English Dictionary*, mientras que la heurística supervisada se basa en las listas de decisión al estilo de Yarowsky (1994). Para la combinación de las heurísticas, se suman los votos normalizados[169] de cada heurística. La evaluación del sistema mixto sobre los datos de prueba de Senseval-1 para nombres, en inglés, indica 66,9% de precisión absoluta, 68,3% de precisión relativa, 98% de cobertura. La experimentación demuestra, por una parte, que la combinación obtiene una calidad en la desambiguación sensiblemente superior a las heurísticas individuales, con más de 15% por encima de la mejora heurística; por otra parte, que las heurísticas supervisadas (todas juntas con 40,4% de precisión absoluta, 43,5% de precisión relativa, 93% de cobertura) y la no supervisada (con 51,3% de precisión absoluta, 71,6% de precisión relativa, 71,6% de cobertura) colaboran bien, mejorando sus resultados separados, en todas las tres medidas.

4.4.5 Métodos basados en técnicas de combinación de clasificadores y en aprendizaje computacional

Una tendencia cada vez más marcada en los últimos años es utilizar las técnicas del área de la combinación de clasificadores para construir sistemas de DSA que incorporan diferentes clasificadores: Mooney (1996), Rigau *et al.* (1997), Escudero *et al.* (2000a, 2000b), Brill *et al.* (1998), Pedersen (2000), Yarowsky *et al.* (2001), etc.

Ilustramos esta estrategia con la propuesta de Florian *et al.* (2002). En este trabajo, se amplía el sistema de Yarowsky *et al.* (2001), que obtuvo los mejores resultados en Senseval-2 en las tareas supervisadas en que ha participando: *all words* para estoniano y checo y *lexical sample* para inglés, español, sueco y vasco. Aunque los algoritmos y técnicas presentados son aplicables igualmente a las tareas *all words* y *lexical sample*, el estudio de 2002 se desarrolla exclusivamente en el ámbito de la última. El trabajo investiga varias cuestiones fundamentales relacionadas con la combinación de sistemas para la tarea de DSA, desde la estructura algorítmica hasta la estimación de los parámetros. Se ofrece una descripción y una evaluación comparativa del problema de la combinación de clasificadores sobre un conjunto de clasificadores heterogéneos en cuanto a estructura y procedimientos:

a) un clasificador bayesiano simple;
b) un clasificador coseno (ambos modelos, a y b, extendidos con atributos más ricos y con peso para los diferentes tipos de atributos;
c) un clasificador bayesiano con enfoque bolsa de palabras (*BayesRatio*);
d) un clasificador basado en listas de decisión no jerárquicas (Yarowsky, 1996);
e) un clasificador con aprendizaje basado en transformación y
f) un modelo mixto de clasificadores basados en el método de corrección de la variación máxima[170].

[167] Ver el apartado 4.2.2.4.
[168] Ver el apartado 5.2.1.
[169] Es decir los votos de cada heurística para los diferentes sentidos se dividen por el voto más alto entre los que la heurística da a los sentidos.
[170] Para detalles sobre los modelos enumerados, enviamos al trabajo de Florian *et al.* (2002).

Por una parte, los seis clasificadores tienen un buen nivel individual de precisión y resultados variables de una lengua a otra, lo que aporta una buena base de partida en la combinación. Por otra parte, hay un grado de acuerdo relativamente bajo entre ellos, lo que permite explotar las diferencias de manera sistemática, para mejorar la precisión del conjunto.

Para la combinación de los clasificadores, se aplican varias estrategias:

1) votación como suma de votos de parte de los clasificadores que separadamente eligen sólo el sentido más probable;
2) votación a través de la interpolación de los votos de los clasificadores[171], usando pesos para los clasificadores inversamente proporcionales con el lugar que ocupan en su ordenamiento según la precisión;
3) combinación basada en confidencia, o sea los pesos de los clasificadores en la votación varían en función de la confidencia combinada para cada muestra;
4) combinación basada en los resultados, en cual caso a los clasificadores se les asignan pesos basados en la precisión estimada sobre el corpus de entrenamiento;
5) metavotación, que tiene en cuenta los resultados diferentes, pero paralelos, de los clasificadores de una lengua a otra y además combina con pesos iguales las salidas de los k clasificadores con mejores precisiones para las distintas lenguas, obteniéndose así un clasificador con los resultados próximos al mejor clasificador para cada lengua.

La evaluación sobre los datos de prueba en la *tarea lexical sample* de Senseval-2 para cuatro lenguas (español, inglés, sueco y vasco) revela que la combinación mediante la metavotación entrena una mejora sustancial de la precisión de los mejores clasificadores individuales en cada lengua: para inglés, 66, 3%; para español, 72,4%; para sueco, 71,9%; para vasco, 76,7%. Los resultados representan el nivel más alto de precisión sobre estos datos hasta la fecha.

Cerramos la presentación de los métodos mixtos de DSA con las conclusiones de algunos experimentos desarrollados últimamente. Éstos parecen indicar que la combinación de fuentes de información o de clasificadores, como los de Wilks y Stevenson (1996, 1997), Stevenson y Wilks (2000, 2001) o de Florian *et al.* (2002), permiten alcanzar un nivel de desambiguación muy por encima de los sistemas “simples”.

En el presente capítulo hemos realizado una síntesis de los métodos propuestos para la resolución de la ambigüedad léxica. Hemos introducido la presentación con unos preliminares destinados a sentar las bases terminológicas, taxonómicas y metodológicas de los métodos. Hemos descrito las características y los ejemplos representativos de cada uno de los tres enfoques básicos a la tarea: DSA basada en fuentes léxicas estructuradas, DSA basada en corpus y DSA mixta entre los dos enfoques previos. A la vez, hemos resaltado las limitaciones de cada enfoque y hemos recordado propuestas para su solución.

[171] Hemos definido la interpolación de métodos en el apartado 4.1.4.

5 Metodología de la DSA (III): evaluación y comparación de los sistemas de DSA

5.1 El problema

La calidad de los resultados es fundamental para la DSA, como para toda tarea de PLN, con lo cual la evaluación de los sistemas de desambiguación es un problema central en el área. En el caso específico de la DSA, la comparación de los resultados obtenidos en diversas evaluaciones se ha visto dificultada por las muchas variables incontroladas implicadas en el proceso: medidas de evaluación, material de entrenamiento, granularidad de las distinciones de sentido, la capacidad de desambiguar las categorías morfosintácticas (POS), las lenguas usadas, etc. (Rigau, 1998).

Stevenson y Wilks (2001) consideran que la evaluación en el campo de la DSA es distinta de otras tareas del PLN, en diferentes aspectos: la cobertura de los sistemas (un pequeño conjunto de palabras *vs.* vocabulario no restringido); la controversia alrededor de la existencia misma de la DSA; la novedad de sentidos, que hace inadecuada la metodología estándar de evaluación (etiquetar, modelizar, evaluar) usada en PLN; la variedad y multiplicidad tipológica de la información explotada para realizar la tarea de desambiguación.

*Estrategias de evaluación***.** El hecho de que la DSA sea una tarea intermedia que contribuye a una tarea global hace que se puedan delinear dos modalidades para su evaluación (Ide y Véronis, 1998): evaluación *in vitro*, cuando los sistemas se evalúan independientemente de alguna aplicación, y evaluación *in vivo*, cuando los resultados se evalúan en cuanto a su contribución a los resultados generales de un sistema diseñado para una aplicación particular.

La evaluación *in vitro*, si bien artificial, permite examinar los problemas que plantea la desambiguación. El desarrollo reciente de sistemas de DSA en buena medida independientes de aplicaciones particulares ha permitido el progreso de la evaluación *in vitro*, que parece dominante de momento. Esta evaluación se puede realizar: 1) de manera *declarativa*, mediante la comparación del salida (en inglés, *output*) de un sistema para una entrada (en inglés, *input*) dada, a base de diferentes medidas[172]; 2) de manera *tipológica* o *diagnóstica*, mediante el estudio del comportamiento del sistema sobre una serie de casos distintos, correspondientes a diferentes problemas lingüísticos relacionados con la desambiguación. La segunda estrategia de evaluación necesita un entendimiento previo, de considerable profundidad, de los factores implicados en la DSA. Con lo cual, la línea principal seguida actualmente en la evaluación de la DSA es de tipo declarativo.

Heurísticas de referencia. En los años noventa se ha perfilado un interés creciente en la línea de delinear métodos para la evaluación de la DSA que ofrezcan un punto de referencia para evaluar la calidad del trabajo que se realiza. Un punto de partida lo constituye el trabajo de Gale *et al.* (1992), quienes proponen establecer un límite inferior y uno superior para los resultados de la DSA: el límite inferior corresponde a la elección del sentido más frecuente, mientras que el límite superior corresponde al etiquetado humano. Los autores aseveran como correcta la elección del sentido más frecuente en un 75% de los casos. Respecto al límite superior, el etiquetado humano, hay muchas controversias sobre su fiabilidad, alimentadas por resultados contradictorios obtenidos en varios experimentos[173]. Otra contribución es la de Miller *et al.* (1994): usando el etiquetado morfosintáctico y semántico de las palabras de clase abierta del corpus *Brown*, crean criterios de referencia (en inglés,

[172] En el apartado 4.1.6., hemos presentado algunas de las medidas más usadas en la evaluación de los sistemas de DSA: precisión absoluta y relativa, cobertura, medida F. Para otras métricas de evaluación propuestas para la DSA, enviamos a (Melamed y Resnik, 1997), (Resnik y Yarowsky, 2000), (Stevenson y Wilks, 2001).
[173] V. el capítulo 1.

benchmarks) para los sistemas de DSA en la elección de los sentidos: la casualidad, la frecuencia (el sentido más frecuente), la coocurrencia. En este caso la heurística más frecuente lleva al sentido correcto el 58% de las veces.

Tales técnicas básicas de desambiguación, que usualmente no implican ningún tipo de conocimiento lingüístico, se suelen tomar como punto de referencia para la evaluación de los sistemas de DSA. En la bibliografía de lengua inglesa se llaman *baselines*; en la presente tesis se denominarán *heurísticas de referencia*.

5.2 Senseval[174]

En 1997, bajo la supervisión del grupo SIGLEX[175], se sentaron las bases de una competición libre y voluntaria, denominada Senseval[176], con el propósito de explorar los aspectos científicos y técnicos de la desambiguación semántica automática y así poder establecer unas bases objetivas para la evaluación de estos sistemas.

5.2.1 Senseval-1

Desarrollado en la línea de la evaluación cuantitativa, Senseval (1998) ha sido la primera evaluación abierta de los sistemas de DSA. Para una presentación general de la competición nos remitimos al número especial de *Computers and Humanities*[177]. Nos limitamos aquí a ofrecer las características generales de la competición y a sintetizar las conclusiones que ésta ha permitido perfilar, siguiendo principalmente a Kilgarriff y Rosenzweig (*idem*).

Tareas, datos y fuentes léxicas, desarrollo de la prueba. En esta primera edición, se ha optado por la tarea de DSA limitada a un conjunto restringido de palabras, o sea la variante *lexical sample*. Como fuente de referencia para el inventario de sentidos, se ha elegido la base de datos léxica HECTOR (Atkins, 1993, *apud* Kilgarriff y Rosenzweig, 2000). Además, para los sistemas cuya salida consistía en sentidos del *WordNet*, se ha asegurado el enlace (en inglés, *mapping*), entre los sentidos de *WordNet* y de HECTOR.

De cara a la controversia sobre si separar el etiquetado morfosintáctico (en inglés, POS *tagging*) de la DSA[178], generalmente se separaron ambas tareas: la clase de la palabra (nombre, verbo, adjetivo) formaba parte de la entrada del sistema de desambiguación. A cada ocurrencia por desambiguar se le añadió una etiqueta sobre la clase: *-n* (nombre), *-v* (verbo), *-a* (adjetivo) o *-p* (para 'categoría no provista'). Los dos tipos de datos, entradas léxicas en el diccionario e instancias en el corpus etiquetadas a mano, estaban destinados a cubrir las necesidades de ambas clases de sistemas de DSA participantes en la competición: los sistemas basados en el conocimiento y sistemas basados en corpus. Estos datos se suministraron a los sistemas en tres fases sucesivas: para la adaptación de los sistemas al formato y estilo del ejercicio, para el entrenamiento, y para la evaluación respectivamente.

Gold standard. Como nivel de referencia para la evaluación, se ha creado un *estándar de referencia* (en inglés, *gold standard*), basado en las "respuestas de control", o sea las respuestas de los anotadores humanos. El *gold standard* constituye un aspecto crítico del sistema de evaluación, ya que de él depende la validez del ejercicio mismo. Por esta razón, su creación ha requerido esfuerzos y atención particular, concretados en un nivel alto de validez del *gold standard* que se estima en el 95%. Los materiales de este proceso de etiquetado manual, sin la correspondiente desambiguación de las palabras ambiguas, fueron enviados a los diferentes sistemas sobre los que se realizaba el experimento, esperándose las repuestas en un intervalo de tiempo predefinido. Las respuestas recibidas por parte de cada sistema fueron evaluadas según la métrica (*score*) establecida y los resultados fueron presentados y analizados en un *workshop*.

---

[174] Consultar el sito http://www.Senseval.org/ para más detalles.

[175] Abreviación de Special Interest Group of the Association for Computational Linguistics, que provee un marco de referencia para la investigación en lexicografía, semántica léxica computacional y otras áreas afines. Para una presentación general, consúltese el sito: http://www.siglex.org/.

[176] De "SENSe EVALuation".

[177] Ide y Mylonas (2000) (eds.), *Computers and Humanities*, **34 (1-2)**, y Kilgariff y Rosenzweig (2000).

[178] V. el apartado 1.5.

*Procedimiento de evaluación***.** La evaluación (en inglés, *scoring*) se realizó teniendo en cuenta tres niveles de granularidad: 1) *granularidad fina* (en inglés, *fine-grained*), donde han contado sólo las etiquetas exactas, idénticas con las respuestas de control; 2) *granularidad basta* (en inglés, *coarse-grained*), donde las etiquetas de subsentidos se han asimilado a las de sentidos, con lo cual se ha restado importancia a la identificación de los subsentidos, y se ha valorado tan sólo la anotación a nivel de sentido; 3) *granularidad mixta*, donde se ha asignado todos los puntos posibles al etiquetado propuesto si era subsumida por la respuesta de control, y una parte de los puntos si subsumía la respuesta de control.

Tipología de los sistemas participantes. En la competición para el inglés, que aquí presentamos, han participado dieciocho sistemas, muy distintos en cuanto a los datos de entrada y a la metodología seguida. Para la comparación, fueron divididos en dos categorías: supervisados y no supervisados. Algunos de los sistemas no supervisados eran flexibles, con posibilidad de transformarse, en mayor o menor grado, en supervisados. Otras lenguas implicadas en la competición fueron, además del inglés, el francés y el italiano, reunidas en un ejercicio paralelo, Romanseval. Para más detalles nos remitimos a (Segond *et al.*, 2000) y a (Calzolari *et al.*, 2000).

Heurísticas de referencia. Los resultados de los sistemas se han medido con respecto a dos conjuntos de heurísticas de referencia (en inglés, *baselines*): las que usan datos del corpus de entrenamiento para la comparación con los sistemas supervisados, y las que usan definiciones y ejemplos ilustrativos, para la comparación con los sistemas no supervisados. Ninguna de las de heurísticas de referencia se basa en alguna forma de conocimiento lingüístico, a excepción de las empleadas a la vez con el filtro de sintagmas (en inglés, *phrase filter*), que reconoce las formas flexionadas y aplica restricciones elementales de orden para las expresiones multipalabra (en inglés, *multiwords*). Las heurísticas de referencia usadas son las siguientes:
- RANDOM: se da peso igual a todas las etiquetas de sentido que corresponden a la forma raíz de la palabra o, para las tareas -*nva*[179], a su categoría sintáctica;
- COMMONEST: se elige siempre el sentido más frecuente de las etiquetas de sentido posibles en el corpus de entrenamiento que corresponde a la forma base (raíz) de la palabra o, para las tareas -*nva*, a su categoría sintáctica;
- LESK: se elige el sentido de la palabra raíz cuya definición, tanto en el diccionario como en los textos de los ejemplos, tiene el mayor número de palabras en común con las palabras alrededor de la instancia por desambiguar;
- LESK-DEFINITIONS: como LESK, pero usando sólo las definiciones, sin el texto de los ejemplos;
- LESK-CORPUS: como LESK, pero considerando también los datos etiquetados del corpus de entrenamiento, cuando están disponibles, para una comparación con los sistemas supervisados;
- FILTER-PHRASE: todas las anteriores fueron conectadas también con este filtro de sintagmas, diseñado para identificar, mediante un análisis superficial, las expresiones multipalabra a modo de preprocesador: si no encuentra evidencia de que se trata de una expresión multipalabra, inhibe sus sentidos correspondientes; si encuentra evidencia para una de las unidades del diccionario, prohíbe los sentidos correspondientes a las expresiones multipalabra.

Resultados. Sin entrar en los detalles del cálculo de los resultados, nos interesa la interpretación en su conjunto. Sintetizamos algunas de las observaciones más importantes a continuación. Se logró un éxito notable para el etiquetado manual de los sentidos (*gold standard*), de hasta un 95%. El nivel de la DSA para granularidad fina, con datos de entrenamiento disponibles, fue del 75% (o incluso de hasta el 80%). En este último caso, se observó que los sistemas supervisados muestran resultados considerablemente mejores que los que los no supervisados. Los sistemas no supervisados pensados para ser tratados en técnicas supervisadas, o bien para apoyarse en ejemplos del diccionario si no hay datos disponibles en el corpus de entrenamiento, obtuvieron mejores resultados en la variante supervisada. Todo ello demuestra que, si se usan datos para entrenamiento, el resultado es mucho mejor. Para los nombres, los mejores resultados se situaron por debajo del 80%; para los verbos, los

[179] Palabras con la clase especificada en la entrada que se proporciona a los sistemas de DSA.

mejores resultados alcanzaron un 70%; para los adjetivos o categoría indeterminada, los mejores resultados oscilaron entre el 70 y el 80%.

Los sistemas fueron comparados con las heurísticas de referencia Lesk correspondiente a su tipo de sistema, pero la mejora fue de tan sólo del 2% (para verbos); un algoritmo simple LESK bien implementado es difícil de ser superado.

Hubo una limitación notable en el caso de etiquetado según otro inventario de sentidos, diferente de Hector; por ejemplo: los sistemas cuyas respuestas estaban relacionadas con *synsets* de *WordNet*, tuvieron una importante desventaja: el mismo *gold standard*, en el proceso de "traducción" (en inglés, *mapping*) desde HECTOR a *WordNet* y de nuevo a HECTOR, estableció un límite máximo para estos sistemas del 79%.

Análisis. Respecto de la evaluación, una limitación fue la no graduación de los ítems según la dificultad: como casi todos los sistemas etiquetan sólo una parte del conjunto de datos, no se puede decir si han elegido subconjuntos igual de difíciles para comparar correctamente sus resultados.
Otras observaciones, en un plano más teórico, tienen que ver con la polisemia, la entropía[180] y la dificultad de la tarea: a) la distribución de las etiquetas de sentido en los datos de entrenamiento y de evaluación está altamente distorsionada, con muy pocas etiquetas comunes y una gran cantidad de etiquetas raras; esto sugiere que las distribuciones de las etiquetas de sentido para palabras individuales en los datos será también distorsionada y que la entropía de estas distribuciones será relativamente baja; en todo caso, hay una consistente variación de la entropía a través de las palabras; b) la polisemia y la entropía a menudo varían juntas, pero no siempre: los nombres suelen tener polisemia más alta, mientras que los verbos suelen tener entropía más alta; c) los sistemas tienden a tener resultados mejores para los nombres que para los verbos; lo que indica que la entropía es la mejor medida para la dificultad de la tarea.

5.2.2 Senseval-2

La reedición del ejercicio, en Tolouse, 5-6 de julio de 2001, se realizó sobre bases algo diferentes. Esta vez el objetivo era evaluar los problemas de los sistemas de DSA respecto de diferentes tipos de palabras, diferentes variedades de lenguaje y diferentes lenguas. Preiss e Yarowsky (2002) ofrecen una presentación de la competición en su conjunto que seguimos en este apartado. Sintetizamos aquí algunos aspectos relevantes, focalizando las novedades con respecto a la primera edición.

Tareas. Para esta edición se definieron tres tareas: 1) *léxico no restringido* (en inglés, *all-words*): etiquetar la mayoría de las palabras de clase abierta de una muestra de texto; 2) *inventario limitado de palabras* (en inglés, *lexical sample*): para un pequeño conjunto de palabras seleccionadas, etiquetar varias instancias suyas en breves fragmentos de texto; 3) *traducción* (en inglés, *translation*): como en el caso precedente, con la diferencia de que las palabras se definen de acuerdo con su traducción.

Lenguas. Uno de los propósitos de esta edición fue promover la participación de nuevas lenguas, con lo cual, los 93 sistemas participantes trataron 12 idiomas. De acuerdo con las tres tareas, la participación de las lenguas fue la siguiente: 1) *all-words:* checo, holandés, inglés, estoniano; 2) *lexical sample:* español, inglés, italiano, japonés, coreano, sueco, vasco; 3) traducción: japonés.

Datos y fuentes léxicas. Los tipos de datos proporcionados variaron ligeramente frente a Senseval-1: a) un lexicón con correspondencias (en inglés, *mappings*) entre palabras y sentidos, con la posibilidad de información suplementaria para explicar, definir o distinguir los sentidos (p.ej. *WordNet*); b) un corpus de texto o muestras de texto etiquetadas a mano, como *gold standard*, que se podía dividir

[180] La *entropía* es la medida de la incertidumbre sobre lo que un mensaje transmite. Para la interpretación computacional y para la definición formal de la entropía, consúltese, por ejemplo, (Charniak, 1993: 27-31). La entropía semántica se define en (Melamed, 1997) como una medida de la ambigüedad y de la no informatividad de una palabra, o sea se le ve inversamente proporcional con el contenido de información, con el peso semántico y con la consistencia en la traducción. Melamed propone una medida para su medición basada en las distribuciones de las traducciones de una palabra en un corpus bilingüe paralelo. Sin embargo, según el autor, cuando sólo se dispone de datos monolingües, el logaritmo de la frecuencia en un corpus es un buen indicador para la entropía semántica.

opcionalmente en corpus de entrenamiento y corpus de prueba (en inglés, *test*); c) una jerarquía o agrupamiento de sentidos (opcionales), para permitir distinciones finas o bastas en el cálculo (en inglés, *scoring*) de las respuestas. Para la tarea *all-words* se proporcionó un texto de 5.000 de palabras, con las de clase abierta etiquetadas, y para la tarea *lexical sample*, un mínimo de 75 + 15*n* ocurrencias etiquetadas para cada palabra, donde *n* es el número de sentidos de la palabra. Para la versión inglesa, se eligió una combinación entre el *British National Corpus* (la edición nueva) con subcorpus limitados del *Wall Street Journal* para el inglés americano. Una novedad con respecto a la edición anterior, posiblemente la más importante, fue el uso del *WordNet 1.7*, y del *EuroWordNet*, en sus versiones castellana, italiana y estoniana, como lexicón de referencia para el inventario de sentidos.

Para cada tarea, los datos se proveyeron en tres etapas: ensayo (en inglés, *trial*), entrenamiento (en inglés, *training*) y prueba (en inglés, *test*). Los equipos tuvieron a disposición veintiún días para trabajar con los datos de entrenamiento y siete con los datos de prueba.

Evaluación. En Senseval-2 se usó la modalidad de evaluación establecida en la edición anterior, con ligeras modificaciones. Así, se aplicó la *evaluación de granularidad fina* para todos los sistemas. En este caso, los sistemas deben proponer al menos uno de los sentidos del corpus *golden standard*. Si hubo disponible una jerarquía o un agrupamiento de sentidos, se aplicó también la *evaluación de granularidad basta*. En esta evaluación, todos los sentidos devueltos como respuesta por el sistema se colapsan al más alto ancestro común o bien al identificador del grupo de sentidos[181]. Para las jerarquías de sentido, se aplicó además una *evaluación de granularidad mixta*: se asignan puntos a las respuestas que eligen un sentido relacionado con el sentido requerido[182].

Resultados y análisis. El *workshop* que concluyó el ejercicio se estructuró alrededor de una serie de problemas de la DSA y su evaluación: desambiguación en dominios específicos; el diseño de la tarea para nuevas lenguas en SenseEval; distinciones de sentido; aplicaciones de la DSA; estandardización de los *WordNets*. Los resultados de Senseval-2, en su conjunto, representan un retroceso frente a Senseval-1, tanto para nombres y adjetivos, en media con un 14% (Kilgarriff, 2002), como para los verbos (Palmer *et al.*, 2002). Los resultados han sido confirmados también para el italiano (Bertagna *et al.*, 2002). Kilgarriff (2001) atribuye el descenso al uso del *WordNet*: en su elaboración, se ha dado prioridad a la construcción de los *synsets* frente al análisis coherente de los diferentes significados de una palabra, mientras que la DSA necesita unas distinciones de sentido claras y bien motivadas. Se acordó que esta cuestión debe constituir la base de investigaciones futuras en DSA. Palmer *et al.* (2002) añaden como causa la cantidad inferior de material de entrenamiento y la dificultad superior de las palabras de test. Sin embargo, a diferencia de lo ocurrido en Senseval-1, para los verbos hubo sistemas que superaron la heurística de referencia más alta, *LESK-CORPUS*, con un nivel del 70% para las distinciones finas y del 90% para las distinciones bastas. Desde la perspectiva de la variante castellana del ejercicio (Rigau *et al.*, 2002), se señaló la diferenciación de los resultados en función de la clase de la palabra: algunos sistemas etiquetaron mejor para nombres y verbo, otros para adjetivos. A la vez, se propuso incluir un criterio de evaluación sensible al dominio en que se efectúa la desambiguación (en inglés, *cross-domain*) en las futuras ediciones de Senseval.

Senseval-2 abrió nuevas vías en la investigación, igualmente de la DSA y de la polisemia, en una relación dialéctica. Partiendo de la constatación de que los sistemas basados en aprendizaje supervisado obtienen los mejores resultados, dos focos de interés son el diseño de métodos para la obtención de corpus etiquetados a gran escala y la selección de los atributos en relación con el tipo de polisemia a tratar. Se espera que el análisis del impacto que un conjunto de atributos y algoritmos han tenido sobre la desambiguación de diferentes palabras permita identificar tipos de polisemia. La comparación entre los resultados de las dos ediciones ha puesto de manifiesto la necesidad de identificar unos criterios y una metodología rigurosamente para la elaboración de los inventarios de sentidos que se toman como punto de referencia en la DSA. La cuestión de la discriminación de los sentidos está íntimamente vinculada con el desarrollo de la tarea de la DSA fuera *vs.* dentro de una aplicación o de un dominio particular, debido a que en este último caso se cuestiona la necesidad de una distinción entre los sentidos y de un módulo separado de DSA.

[181] Cf. *ELRA Newsletter, 7 (3),* 2002.
[182] Según la métrica propuesta por Melamed y Resnik (1997).

5.2.3 Senseval-3

La tercera edición de Senseval se ha desarrollado entre febrero y abril de 2004. El *workshop* para la presentación, análisis y comparación de los resultados obtenidos por los sistemas de DSA participantes en el ejercicio tuvo lugar el 25 y el 26 de julio de 2004, en el marco de la conferencia ACL de Barcelona.

Tareas. Respecto a las ediciones anteriores, Senseval-3 aportó una serie de novedades, ante todo en cuanto a las tareas. Así, se inauguraron las tareas de adquisición automática de subcategorización, inventario multilingüe de palabras, DSA de glosas de *WordNet*, papeles semánticos, formas lógicas, que detallamos a continuación.

La tarea de adquisición automática de subcategorización supone la evaluación de los sistemas de DSA en el contexto de este proceso. La tarea se organizó, en inglés, para 30 verbos difíciles, o sea altamente frecuentes y con muchos sentidos, cada verbo con unas 1000 ocurrencias. Las asignaciones de sentido recibidas de parte de los participantes, en términos de *synsets* de *WordNet 1.7.1*., se traducen a clases verbales al estilo de Levin, y se introducen, como entrada, en el sistema de adquisición de subcategorización de Anna Korhonen. Los esquemas (*frames*) así adquiridos se evalúan luego contra un conjunto de esquemas *gold standard* obtenidas a mano, lo que proporciona la clasificación de los sistemas de DSA.

La tarea multilingüe de inventario limitado de palabras (en inglés, *multilingual lexical sample*) tuvo como objetivo crear un marco para la evaluación de sistemas de Traducción Automática. En vez de usar el inventario de sentidos de un diccionario, se usan las traducciones de las palabras por desambiguar en una segunda lengua. Los contextos son en inglés y las etiquetas de las palabras por desambiguar son sus traducciones a una segunda lengua. Se eligieron palabras con diferentes grados de ambigüedad interlingüe. La tarea se organizó para dos pares de idiomas, inglés y francés, respectivamente inglés e hindi, con aproximadamente cincuenta palabras por desambiguar en cada caso. Los datos se coleccionaron a través de *Open Mind Word Expert* (edición bilingüe)[183].

La desambiguación de las glosas de WordNet se desarrolló usando el etiquetado manual de glosas realizado dentro de los proyectos *WordNet 2.0* y *Extended WordNet*[184] como corpus de entrenamiento y de prueba. La tarea se concebió como *all-words*, o sea se debían desambiguar todas las palabras de contenido léxico de las glosas: nombres, adjetivos, verbos, adverbios.

El etiquetado de papeles semánticos se desarrolló en el marco de *FrameNet*[185], que proporcionó también los datos etiquetados a mano que se tomaron como punto de referencia. La tarea se organizó sobre las bases del trabajo de Gildea y Jurafsky (2000), en que se proponen un conjunto de métricas para la evaluación de los sistemas.

La identificación de formas lógicas se organizó sólo para inglés, con el propósito de transformar las oraciones del inglés en una notación de la lógica de primer orden. Las palabras de contenido léxico corresponden a predicados, mientras que las conjunciones, las preposiciones y los argumentos tienen valores sintácticos. Los resultados se evaluaron al nivel de la oración y del predicado, con la ayuda de unas medidas de precisión absoluta y relativa contra un *golden standard* construido a mano.

Además, el ejercicio Senseval se abrió hacia otras lenguas (como el chino y el rumano) en la tarea de DSA para inventario limitado, mientras que para el italiano se organizó por primera vez la tarea para inventario ilimitado. Como novedad también, por primera vez se coordinaron (parcialmente) las tareas de inventario limitado en varias lenguas; así, se han eligieron diez palabras comunes para el catalán, el español, el inglés, el italiano, el rumano y el vasco.

[183] Ver nota 188.
[184] Hemos presentados ambas fuentes léxicas en el apartado 3.1.1.3.
[185] Ver el apartado 3.1.1.3.

Nuestro principal interés se dirige hacia las tareas "clásicas" de Senseval, de relevancia para la presente tesis, por lo que a continuación nos centramos en la tarea de inventario limitado[186]. Para el inglés, el ejercicio utilizó como fuentes léxicas de referencia *WordNet 1.7.1.* (para nombres y adjetivos) y *Wordsmyth*[187] (para los verbos). El corpus etiquetado con sentidos se obtuvo mediante el sistema *Open Mind Word Expert*[188], con un nivel de acuerdo entre los anotadores de 67,3%. La heurística de referencia del sentido más frecuente alcanzó el 55,2% para la granularidad fina y el 64,5% para la granularidad basta. Los resultados de los sistemas participantes, en su gran mayoría, superaron sensiblemente estos niveles. Para los sistemas supervisados, se alcanzó el 72,9% para la granularidad fina y el 79,3% para la granularidad basta, mientras que los sistemas no supervisados obtuvieron una precisión absoluta del 65,7% y del 74,1% respectivamente. Varios sistemas mejor clasificados consistieron en una combinación de clasificadores, lo que confirmó que los sistemas complejos superan a los clasificadores individuales. A la vez, el ejercicio demostró que se pueden obtener sistemas no supervisados de buena fiabilidad (Mihalcea *et al*., 2004).

Respecto a la edición anterior de Senseval, el progreso de la calidad de los sistemas no supervisados es de 25,6% frente al progreso de 8,7% en el caso de los sistemas supervisados. A la vez, esto significa la reducción drástica de la distancia entre los sistemas supervisados y los sistemas no supervisados, de 24,1% en Senseval-2 a sólo 7,2% en Senseval-3. En nuestra opinión, la evolución comentada indica que el enfoque no supervisado es una línea de investigación con un potencial todavía por explorar, mientras que el enfoque supervisado parece haber encontrado cierto tope. Significativamente, los primeros catorce sistemas supervisados en la clasificación (de los 37 participantes) ocupan un intervalo de sólo 2%.

La organización de la tarea española de inventario limitado puso una atención especial en la preparación de los recursos lingüísticos. El diccionario *Minidir 2.1*.[189] se elaboró específicamente para la tarea de DSA, lo que permitió obtener un alto grado de acuerdo entre los anotadores en el etiquetado del corpus, del 90% para nombres, del 83% para los adjetivos y del 83% para los verbos. Los resultados obtenidos en esta tarea alcanzaron el 84,2% de precisión relativa y absoluta, mucho por encima de la heurística de referencia del sentido más frecuente (67,72%).

En nuestra opinión, las dos conclusiones más importantes de Senseval-3 son, primero, los resultados obtenidos para las tareas del inglés y del español y, segundo, la evolución de estos resultados en la última edición del ejercicio respecto a la edición anterior. Así, en Senseval-3 la precisión alcanzada en el ejercicio para el español (84,2%) supera la precisión obtenida para el inglés (el 72,9% para la granularidad fina y respectivamente el 79,3% para la granularidad basta). El salto cualitativo respecto a Senseval-2 fue de 8,7% para la granularidad fina y respectivamente de 8% para la granularidad basta en la tarea inglesa supervisada, frente a un salto de 13% en la tarea española. Consideramos que estas diferencias evidencian el impacto que la calidad de las fuentes léxicas ha tenido sobre el nivel de la desambiguación en el caso del español[190].

El experimento de Màrquez *et al.* (2004a), posterior a Senseval-3, confirma estas conclusiones, demostrando que efectivamente hay un paralelismo entre, por una parte, el grado de acuerdo entre los anotadores humanos en el etiquetado de un corpus en base a una fuente léxica dada y, por otra parte, la calidad de la desambiguación obtenida por un algoritmo entrenado sobre el mismo corpus etiquetado[191]. Los autores expresan cierta reserva acerca de la causa real que determina el aumento de precisión en la desambiguación: la mejora cualitativa de la fuente léxica o la reducción de la granularidad. Sin embargo, el análisis de los resultados obtenidos, por una parte en la tarea española de Senseval-2 y de Senseval-3 (71,2% *vs.* 84,2%) y, por otra parte, en la tarea inglesa de Senseval-2 y 3

[186] Para los resultados y las conclusiones de las tareas nuevas introducidas en Senseval-3, consultar *Proceedings of Senseval-3* (2004).
[187] http://www.wordsmyth.net/
[188] *Open Mind Word Expert* es un sistema que permite coleccionar corpora anotados manualmente a través de la red (Chklovsky y Mihalcea, 2002). Se puede accesar al sito: http://teach-computers.org.
[189] Ver el capítulo 1.
[190] La tarea catalana se ha organizado por el mismo equipo y por lo tanto la metodología para la elaboración del diccionario y del corpus anotado es igual. El nivel de precisión alcanza el 85,82% (Màrquez, 2004b).
[191] En el capítulo 1, hemos remitido a otro experimento parecido en que se comparaba el grado de acuerdo en la anotación manual en base a tres fuentes léxicas.

para la granularidad fina frente a la granularidad basta (en Senseval-2, 64,2% *vs.* 71,3%, y en Senseval-3, 72,9% *vs.* 79,3%), parece indicar que el aumento cualitativo no se debería sólo a la diferencia de granularidad. En la tarea inglés, la agrupación de los sentidos en *WordNet* lleva a un aumento cualitativo de un 7-8% para los métodos supervisados. Además, el nivel así alcanzado (79,3%) queda por debajo del nivel de la tarea española (84,2%).

5.3 Otras evaluaciones y comparaciones

Fuera del ámbito de Senseval, se han desarrollado varios experimentos de carácter comparativo, orientados hacia el estudio de determinados aspectos relacionados con la tarea de DSA. En el presente apartado, sintetizamos algunas de estas investigaciones.

Gale *et al.* (1992) analizan un único algoritmo, un clasificador de tipo bayesiano, sobre doce palabras del inglés con sólo dos sentidos, desde la perspectiva de parámetros como la cantidad de datos de entrenamiento o la dimensión de la ventana contextual. Usando igualmente un único algoritmo, listas de decisión, y palabras con discriminación binaria (homónimos), Yarowsky (1993) analiza la contribución de los diferentes tipos de atributos en la desambiguación y también la relación entre la categoría de la palabra por desambiguar y la dimensión de la ventana contextual. Leacock *et al*. (1993), en cambio, comparan tres algoritmos: uno basado en el modelo del vector de contenido, un clasificador bayesiano y una red neuronal. El contraste se efectúa fijando una palabra, altamente polisémica, *line*, para diferentes dimensiones del corpus de entrenamiento. Mooney (1996) extiende el experimento de Leacock *et al*. a otros algoritmos: clasificadores bayesianos simples, clasificadores con los tres vecinos más próximos (en inglés, *3-nearest neighbors*), *Perceptron*, listas de decisión, variantes de inducción de programas lógicos. Su interés comparativo vierte sobre la dimensión del corpus de entrenamiento, el tiempo necesario para el entrenamiento y para la prueba. Experimentos más recientes estudian las curvas de aprendizaje, o sea la correlación entre la precisión y la dimensión de los datos de entrenamiento, a través de la variación de la última. Un parámetro más es investigado es el corpus de entrenamiento en cuanto a la dependencia de los resultados de este corpus; por ejemplo, Ng (1997) compara las precisiones de un mismo sistema entrenado sobre *Brown Corpus* y respectivamente sobre *Wall Street Journal*. Pedersen (2001), en un experimento más amplio, sobre los datos de Senseval-1, compara dos algoritmos, el clasificador bayesiano y el árbol de decisión y además analiza la precisión del último en relación con los atributos usados y la modalidad para seleccionarlos a partir del corpus de entrenamiento. En la comparación de diferentes algoritmos de aprendizaje supervisado, Mooney (1996) y Pedersen (2002) aseveran la superioridad de los algoritmos de tipo bayesiano por ser más efectivos en un espacio altamente dimensional, como el que trata el *DSA* (cf. Yarowsky, 2000b). Stevenson y Wilks (2001) estudian la eficiencia de diferentes tipos de conocimiento y de distintas combinaciones suyas para la resolución léxica, en función de la categoría morfosintáctica de la palabra y de la granularidad de los sentidos. Escudero *et al*. (2000) comparan dos métodos de aprendizaje automático supervisado: la clasificación bayesiana simple y el aprendizaje basado en memoria. En (Màrquez, 2002), se amplía esta comparación al algoritmo de aprendizaje incremental[192] *SNoW* (de *Sparse Network of Winnows*) y al algoritmo de *boosting* llamado *LazyBoosting*, sobre dos corpus. El propósito de la experimentación es el análisis de la dependencia de los resultados del corpus usado y la portabilidad de los sistemas de un corpus a otro. Hoste *et al*. (2002) exploran el espacio de parámetros para expertos de palabras individuales, usando algoritmos de aprendizaje basado en memoria. La gran variabilidad de los resultados según los parámetros seleccionados y a la vez la eficiencia distinta de los atributos usados en el proceso de desambiguación indican que ambas características influyen decisivamente sobre la calidad de un sistema de DSA y de aquí que toda clasificación de métodos es relativa. En varios experimentos, se han analizado los resultados obtenidos por los sistemas de DSA en relación con la categoría morfosintáctica de las palabras por desambiguar (Pedersen, 2002; Hoste *et al*., 2002; Mihalcea, 2002b), con la dificultad de

[192] En la estrategia incremental, los algoritmos aprenden los diferentes datos de manera secuencial, uno a la vez, mientras que, en la estrategia no incremental, los aprenden de manera simultánea (Schlimmer y Langley, 1990, *apud* Màrquez, 2002).

la tarea, medida a través de los conceptos de polisemia y de entropía (Resnik y Yarowsky, 2000), o con el tipo de información usado para la DSA (Pedersen, 2002).

Nos detenemos con más detalles en el experimento de Yarowsky *et al*. (2002), probablemente el más comprehensivo estudio de los enfoques supervisados desde la perspectiva del efecto que las diferentes características de los datos de entrenamiento puedan tener sobre los resultados. La evaluación se desarrolla sobre seis algoritmos supervisados: tres variantes de clasificadores bayesianos, un modelo de coseno, listas no jerárquicas de decisión y una extensión del modelo de aprendizaje basado en transformación. Los parámetros en base a los cuales se analizan los sistemas son los siguientes: lengua; categoría morfosintáctica; granularidad de los sentidos; inclusión o no de clases de atributos básicos; dimensión de la ventana contextual, donde el contexto se entiende al nivel de categorías morfosintácticas o bien de lemas; número de ejemplos de entrenamiento; probabilidad del sentido mayoritario; entropía distribucional de los sentidos; número de sentido por palabra; divergencia entre los datos de entrenamiento y de prueba; grado de ruido en los datos de entrenamiento, introducido artificialmente; efectividad de las ordenaciones de confidencia de un algoritmo; el informe detallado por palabras de los resultados de cada algoritmo, en su totalidad. La experimentación se desarrolla en un marco unitario, con un el mismo espacio de atributos, complejo, que incluye: palabra; categoría morfosintáctica; lema en un contexto amplio, de dimensiones variables, considerado como "bolsa de palabras", o bien en colocaciones locales vistas como n-gramas; asociaciones de atributos en relaciones sintácticas. El estudio consiste prácticamente en observar el impacto de la variación en este espacio de atributos sobre los resultados de los algoritmos. Para la evaluación, se usa el corpus de entrenamiento de Senseval-2, en una estrategia de evaluación transversal (en inglés, *cross-validation*)[193].

Las pruebas demuestran la sensibilidad de la calidad de la desambiguación a todos estos parámetros: las diferencias sustanciales en el espacio de atributos tienen mayor influencia sobre los resultados de los algoritmos que las variaciones en su arquitectura. En consecuencia, la prioridad más alta en el diseño de los algoritmos debe ser el enriquecimiento y la mejora del espacio de atributos usado. Si, en cambio, se mantiene fijo el espacio de atributos, los algoritmos investigados demuestran diferencias marcadas en el comportamiento empírico, induciendo a los autores a clasificarlos en dos categorías mayores: discriminativos y agregativos. Los algoritmos "agregativos" juntan toda la información disponible a favor de cada uno de los sentidos y posteriormente seleccionan el sentido con el mayor apoyo acumulado, mientras que los algoritmos "discriminativos" tienden a basarse en el mejor o los mejores, pocos atributos en el contexto para la asignación de sentido. La agrupación tiene fundamento empírico; está basada en el grado de acuerdo entre los algoritmos sobre la tarea *lexical-sample* para el inglés en el marco de Senseval-2. Se asevera la necesidad de combinar los clasificadores de acuerdo con estas diferencias, para obtener la sinergia de sus potenciales individuales, a menudo complementarios, en la tarea de DSA. La principal conclusión del estudio, resaltamos una vez más, es que la calidad que se obtiene en el proceso de DSA depende en grado significativamente superior de la selección del espacio de atributos que del algoritmo. Las observaciones de Yarowsky *et al*. son convergentes con las de otros autores, como Hoste *et al*. (2002), Mihalcea (2002b), Pedersen (2002).

[193] La estrategia *n-folder cross-validation* consiste en realizar una evaluación iterativa sobre un mismo corpus anotado con sentidos, a través de su división en *n* partes, de las cuales *n-1* se consideran como corpus de entrenamiento y una como corpus de prueba (Manning y Schütze, 1999).

II PROPUESTA:
DESAMBIGUACIÓN SEMÁNTICA AUTOMÁTICA BASADA EN LA EXPLOTACIÓN DE PATRONES LÉXICO-SINTÁCTICOS

6 Método de DSA

6.1 Espacio de análisis

En el apartado 2.5., hemos perfilado las coordenadas linguísticas prototípicas de la DSA: (i) semántica; (ii) semántica léxica; (iii) interacción del significado de las palabras con el contexto lingüístico; (iv) enfoque contextual a la polisemia; (v) variante fuerte del Principio de Composicionalidad en la construcción del significado; (vi) enumeración de sentidos como modelo del significado. Nuestra investigación en la presente tesis se desarrollará a lo largo de estas coordenadas. En el presente apartado definimos nuestra posición específica dentro del espacio teórico así delimitado para la DSA, precisando ciertas opciones teóricas que adoptaremos en nuestra propuesta para la DSA. Mantenemos, en la numeración ((iii')-(vi')) a continuación, el paralelismo con las coordenadas previamente enumeradas ((i)-(vi)).

(iii') El contexto lingüístico: sintagma u oración. Como contexto estrictamente lingüístico de una palabra por desambiguar, tenemos en vista precisamente dos niveles: el sintagma al cual pertenece la palabra y la oración en que aparece la palabra. El término *contexto* se deberá entender a lo largo de nuestra exposición como contexto lingüístico, en una de las dos formas, sintagma u oración.

(iv') Composicionalidad y consistencia en la construcción del significado. En la presente tesis, asumimos el paralelismo entre sintaxis y semántica en que se funda la semántica composicional, en concreto, la variante fuerte del Principio de Composicionalidad.

Nos mantenemos lejos de la controversia sobre la modalidad en que actúa el Principio de Interpretación Consistente: simultáneamente con la composición o bien posteriormente. Sin embargo, consideramos que ambos principios actúan a cada nivel de la estructura sintáctica, por lo tanto también dentro del sintagma. En base de las consideraciones sintetizadas en la sección 2.4., en la presente tesis asumimos que en la comunicación el flujo del significado está sometido a una doble acción, de composicionalidad y de consistencia. O sea, asumimos: 1) la combinación, dirigida por la sintaxis -a todos los niveles de estructuración sintáctica-, de los significados de las expresiones lingüísticas simples para formar el significado de las expresiones lingüísticas complejas; 2) la resolución de la ambigüedad de las expresiones simples en base a criterios de consistencia con el significado de la expresión lingüística superior que la incorpora y con los otros elementos que en ésta hay. Consideramos que en el proceso de DSA se deben reflejar ambas operaciones.

(v') Enfoque contextual a la polisemia en una visión dialéctica. Uniendo las opciones de (iv) y de (v), la perspectiva que adoptamos en la presente tesis es la dialéctica y la complementariedad entre la determinación del significado de la palabra y la determinación del significado de estructuras sintácticas como el sintagma y la oración.

(vi') (1) Visión dinámica del léxico: lexicón y corpus. Condicionados por la tipología de las fuentes léxicas existentes, también en nuestra tesis adoptamos el modelo estático del lexicón, de enumeración de sentidos.

Sin embargo, en nuestra investigación nos aproximamos a una visión dinámica del léxico, debido a que explotamos intensivamente la información proporcionada por los corpus. Tal como se expondrá a lo largo del presente capítulo, esta información es igualmente de tipo paradigmático y sintagmático. La información paradigmática se extrae del corpus partiendo con un patrón léxico-sintáctico que corresponde a un sintagma nominal (una determinada relación sintáctica), mediante una mirada

transversal en el corpus: se buscan sustitutos de la palabra por desambiguar dentro del sintagma de partida. El conjunto de sustitutos que se extraen del corpus para la palabra ambigua dentro del patrón léxico-sintáctico refleja, aunque de manera implícita, las restricciones que el otro constituyente del sintagma nominal impone sobre el nombre por desambiguar. Es decir, se hace un uso indirecto de las restricciones semánticas internas a un sintagma, en este caso sintagma nominal. En cambio, la información sintagmática se obtiene siempre del corpus, buscando ocurrencias del patrón léxico-sintáctico que incorpora el nombre por desambiguar dentro de diferentes oraciones, y extrayendo palabras frecuentemente co-ocurrentes con el patrón. De esta manera, en la parte de adquisición de información relacionada a la ocurrencia por desambiguar, nuestro método tiene un marcado carácter dinámico. La visión dinámica que hemos adoptado, manteniendo siempre como parámetro fijo el patrón léxico-sintáctico de partida, nos permite identificar lo que hay constante y por lo tanto característico para el significado del nombre por desambiguar dentro de su contexto mínimo, el sintagma nominal de partida. Si bien no seguimos el modelo generativo del lexicón propuesto por Pustejovsky (1995), sí logramos identificar, aunque de manera implícita, mecanismos composicionales internos al sintagma en que participa la palabra ambigua, al estilo de los propuestos por Pustejovsky. En conclusión, en nuestra investigación combinamos un modelo estático del léxico con información obtenida del corpus, lo que supone una visión dinámica al significado.

(vi') (2) Modelo relacional del significado. Debido a las características de la fuente adoptada - la red semántica *EuroWordNet* -, el modelo particular del significado que usamos es un *modelo relacional*[194].

6.2 Punto de partida: la intervención del conocimiento lingüístico en la DSA

Motivación del estudio del conocimiento lingüístico en la DSA. En el capítulo anterior, hemos presentado el estado de la cuestión en DSA. La dificultad de la tarea y el nivel insuficiente de los métodos disponibles, según revelan las competiciones científicas Senseval, determinan que la DSA siga siendo un problema abierto dentro del área del PLN. La falta de resultados satisfactorios en la Desambiguación Semántica Automática reclama un examen profundo de la tarea, con el objetivo no sólo de mejorar los sistemas actuales de DSA sino también de explorar vías nuevas para desarrollar el proceso en su conjunto. En esta línea, el conocimiento lingüístico puede jugar un papel importante, debido a que el interés en el área de la DSA se ha centrado principalmente en los aspectos computacionales de la tarea. En los últimos años se ha puesto de manifiesto repetidamente la necesidad de integrar más conocimiento lingüístico en los sistemas de DSA (Manning y Schütze, 1999; Corazzari *et al*., 2000, etc.). En la visión de dichos autores, que compartimos plenamente, las limitaciones actuales en la Desambiguación Semántica Automática se deben en general a una insuficiente explotación de la información lingüística. Algunos experimentos realizados recientemente ofrecen evidencia a favor de esta posición, puesto que muestran la contribución más importante de la información frente a la de los algoritmos usados en el proceso de desambiguación (Hoste *et al*., 2002; Pedersen, 2002; Yarowsky *et al*., 2002).

El objetivo general de nuestra tesis es responder a esta necesidad e identificar modalidades eficientes de incorporar más conocimiento lingüístico en la tarea de DSA.

El conocimiento lingüístico útil para la DSA: teoría y uso. El conocimiento lingüístico puede ayudar al proceso de DSA no sólo como aportación teórica, de orden general, sobre el lenguaje y las lenguas, sino también como información particular, relacionada con el uso de las palabras individuales. Así, la investigación en la DSA debe tener una visión más consistente con la teoría lingüística reciente (Ide y Véronis, 1998). A la vez, se debe explotar la visión complementaria de la lingüística del corpus, fundada en grandes cantidades de texto, que aportan datos concretos sobre las características individuales de las palabras por desambiguar. Los corpus constituyen una fuente de conocimiento lingüístico valioso para la DSA, insuficientemente explotada. Los intentos de utilizar corpus se han limitado más bien a palabras aisladas. Estimamos necesaria una visión sistemática, de carácter lingüístico, sobre la explotación de la información textual para que se redimensione su eficiencia.

[194] Ver el apartado 2.2.1.5.

Modalidades de intervención del conocimiento lingüístico en la DSA. Líneas de investigación. Nuestro intento de identificar modalidades eficientes para incorporar más conocimiento lingüístico en la Desambiguación Semántica Automática toma como puntos de referencia los tres elementos que participan en el proceso: fuentes de conocimiento léxico, texto por desambiguar al nivel de sentidos y algoritmos.

Consideramos que el conocimiento lingüístico es útil para:

a) ajustar y adecuar los elementos que participan en el proceso de Desambiguación Semántica Automática a las necesidades de la tarea;
b) investigar las relaciones entre éstos en el proceso;
c) sistematizar y guiar el proceso de Desambiguación Semántica Automática en su conjunto.

A continuación, detallamos nuestra perspectiva sobre estas tres posibilidades de intervención del conocimiento lingüístico en el proceso de DSA, que corresponden a otras tantas líneas de investigación. Para la claridad de la exposición, las estrategias mencionadas se presentan como distintas, aunque en realidad hay solapamiento entre ellas bajo múltiples aspectos.

a) La primera estrategia consiste principalmente en investigar el impacto que puede tener, sobre el nivel de la DSA, la variación en el tipo de información lingüística relacionada con las palabras y con los sentidos en la fuente léxica, en el contexto y en la heurística. Es decir, investigar la variación alternativa de uno de los tres factores manteniendo el resto invariable. Dicho estudio está orientado a identificar y potenciar, mediante el uso de conocimiento lingüístico, la contribución de cada uno de estos elementos al proceso de DSA.

a1) Respecto de la fuente de conocimiento léxico utilizada en el proceso de DSA, podemos distinguir dos tareas. Una tarea trata de enriquecer y explotar mejor y más adecuadamente las fuentes léxicas actuales, a través del etiquetado morfosintáctico, la lematización, la desambiguación semántica parcial de las unidades léxicas y la explicitación de la información léxico-semántica implícita. Otra consiste en construir fuentes léxicas más adecuadas mediante la definición de sentidos sobre unas bases lo más objetivas posibles y con información asociada realmente útil para la desambiguación.

a2) Buena parte de la labor en DSA se ha orientado hacia el texto por desambiguar al nivel de sentidos. Sin embargo, se ha investigado menos tanto la delimitación como la explotación del contexto según criterios lingüísticos. Otra vía para la explotación del contexto es añadirle información lingüística. Se trata de realizar un preproceso variado de las palabras por desambiguar y de su contexto: etiquetado morfosintáctico, lematización y agrupación (en inglés, *chunking*).

a3) El conocimiento lingüístico teórico puede contribuir en la mejora del proceso de DSA en la definición de nuevas heurísticas y en la incorporación idónea de conocimiento lingüístico específico.

b) El criterio lingüístico puede intervenir en la identificación y en la optimización de la interrelación que hay entre los elementos que participan en el proceso de DSA. Este estudio está orientado hacia la particularización del proceso según el perfil lingüístico de la ocurrencia por desambiguar, a través de la caracterización específica de los sentidos, del tratamiento individual del contexto, de las heurísticas personalizadas, etc. Un primer factor de diferenciación en este sentido es la categoría morfosintáctica de la palabra por desambiguar.

En un análisis más profundo, se trata de obtener una colaboración óptima entre los dos tipos de conocimiento relacionado con los sentidos que intervienen en la DSA: el conocimiento para caracterizar los sentidos y el conocimiento asociado a la ocurrencia ambigua. En esta línea se puede establecer una correspondencia entre los dos tipos de conocimiento mencionados o aproximarlos lo máximo posible. De esta manera se potencia la eficiencia de las heurísticas actualmente utilizadas en DSA o se pueden identificar nuevas heurísticas de calidad y fiables.

c) Una modalidad complementaria de aportar conocimiento lingüístico a la tarea de DSA es guiar el proceso en su conjunto, mediante la intervención en la implementación propiamente dicha de las

heurísticas. Entendemos que se puede proceder de diferentes maneras en la desambiguación semántica, como por ejemplo la desambiguación por clases de palabras establecidas en base de diferentes criterios lingüísticos, morfosintácticos o semánticos. Es ésta, a nuestro entender, una vía casi inexplorada en la DSA. Además, presenta interés en cuanto a la posibilidad de identificar modalidades de desambiguación con aplicabilidad más amplia, no restringida a una sola ocurrencia.

Una opción en esta dirección es delimitar criterios para establecer tipos de ambigüedad y reglas para reducir los casos nuevos a alguno de estos tipos. Así, en el proceso de DSA se aplicaría un determinado grupo de heurísticas para cada tipo de ambigüedad, es decir se usarían ciertas claves del contexto para la DSA de aquella clase de ambigüedad. Aquí, se deben estudiar cuestiones como: en qué fuentes léxicas se encuentran dichas claves; qué tipo de conocimiento es útil para la DSA de un tipo dado de ambigüedad; qué porcentaje de casos que se pueden resolver por estas heurísticas (cobertura), etc.

Foco de interés en la presente tesis. En este trabajo intentaremos abordar diferentes líneas de las mencionadas que interfieren. Sin embargo, debido al objetivo de usar intensivamente el conocimiento lingüístico en DSA, nuestra atención se dirige preponderantemente hacia la exploración de aspectos menos investigados y supuestamente con mayor impacto sobre el proceso. Así, enfocamos el uso intensivo de información implícita presente en los corpus textuales no anotados semánticamente. También analizaremos posibilidades de desambiguar palabras agrupadas según criterios de índole lingüística (estrategia c) y alcanzar así un nivel de generalidad superior y por lo tanto un desambiguador potencial en el proceso de DSA. Opinamos que esta reorganización del proceso de DSA puede tener a la vez implicaciones sobre cada uno de los elementos participantes en la tarea (estrategia a) como sobre la relación entre ellos (estrategia b).

6.3 Principios-guía en el estudio

De acuerdo con el objeto de estudio de la tesis doctoral -el conocimiento lingüístico en la DSA-, y con el propósito de explorar modalidades de incorporación intensiva de esta información en el proceso de desambiguación, nos hemos guiado en nuestra investigación según los siguientes principios:

P1. Enfoque lingüístico. Se considera que la explotación de la información lingüística para el proceso de DSA se puede hacer de manera óptima si la explotación se realiza en conformidad con la teoría lingüística.

P2. Relevancia teórica. El estudio empírico sobre el corpus y la experimentación, que la tarea de DSA supone, se consideran como una metodología orientada hacia conclusiones de orden teórico sobre los hechos lingüísticos relacionados con el establecimiento y la asignación de sentido a una palabra en el contexto.

La metodología adoptada en el estudio empírico sobre el corpus y en la experimentación, que la tarea de DSA supone, está orientada hacia conclusiones de orden teórico sobre los hechos lingüísticos relacionados con el establecimiento y la asignación de sentido a una palabra en el contexto.

P3. Minimización del uso de la estadística. Se intenta de esta manera, primero, poner de relieve el potencial de la información lingüística pura para la desambiguación léxica, y, segundo, evitar posibles desvíos en el proceso de DSA debido al uso de parámetros de tipo estadístico, como frecuencias, medidas de similitud, etc., y valores arbitrarios de estos parámetros. No se pretende negar la utilidad de la estadística para la desambiguación léxica sino simplemente investigar hasta que punto la incorporación de conocimiento lingüístico incide en la mejora de los sistemas de DSA, independientemente de la algorítmica utilizada.

P4. Simplicidad. El objetivo de este estudio es identificar un método de desambiguación semántica de fácil implementación, que no requiera un procesamiento complejo.

P5. Independencia de herramientas externas o de procesos previos de adquisición de conocimiento. En íntima vinculación con el principio precedente, este requerimiento está orientado igualmente hacia la fácil implementación, con un preprocesamiento mínimo.

P6. Portabilidad. La condición es consecuencia de los principios precedentes (principalmente P4 y P5) e impone la obtención de un método de desambiguación que sea generalizable y transferible a otros idiomas mediante la adaptación a sus sistemas lingüísticos. Para ello, se delimitarán los procesos que son de carácter general de aquellos que son dependientes de la lengua.

P7. Generalidad. El propósito de este principio es alcanzar un nivel de generalidad del método lo más alto posible, que permita una buena cobertura y la reutilización de los resultados previamente obtenidos.

P8. Reutilización de los resultados previamente obtenidos. En directa dependencia del principio anterior, éste tiene como objetivo la reducción de los cálculos necesarios para la desambiguación de casos nuevos, mediante la reutilización de los resultados obtenidos previamente. Como consecuencia, se reduce el problema de la escasez de datos y se mejora cualitativamente el proceso de DSA.

P9. Alta calidad en el proceso de desambiguación. La obtención de una alta precisión es el objetivo en la presente tesis, si es necesario en detrimento de la cobertura. La mejora cualitativa en la DSA se ve como resultado de la reutilización de los resultados previos, de la incorporación de información lingüística en el proceso o bien de la combinación de varias heurísticas altamente fiables.

P10. Explotación máxima de la información lingüística implícita en los corpus. Se trata de un conocimiento insuficientemente utilizado, según nuestra opinión, para la identificación de los sentidos. Se espera alcanzar, mediante su explotación intensiva, un nivel cualitativamente superior, una mayor objetividad y una cobertura más amplia en el proceso de desambiguación, así como superar el problema de la adquisición de conocimiento.

P11. Explotación variada de la información lingüística contenida en las fuentes léxicas estructuradas. El principio alude al uso variado de un mismo lexicón así como a la combinación de diferentes fuentes léxicas. Se espera, de esta manera, reducir el desvío del proceso de desambiguación debido a las limitaciones de las fuentes léxicas estructuradas usadas.

6.4 Enfoque a la DSA

6.4.1 Aspectos críticos de la DSA que investigamos

En nuestra investigación, hemos partido del análisis de los factores implicados en el proceso de DSA -a) fuentes de conocimiento léxico, b) texto por desambiguar al nivel de sentidos, c) algoritmos para la asignación de sentido– y sus relaciones, así como de las dificultades de los sistemas actuales de desambiguación. A continuación, resumimos los diferentes problemas expuestos en el capítulo 3, en relación con cada uno de los factores mencionados.

a) Un aspecto de mayor incidencia en la calidad de la DSA son las fuentes léxicas usadas y el concepto de sentido en ellas asumido. La experiencia acumulada en el área de la Desambiguación Semántica Automática ha puesto de manifiesto la falta de criterios claros y objetivos en la definición y delimitación de los sentidos. Las fuentes léxicas actuales son inadecuadas para sustentar este proceso, debido a la falta de un nivel de granularidad idóneo o a la discrepancia entre el tipo de conocimiento sobre los sentidos ofrecido por las fuentes léxicas y el necesario para desambiguar ocurrencias en el texto (Ide y Véronis, 1998; Ide, 2000; Palmer, 1998, etc.). En consecuencia, es necesario idear un modelo del significado y de fuente léxica adecuado para tareas del Procesamiento del Lenguaje Natural y definir los sentidos a partir de los usos de las palabras en los textos (Kilgarriff, 1997; Véronis, 2002; Agirre y Martínez, 2001a). Para el alcance del objetivo de definir los sentidos de manera objetiva y adecuada a la DSA, consideramos indispensable un examen previo profundo de la información asociada a ellos en las fuentes léxicas disponibles actualmente (Taulé *et al*., 2004). Nos

referimos aquí a varios aspectos relacionados con esta información: tipología, explicitación, impacto sobre el proceso de DSA, uso óptimo para la desambiguación, etc. Vemos una posibilidad más de superar las limitaciones intrínsecas, previamente mencionadas, en el uso variado de las fuentes léxicas existentes.

b) De las dos categorías básicas de contexto que se suelen diferenciar en la DSA -local y temático-, en la presente tesis la atención se focaliza en el contexto local, en su delimitación y tratamiento. El contexto local sigue siendo objeto de estudio y experimentación dentro de la DSA, ya que no se ha identificado un contexto local idóneo y su uso óptimo para la DSA.

El contexto local se ha definido frecuentemente como una ventana de dimensión predefinida, centrada en la ocurrencia por desambiguar. Los experimentos han demostrado que los nombres y los verbos se comportan de manera distinta en relación con el tamaño del contexto local (cf. Ide y Véronis, 1998), como también las diferentes palabras de una misma categoría morfosintáctica (Mihalcea, 2002; Yarowsky y Florian, 2002; Pedersen, 2002). Por lo tanto, el contexto local de dimensión predefinida para toda ocurrencia ambigua parece no satisfacer los requisitos de una DSA de alta calidad.

En cuanto al tratamiento del contexto, los diferentes experimentos que han explotado las palabras funcionales contiguas a la ocurrencia ambigua demuestran su utilidad para la asignación de sentido (Yarowsky y Florian, 2002; Mihalcea, 2002; Hoste *et al*., 2002, etc.), y por lo tanto la ventaja del enfoque relacional –n-gramas o funciones sintácticas– frente al enfoque "bolsa de palabras". Sin embargo, hasta el momento, el uso de las palabras de contenido gramatical en el proceso de DSA se ha realizado especialmente en el enfoque basado en ejemplos, con lo cual ha estado vinculado y dependiente de un corpus etiquetado a nivel de sentido. En el caso de que se quieran explotar las relaciones y funciones sintácticas, el corpus de entrenamiento debe ser anotado también a nivel sintáctico. Hasta el presente, la información sintáctica usada para la DSA se ha limitado por lo general a las relaciones verbo-sujeto y verbo-objeto (Ng, 1996; Leacock *et al.*, 1998; Federici *et al*., 2000; Agirre y Martínez, 2001b, Martínez *et al*., 2002, etc.), con pocas excepciones (Hearst, 1991; Yarowsky, 1993; Lin, 1997; Stetina *et al*., 1998).

Falta, en consecuencia, un estudio sistemático de la contribución de los diferentes tipos de información para cada uno de los diferentes tipos de palabras por desambiguar (Ide y Véronis, 1998). Como alternativa, desde la perspectiva computacional, además de la solución ya clásica de la combinación de algoritmos, últimamente se intenta usar varios parámetros relacionados con el contexto y, por otra parte, algoritmos para identificar el parámetro más informativo respecto del sentido (Mihalcea 2002; Yarowsky y Florian, 2002). Sin embargo, se ha investigado menos tanto la delimitación como la explotación del contexto con criterios lingüísticos.

c) Ambos enfoques de la DSA -el basado en conocimiento (en inglés, *knowledge-driven WSD*), y el basado en corpus (en inglés, *corpus-based WSD*)- tienen sus limitaciones.

Por una parte, los sistemas de DSA basados en conocimiento usan la información, principalmente paradigmática, contenida en las fuentes léxicas estructuradas, mientras que en el contexto de la ocurrencia ambigua se halla sobre todo información sintagmática. En consecuencia, estos sistemas se ven afectados por el problema del "vacío" que hay entre el lexicón y el corpus (Kilgarriff, 1998). La solución comúnmente adoptada es el acercamiento de las fuentes léxicas estructuradas a los corpus mediante la incorporación de información sintagmática (Véronis, 2002). El enfoque está limitado por el coste elevado del proceso de extracción y de representación de tal conocimiento.

Por otra parte, los sistemas de DSA basados en corpus, que sí utilizan información sintagmática, obtuvieron mejores resultados en las competiciones Senseval. No obstante, son altamente dependientes de la disponibilidad de ejemplos etiquetados con sentidos, ya que hay una proporcionalidad directa entre la cantidad de tales ejemplos y el nivel de los resultados. La necesidad de grandes cantidades de casos anotados con sentidos se debe al problema de la dispersión de datos: el contexto local se repite poco y por lo tanto la información contextual aprendida sobre el corpus etiquetado es raramente utilizable para casos nuevos (Leacock *et al*. 1998; Ide y Véronis, 1998; Manning y Schütze, 1999). La obtención de textos etiquetados semánticamente a mano es difícil y costosa, por lo que se ha propuesto, como alternativa, su obtención de manera automática (Gale *et al*., 1992; Yarowsky, 1992, 1995; Leacock *et al*., 1998; Mihalcea y Moldovan, 1999, etc.). Las pocas evaluaciones de los corpus

anotados a nivel de sentido no son, sin embargo, muy positivas (por ejemplo, Agirre y Martinez, 2000).

En síntesis, los puntos críticos en la tarea de DSA son los siguientes:

- la caracterización de los sentidos;
- la distancia entre el lexicón y el corpus;
- la dificultad de la adquisición de conocimiento (en inglés, *knowledge acquisition bottleneck*);
- la escasez de datos (en inglés, *data sparseness*) y
- el contexto local.

En la presente tesis, nos centramos en el estudio de cada una de estas cuestiones. A continuación, exponemos nuestras propuestas para solucionar los problemas mencionados.

6.4.2 Análisis y soluciones propuestas

En esta sección, analizamos las limitaciones del proceso de DSA señaladas en el apartado previo, siendo estas limitaciones estrechamente relacionadas entre sí. El objetivo del análisis es la identificación de las causas de estas dificultades y su eliminación. Debido a que estas causas corresponden a operaciones u opciones internas al proceso de DSA, cuestionamos ciertas características del enfoque actual de la DSA y exploramos alternativas para el desarrollo de la tarea.

6.4.2.1 Caracterización de la ocurrencia ambigua con información paradigmática extraída del corpus

Las dos categorías de sistemas de DSA (basados en conocimiento y basados en corpus) se enfrentan esencialmente con el mismo problema: la falta de información sintagmática relacionada con los sentidos de las palabras. Esta dependencia se debe a que actualmente los métodos de DSA se limitan a extraer información sintagmática relacionada con la ocurrencia ambigua, y en base a esta información proceden a la identificación del sentido. Además, la falta de información de tipo sintagmático asociada a los sentidos en el lexicón dificulta y limita la operación de asignación de sentido.

Si la información sintagmática relacionada con los sentidos de las palabras es difícil de obtener, nos hemos planteado operar con información paradigmática en el proceso de DSA. Es decir, para una ocurrencia por desambiguar fijada, identificar información paradigmática relacionada con ella; esta información sí que es más fácil de proyectar en las fuentes léxicas, en las cuales predomina la información de tipo paradigmático.

6.4.2.2 Adquisición de conocimiento mediante el uso de relaciones paradigmáticas y sintagmáticas

Respecto al problema de la adquisición de conocimiento léxico necesario para la DSA, en la presente tesis proponemos la explotación automática de corpus no anotados, como alternativa a la dependencia actual de la intervención humana en este proceso. Contrariamente a la solución habitual de incorporar información sintagmática en las fuentes léxicas para llenar el vacío lexicón-corpus, planteamos la posibilidad inversa, de acercar el corpus al lexicón. Consideramos que en los corpus hay información implícita, aprovechable para la DSA. Se puede hacer un uso cualitativo de los corpus, para la extracción de información, y un uso estadístico, para filtrar los datos obtenidos. Basándonos en las consideraciones de la sección anterior y motivados por la dificultad de obtener información sintagmática relacionada con los sentidos, investigamos la extracción de información de tipo paradigmático de los corpus, asociada a una ocurrencia por desambiguar.

Esto nos lleva a otra cuestión de la DSA. La obtención de información paradigmática para una ocurrencia ambigua no es posible usando exclusivamente su contexto. Las relaciones paradigmáticas se establecen con elementos virtuales, que podrían sustituir a una palabra dada en el mismo contexto oracional. Por lo tanto, estos elementos se hallan usualmente en otros fragmentos textuales; su identificación requiere una mirada transversal en los corpus[195], lo que excede los límites de la DSA en su forma actual.

[195] Se hace una búsqueda en diferentes oraciones del corpus para hallar ocurrencias de otras palabras en un

En consecuencia, cuestionamos el desarrollo -dominante actualmente- del proceso de DSA ocurrencia por ocurrencia y palabra por palabra, en que se resuelven los casos de ambigüedad de manera independiente uno del otro. Consideramos que la desambiguación debe ampliar su visión caso por caso a una perspectiva más amplia: grupos de ocurrencias o grupos de palabras[196]. Nos acercamos, en este aspecto, a los métodos de DSA basados en clases o en similitud (cf. Ide y Véronis, 1998).

Creemos que la información implícita en los corpus es explotable para la DSA mediante diferentes agrupaciones de palabras. En el presente trabajo, para establecer la base de agrupación, hemos partido de una propiedad fundamental del lenguaje natural: la interacción entre los ejes sintagmático y paradigmático. Las palabras que se suceden en la cadena comunicativa, oral o escrita, se sitúan en el eje sintagmático, estableciendo relaciones que aseguran la coherencia del acto comunicativo. A la vez, un elemento fijado en un punto de la cadena sintagmática puede ser sustituido por otras palabras, obteniéndose enunciados igualmente coherentes. Estos elementos virtuales, que pueden sustituir un elemento dado del eje sintagmático, junto con el mismo elemento sustituible pertenecen a un eje paradigmático, y entre ellos establecen relaciones paradigmáticas. De esta manera, condiciones sintagmáticas idénticas delimitan conjuntos de tipo paradigmático de palabras: partimos de la hipótesis de que las diferentes palabras que pueden aparecer en una determinada posición de una secuencia sintagmática fijada tendrán sentidos relacionados, pertenecientes a una o más zonas conceptuales comunes[197].

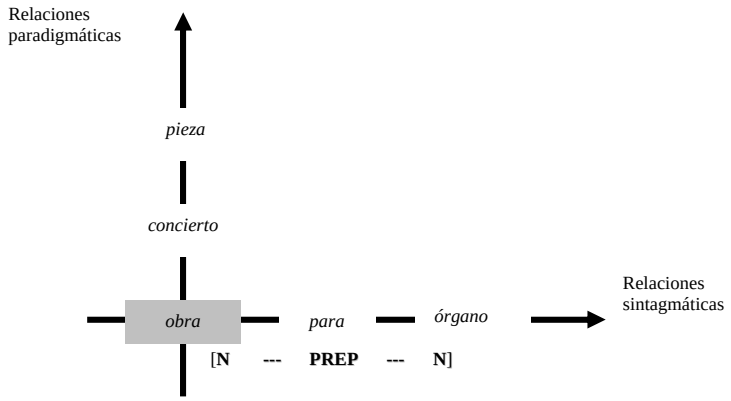

Figura 6.1. Interacción entre los ejes sintagmático y paradigmático

Ilustramos esta propiedad del lenguaje natural con un ejemplo (figura 6.1.). Si en el eje sintagmático se fija la secuencia *obra para órgano* y se deja variable sólo el punto correspondiente a la posición de *obra*, se pueden identificar otras palabras como sustitutos suyos: *concierto, pieza*, etc. Se observa que las palabras que pueden intercambiarse (*obra, concierto, pieza*), o sea que pertenecen al eje paradigmático, están relacionadas semánticamente y ocupan una zona conceptual común, en este caso, 'instrumentos musicales'.

Los experimentos de Miller y Charles (1991) confirman que los humanos determinan la similitud semántica de las palabras en base de la similitud de los contextos en que se usan.

6.4.2.3 Caracterización alternativa de los sentidos. Adaptación de *EuroWordNet* con discriminadores de sentido

Para la asignación de sentidos, se debe establecer una correspondencia entre la información paradigmática, asociada a una ocurrencia ambigua, que se extrae del corpus, y la información

---

contexto similar al contexto que tiene la palabra ambigua dentro de la oración de partida en la que se tiene que desambiguar.

[196] Complementaria a este enfoque es la visión de la DSA (en la tarea para todas las palabras, *all-words*) como un proceso desarrollado en oleadas: la desambiguación de los nombres, seguida por la desambiguación de los verbos, etc. La comprobación del impacto de esta estrategia sobre la calidad de la asignación de sentidos supera los límites de este trabajo.

[197] Cf. el apartado 2.3.

paradigmática de la fuente léxica que caracteriza los sentidos de la palabra. Cuanto más amplia sea la información paradigmática en la fuente léxica, más probabilidad hay de que se realice esta proyección. Con lo cual, es necesario usar una fuente léxica con información paradigmática rica. Frente a otros lexicones, *EuroWordNet* (Vossen, 1998) contiene una información amplia y variada (diferentes tipos de relaciones léxico-semánticas), bien formalizada y estructurada, y por lo tanto es idónea para la DSA que opera con información paradigmática.

Para potenciar la información paradigmática contenida en *EWN* sobre los sentidos, hemos desarrollado una adaptación de la componente española de *EuroWordNet*, que exponemos a continuación[198]. En esta adaptación, cada uno de los sentidos de una palabra polisémica se caracteriza mediante el conjunto de *variants*[199] de los *synsets* con los que está en relación, y que no comparte con ningún otro sentido de la palabra. Así, para cada sentido X_i de una palabra X, se extrae de *EWN* el conjunto de *synsets* con que se relaciona. Para ello, a partir del nodo en que se halla X_i, se atraviesa la red a lo largo de los arcos correspondientes a un mismo tipo fijado de relaciones léxico-semánticas. Posteriormente, se repite la operación para cada uno de los tipos de relaciones léxico-semánticas contenidas en *EWN*: hipo-hiperonimia, mero-holonima, co-hiponimia. De los *synsets* así obtenidos para cada sentido X_i, se extraen los *variants* que contienen, incluidos los *variants* del mismo nodo que contiene a X_i. Así, se recuperan también los *variants* que contraen una relación de sinonimia con X_i. De esta manera, a los sentidos X_i de X se les asocian respectivamente los conjuntos V_i de *variants*.

Se eliminan luego los *variants* que aparecen en más de un conjunto V_i, de modo que los conjuntos obtenidos D_i son disjuntos y, para un *i* fijado, los elementos del conjunto D_i están relacionados en *EWN* exclusivamente con el sentido X_i. Desde la perspectiva de *EWN*, las palabras de los conjuntos D_i son específicas para cada uno de los sentidos correspondientes X_i y se convierten por lo tanto en discriminadores de sentido[200]. Por esta razón, hemos llamado la fuente de conocimiento léxico derivada de *EWN*, a través del procedimiento descrito, *Discriminadores de Sentido*.

Ilustramos la extracción de los discriminadores de sentido para el nombre *órgano*, que tiene cinco sentidos en *EWN 1.5*[201]:

*órgano*_1 (n07977350): 'parte de una planta';
*órgano*_2 (n05302115): 'agencia gubernamental, instrumento';
*órgano*_3 (n03650737): 'parte funcional de un animal'
*órgano*_4 (n02831270): 'instrumento musical'
*órgano*_5 (n04311573): 'periódico'

Así, partiendo del *synset* n07977350, correspondiente al sentido *órgano*_1, extraemos los *variants* de los *synsets* que se encuentran con este *synset* en una de las diferentes relaciones léxico-semánticas:
a) sinonimia: {*órgano vegetal*};
b) hiperonimia: {*objeto inanimado, objeto físico, objeto, cosa, entidad*};
c) hiponimia: {*estructura reproductiva, lámina, raíz, tronco, troncho, tallo, rabillo, pedúnculo, cálamo, caña, cabillo, hoja, follaje, retoño, …*};
d) holonimia: ∅;
e) meronimia: ∅;
f) coordinación o cohiponimia: {*talo, sombrero, sombrerito, sombrerillo, sombrerete, píleo, carpóforo, volva, chalaza, micelio, túbulo, tubo, poro, velo universal, velo parcial, ánulo, anillo, …*}.

Los nombres así obtenidos se juntan en un conjunto asociado a *órgano*_1:

[198] Esta adaptación nos ha permitido diseñar un nuevo algoritmo de DSA, la Prueba de Conmutabilidad (apartado 6.7.3.).
[199] Los *variants* son las palabras que forman un *synset*. Por ejemplo, el *synset* número 03984375 de *EWN* es: {*ensoñación*_3 *visión*_6 *quimera*_3 *imaginación*_2 *ilusión*_2 *fantasía*_5 *ensueño*_3}. Los nombres *ensoñación, visión, quimera, imaginación, ilusión, fantasía, ensueño* son los *variants* de este *synset*.
[200] El proceso de extracción de esta información para caracterizar los sentidos es completamente automático.
[201] Las pseudodefiniciones son nuestras, creadas para la presente exposición. Se han derivado en base del hiperónimo y de la glosa inglesa correspondientes al *synset* de cada sentido.

V_1 = {*órgano vegetal, objeto inanimado, objeto físico, objeto, cosa, entidad, talo, sombrero, sombrerito, sombrerillo, sombrerete, píleo, carpóforo, volva, chalaza, micelio, túbulo, tubo, poro, velo universal, velo parcial, ánulo, anillo, …*}.

El proceso se repite para todos los sentidos de *órgano*, obteniéndose para cada uno el conjunto correspondiente de tipo V. A continuación mostramos el proceso para todos los sentidos.

Para *órgano*_2, partiendo del *synset* n05302115, obtenemos primero los conjuntos:

a) sinonimia: ∅;

b) hiperonimia: {*oficina, agencia, unidad administrativa, unidad, organización, organización, grupo, colectivo, agrupación*};

c) hiponimia: ∅;

d) holonimia: ∅;

e) meronimia: ∅;

f) coordinación o cohiponimia: {*agencia ejecutiva, oficina de la seguridad social, seguridad social, F.B.I., FBI, oficina de aduanas, aduanas, banco central, Banco Central, servicios secretos, servicio secreto, agencia meteorológica, servicio meteorológico, …*}.

Luego reunimos estos conjuntos en el conjunto asociado a *órgano*_2:

V_2 = {*oficina, agencia, unidad administrativa, unidad, organización, organización, grupo, colectivo, agrupación, agencia ejecutiva, oficina de la seguridad social, seguridad social, F.B.I., FBI, oficina de aduanas, aduanas, banco central, Banco Central, servicios secretos, servicio secreto, agencia meteorológica, servicio meteorológico, …*}.

El sentido *órgano*_3 pertenece al *synset* n03650737, a partir del cual obtenemos los conjuntos:

a) sinonimia: ∅;

b) hiperonimia: {*parte del cuerpo, trozo, porción, parte, entidad*};

c) hiponimia: {*oviscapto, patas, ventosa, órgano contráctil, órgano vital, órgano efector, órgano externo, víscera, órgano sensorial, receptor, lengua, órgano del habla, cristalino, órgano segregatorio, órgano secretorio, órganos sexuales, órganos reproductores, …*};

d) holonimia: ∅;

e) meronimia: {*lóbulo*};

f) coordinación o cohiponimia: {*grupa, ambulacro, lomo, ijada, tórax, área, zona, región, parte externa del cuerpo, estructura anatómica, sistema, tejido, hombro, pierna, …*}.

La reunión de estos conjuntos da el conjunto asociado a *órgano*_3:

V_3 = {*parte del cuerpo, trozo, porción, parte, entidad, lóbulo, oviscapto, patas, ventosa, órgano contráctil, órgano vital, órgano efector, órgano externo, víscera, órgano sensorial, receptor, lengua, órgano del habla, cristalino, órgano segregatorio, órganos sexuales, órganos reproductores, …, grupa, ambulacro, lomo, ijada, tórax, área, zona, región, parte externa del cuerpo, estructura anatómica, sistema, tejido, hombro, pierna,…*}.

Las relaciones léxico-semánticas que parten del *synset* n02831270 (en el cual se halla el sentido *órgano*_4) nos permiten la extracción de los siguientes conjuntos:

a) sinonimia: ∅;

b) hiperonimia: {*instrumento de viento, instrumento musical, instrumento, instrumento, mecanismo, dispositivo, aparato, utillaje, artefacto, objeto inanimado, objeto físico, objeto, cosa, entidad*};

c) hiponimia: ∅;

d) holonimia: ∅;

e) meronimia: {*teclado, pedal, campana, embocadura*};

f) coordinación o cohiponimia: {*gaita, cornamusa, metal, corneta, kazoo, ocarina, zampoña*}.

La reunión de estos conjuntos es el conjunto asociado a *órgano*_4:

V_4 = {*instrumento de viento, instrumento musical, instrumento, instrumento, mecanismo, dispositivo, aparato, utillaje, artefacto, objeto inanimado, objeto físico, objeto, cosa, entidad, teclado, pedal, campana, embocadura, gaita, cornamusa, metal, corneta, kazoo, ocarina, zampoña, …*}.

Finalmente, para el sentido *órgano*_5, se parte del *synset* correspondiente, n04311573, extrayendo los siguientes conjuntos:

a) sinonimia: ∅;

b) hiperonimia: {*publicación periódica, periódico, publicación, medio de comunicación escrita, medio de comunicación, método, medio, manera, utillaje, artefacto, objeto inanimado, objeto físico, objeto, cosa, entidad, obra, producto, producción, creación, artefacto, objeto inanimado, objeto físico, objeto, cosa, entidad*};

c) hiponimia: ∅;

d) holonimia: ∅;

e) meronimia: ∅;

f) coordinación o cohiponimia: {*serie, serial, número, ejemplar*}.

La reunión de estos conjuntos es el conjunto asociado a *órgano*_5:

V_5 = {*publicación periódica, periódico, publicación, medio de comunicación escrita, medio de comunicación, método, medio, manera, utillaje, artefacto, objeto inanimado, objeto físico, objeto, cosa, entidad, obra, producto, producción, creación, artefacto, objeto inanimado, objeto físico, objeto, cosa, entidad, serie, serial, número, ejemplar, …*}[202].

Una vez se han obtenido los conjuntos V_i, se eliminan los nombres comunes en los conjuntos V_i, con i = 1,…,5. Se establecen así los conjuntos de discriminadores D_i siguientes:

D_1: {*órgano vegetal_1, lámina_3, raíz_2, tronco_4, tallo_1, hoja_3, …*}
D_2: {*oficina_2, agencia_2, organización_1, colectivo_1, …*}
D_3: {*parte del cuerpo_1, trozo_8, parte_9, lóbulo_2, lengua_3, ojo_4, …*}
D_4: {*instrumento de viento_1, instrumento musical_1, aparato_3, teclado_1, …*}
D_5: {*publicación_2, periódico_4, medio de comunicación_1, obra_5, …*}.

A diferencia de *EWN* en su forma estándar, en esta nueva variante los sentidos se caracterizan de manera plana, mediante unas palabras asociadas, y no por su posición en una jerarquía. Además, si se consideran los conjuntos estrictos de discriminadores de sentido, se pierde la conexión entre los sentidos.

Se ha desarrollado también una variante ampliada de los Discriminadores de Sentido. Con este propósito, dada una palabra y los conjuntos de discriminadores correspondientes que representan a sus sentidos, para cada uno de los discriminadores de un sentido dado, hemos extraído sus hipónimos heredados y los hemos incorporado al conjunto de discriminadores al que pertenece el discriminador de partida. Obtenemos así conjuntos extendidos de potenciales discriminadores, D'_i, asociados a los sentidos. Para asegurar la calidad de discriminadores de sentido de los elementorimin 1.529D /F0 9.96os incorporal

ambigua en su contexto lleva a su desambiguación. En esta integración opera una delimitación sobre el significado potencial de la palabra. Cuanto más amplia sea la integración de la palabra en el contexto más restricciones habrá sobre su significado.

A continuación, recogemos una serie de resultados de orden teórico y empírico que nos dan fundamento en nuestra decisión sobre la delimitación y el tratamiento del contexto. Ante todo, asumimos las consideraciones de la semántica teórica expuestas en el apartado 2.4. sobre la construcción del significado, en particular la interdependencia entre el significado de una palabra y el significado de las estructuras sintácticas superiores que la contienen: sintagma y oración. Esta perspectiva desde la semántica léxica y composicional es convergente con la evidencia empírica obtenida desde la psicolingüística y desde la DSA misma. A continuación, mencionamos algunas aportaciones relevantes.

Los experimentos con humanos (Miller y Charles, 1991) han demostrado que una ventana de pocas palabras alrededor de la ocurrencia ambigua es suficiente para su desambiguación.

Dentro de la misma área de la DSA, algunos experimentos confirman que el uso del contexto local reducido permite obtener un buen nivel de asignación del sentido (cf. Ide y Véronis, 1998). Por lo tanto, la integración de la palabra ambigua en este contexto debería ser una buena aproximación de su sentido. Uno de los experimentos más relevantes es el de Yarowsky (1993), en base al cual se formula la hipótesis de un sentido por colocación (en inglés, "*one sense per collocation*").

Esto nos lleva a considerar la integración de una palabra ambigua en su contexto como una operación gradual: se parte con un contexto mínimo, como primera aproximación a la identificación del sentido de la ocurrencia ambigua, y se amplia posteriormente, hasta obtener una única asignación de sentido.

Creemos que nuestra opción por un contexto local mínimo es oportuna también en relación con el problema de la escasez de datos. En la extracción de información paradigmática relacionada con una ocurrencia ambigua dada, se parte con una secuencia sintagmática que contiene la ocurrencia, se mantienen fijos los demás elementos y se deja variable la posición de la ocurrencia en la secuencia, buscando en el corpus palabras que la ocupan. Una secuencia sintagmática reducida que se tome como punto de partida en esta operación aumentará la probabilidad de que los elementos fijos se repitan y por lo tanto que dentro del corpus se encuentren más palabras en la posición variable, correspondiente a la ocurrencia ambigua.

Pasando al tratamiento del contexto local, en el enfoque "bolsa de palabras" se pierde información. Las palabras funcionales son altamente útiles, ya que estructuran el enunciado y combinan el significado de las palabras individuales de contenido léxico. Por lo tanto, tomamos en consideración las palabras funcionales dentro de nuestro enfoque a la DSA. De esta manera, aprovechamos la información sobre la combinación de las palabras de contenido léxico y su significado, especificada por las palabras funcionales.

Respecto a las palabras de contenido léxico en el contexto, las relaciones sintagmáticas más estrechas son de tipo sintáctico, léxico-semántico o de simple adyacencia, sobre todo cuando ésta última es estable. Nuestra hipótesis es que, por razones que tienen que ver con la comunicación[203], estas relaciones se establecen preponderantemente con las palabras inmediatamente contiguas. Cuando se realizan a distancia en la oración (o incluso entre dos oraciones), se reconocen debido a que previamente se han encontrado realizadas con frecuencia de manera contigua.

Así, en el ejemplo *Cuando almacenamos un archivo, la información va ocupando de una forma ordenada el espacio libre disponible* (CREA), la relación de tipo verbo-objeto entre *ocupar* y *el espacio* se realiza de manera discontinua, pero se reconoce porque se usa con cierta frecuencia de forma seguida. Las secuencias *ocupar el espacio* (con *ocupar* conjugado en el presente del indicativo) aparecen 28 veces en el corpus CREA.

Para el tratamiento formal del contexto inmediato de una ocurrencia ambigua y la identificación de sus relaciones sintáctico-semánticas allí expresadas, introducimos el término de *patrón sintagmático*.

[203] Los principios de cooperación de Grice (1975), las limitaciones de memoria y de computación de los humanos, etc.

Definimos formalmente un *patrón sintagmático* como una tripleta que corresponde a una relación sintáctico-semántica, formada por dos unidades *L1* y *L2* de contenido léxico (nombres, adjetivos, verbos o adverbios) y un patrón léxico-sintáctico *R* que expresa la relación (de dependencia, de coordinación, léxico-semántica o de adyacencia) que contraen las dos unidades léxicas:

$$L_1 - R - L_2$$

Incluimos en este patrón general el caso en que *R* es nulo, como en la relación de adyacencia o entre un nombre y un adjetivo. Ejemplos:

[*grano*-N *de*-PREP *azúcar*-N]
[*pasaje*-N *subterráneo*-ADJ].

Nuestra hipótesis es que el sentido de una palabra en un contexto será principalmente determinado por sus patrones sintagmáticos locales. Desde esta perspectiva, una palabra polisémica, y cada uno de sus sentidos, se podrán caracterizar mediante los patrones sintagmáticos en que participan.

Ofrecen cierto respaldo a nuestro enfoque también las hipótesis que formulan Miller y Charles (1991) partiendo de los experimentos ya mencionados. Partiendo de la constatación que los humanos determinan la similitud semántica de las palabras a base de la similitud de los contextos en que se usan, los autores formulan dos hipótesis que explican este proceso:

> a) hipótesis contextual fuerte: dos palabras son semánticamente similares en la medida en que sus representaciones contextuales son similares, es decir la similitud semántica está determinada por el grado de similitud entre los conjuntos de contextos en los cuales las palabras se pueden usar;
>
> b) hipótesis contextual para los sentidos: dos ocurrencias de una palabra ambigua pertenecen al mismo sentido en la medida que sus representaciones contextuales son similares; es decir los sentidos se basan en la similitud contextual: un sentido es un grupo de ocurrencias (en inglés, *tokens*) de la palabra en cuestión con contextos similares, es decir un grupo de ocurrencias de una palabra similares contextualmente.

En nuestro enfoque, basado en la explotación de patrones sintagmáticos, asumimos unas variantes débiles de estas *hipótesis*:

> *a¢) dos palabras que comparten un contexto local similar (aquí un mismo patrón sintagmático) tienen una probabilidad elevada de ser próximas semánticamente;*
>
> *b¢) dos ocurrencias de una palabra ambigua corresponden con una probabilidad elevada al mismo sentido si aparecen en un contexto similar (aquí un mismo patrón sintagmático) según la hipótesis de la "tendencia hacia un único sentido por patrón sintagmático" (en inglés, "towards one sense per syntagmatic pattern").*

En relación con los patrones sintagmáticos, mostramos una distinción básica entre:

> 1) patrones sintagmáticos que corresponden a relaciones sintácticas, y que llamamos *patrones léxico-sintácticos* (por ejemplo, *corona de santo*), y
> 2) patrones sintagmáticos que corresponden a relaciones léxico-semánticas, y que de manera simétrica denominamos *patrones léxico-semánticos* (por ejemplo, los patrones correspondientes a la relación de meronimia, como *los miembros del comité*).

Vemos los dos tipos de patrones sintagmáticos complementarios y altamente relevantes para la identificación de los sentidos de las palabras en un texto y por lo tanto los sistemas de DSA deberían tratar ambos tipos. Las dos clases de patrones sintagmáticos tienen un tratamiento diferente respecto a su identificación como a su utilización para la DSA. Así, en el caso de los patrones léxico-sintácticos,

el componente relacional *R* se expresa mediante palabras funcionales, mientras que en el caso de los patrones léxico-semánticos, *R* suele tener una forma más compleja, y puede incorporar tanto palabras funcionales como de contenido léxico. Por otra parte, las dos clases de patrones sintagmáticos no son disjuntas sino que hay solapamiento entre ellas: así, el ejemplo anterior, *los miembros del comité*, expresa a la vez una relación sintáctica de dependencia y una relación léxico-semántica de meronimia entre los nombres *miembros* y *comité* a través de la preposición *de* y del artículo determinativo *el*.

En la presente tesis, nos limitamos a estudiar los patrones léxico-sintácticos. Los patrones léxico-semánticos serán objeto de un análisis futuro[204], y aquí se tratan tangencialmente, en la medida en que un mismo patrón es a la vez léxico-sintáctico y léxico-semántico. Sin embargo, parte de las consideraciones del estudio son válidas igualmente en el caso de los patrones léxico-semánticos, debido al solapamiento mencionado.

Las relaciones sintácticas de una palabra se pueden realizar en su contigüidad inmediata o bien a distancia. Uno de los factores que influye en esto es la categoría morfosintáctica de la palabra. Así, los nombres y los adjetivos suelenpalab31ss12 d91en reaveciw (comi2cos,) Tj -12..72 TD7 -12828 Tc n conr una f0.22n

paradigmática proporcionada por la fuente léxica (aquí *EWN*) y la información sintagmática identificable en el contexto de la ocurrencia por desambiguar. En base a los patrones léxico-sintácticos se realiza la transición del eje sintagmático al eje paradigmático: se identifica en un corpus el conjunto de palabras que pueden ocupar la posición de la ocurrencia ambigua en el patrón léxico-sintáctico, lo que define una clase de tipo paradigmático. Sobre esta clase se aplica un algoritmo de desambiguación basado en conocimiento, en concreto con las relaciones paradigmáticas de *EWN*. El algoritmo identificará relaciones léxico-semánticas de *EWN* entre las palabras del conjunto paradigmático, lo que significa su localización en la red y, de este modo, su desambiguación.

El proceso de desambiguación se basa en la colaboración entre fuentes estructuradas de conocimiento léxico y corpus, por lo tanto se trata de un enfoque mixto a la desambiguación semántica respecto de las fuentes léxicas usadas. Por otra parte, debido a que no se necesitan ejemplos etiquetados al nivel de sentido, nuestra estrategia es no supervisada[205].

Con respecto a la bibliografía, nuestra propuesta se acerca más a las de Montemagni *et al.* (1996) y de Federici *et al*. (2000), definida como "enfoque por paradigmas" (en inglés, *paradigm-driven approach*) a la DSA, y de Agirre y Martínez (2001b). En estos métodos se combinan variantes paradigmáticas para las dos posiciones léxicas de lo que aquí llamamos "patrón léxico-sintáctico". Sin embargo, la combinación se realiza sobre patrones que corresponden a relaciones verbo-argumento, etiquetados previamente al nivel de sentido y analizados sintácticamente.

Creemos que la explotación de los patrones léxico-sintácticos de la manera que hemos propuesto previamente abre una línea de investigación, explorable bajo múltiples aspectos. Nuestro principal interés vierte sobre la DSA, y por lo tanto aquí se focalizará sobre todo la relación de los patrones léxico-sintácticos de base sintáctica con la identificación de sentidos.

6.5 Método básico

Según este enfoque, para la asignación de un sentido a una ocurrencia del nombre polisémico X se deben seguir los siguientes pasos:

Paso 1º. Se identifican los patrones léxico-sintácticos P_k en que la ocurrencia ambigua de X aparece en la oración.

La delimitación de los patrones se realiza a nivel de categorías morfosintácticas y de lemas. Por lo tanto, el enunciado en que se halla la ocurrencia por desambiguar se etiqueta previamente con la ayuda de un etiquetador morfológico. Para cada palabra del texto de entrada, el etiquetador devuelve una tripleta que contiene la forma flexiva, el lema y la categoría morfosintáctica.

En base de las propiedades distribucionales y sintácticas de los nombres, hemos predefinido una lista de posibles combinaciones de categorías morfosintácticas en una relación sintáctica, de coordinación o de dependencia, en que intervienen los nombres[206]. Recogemos estos tipos básicos de patrones morfosintácticos en la tabla 6.1:

Patrones básicos	
[**N_1** , N_2]	[N_1 , **N_2**]
[**N_1** CONJ* N_2]	[N_1 CONJ* **N_2**]
[**N** ADJ]	[ADJ **N**]
[**N** VPART]	[VPART **N**]
[N_1 PREP N_2]	[N_1 PREP **N_2**

Tabla 6.1. Patrones básicos

[205] Cf. el apartado 4.1.3.

[206] No hemos considerado en el presente estudio las relaciones –argumentales o sintácticas- entre nombres y verbos. De los verbos, hemos considerado sólo a los participios con valor adjetival. La reducción se debe a que las relaciones entre nombres y verbos se realizan a menudo a distancia y suponen así un tratamiento específico, mientras nuestra investigación se limita a los patrones léxico-sintácticos locales contiguos.

donde CONJ* es una conjunción coordinativa. Hemos marcado en negrita la posición de la palabra por desambiguar.

Una vez hemos encontrado secuencias que corresponden a los patrones morfosintácticos, los consideramos también a nivel de lema, y es con estos patrones de lemas y de categorías morfosintácticas con los que seguimos en la implementación de la estrategia. Por ejemplo, si hemos encontrado la secuencia [*canales-canal*-N *de-de*-PREP *televisión-televisión*-N], consideramos el patrón como [*canal*-N *de*-PREP *televisión*-N].

Paso 2º. Para cada uno de los patrones léxico-sintácticos P_k previamente identificados, se buscan en el corpus nombres que puedan aparecer en la posición del nombre X por desambiguar dentro del patrón y se delimita así el paradigma P_{Pk} correspondiente a esta posición del patrón.

Para la desambiguación de una ocurrencia del nombre X integrada en el patrón P_k [X-R-Y], se extraerán del corpus los nombres sustitutos de X en el patrón P_k, es decir el paradigma correspondiente a la posición de X en P_k. Para esto, se mantienen fijos los demás elementos del patrón, R e Y, a nivel de lema y de categoría morfosintáctica, y se deja variable la posición de X, sólo a nivel de lema, no a nivel de categoría morfosintáctica, que será la misma de X (aquí N). Es necesario, por consiguiente, analizar previamente el corpus de búsqueda a nivel morfosintáctico, mediante el etiquetador morfológico. Los posibles sustitutos de X, en términos de lema, junto con X forman el paradigma asociado a la posición de X dentro de P_k.

Ejemplo: Partiendo con el patrón [*concierto*-N *para*-PREP *órgano*-N], en que se quiere desambiguar *órgano*, para *órgano*-N se obtiene del corpus *LEXESP* el siguiente paradigma de nombres: {*piano, violín, guitarra, solista, órgano, clarinete*}.

Paso 3º. Se identifican relaciones léxico-semánticas de las ofrecidas en *EWN* entre los nombres de cada uno de los paradigmas P_{Pk} previamente extraídos del corpus, con la ayuda de un algoritmo de DSA. Se asigna así a estos nombres un sentido de *EWN*.

Sobre el paradigma P_{Pk}, extraído del corpus, correspondiente a la ocurrencia de X dentro del patrón léxico-sintáctico S_k [X-R-Y], se aplica un algoritmo de DSA. Hemos elegido como algoritmo el de la *Marca de Especificidad Común*, ME (Montoyo y Palomar, 2000). Es un algoritmo basado en conocimiento, que utiliza *WordNet* para el proceso de desambiguación. Hemos introducido el algoritmo en el apartado 4.2.2.3.; recordamos aquí su funcionamiento. En su versión inicial, el algoritmo de las Marcas de Especificidad se aplica sobre los nombres presentes en el contexto oracional de la ocurrencia ambigua, incluida ésta. Se recorre toda la jerarquía de *WordNet*, en búsqueda de aquella marca de especificidad en cuyo subárbol haya la mayor densidad de sentidos de las diferentes palabras de entrada. A las palabras de entrada se les asignan los sentidos que tienen en este subárbol (figura 6.2.).

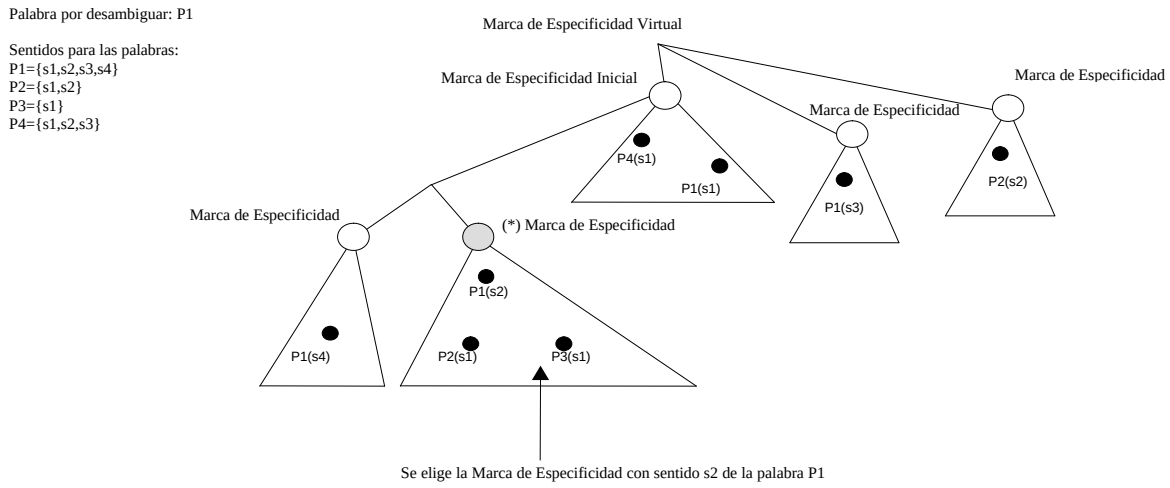

Figura 6.2. El algoritmo de la Marca de Especificidad Común

Debido a que el algoritmo se fundamenta en un enfoque de tipo "bolsa de palabras", el conjunto de entrada no debe necesariamente coincidir con las palabras de una oración. Por lo tanto, el algoritmo de las Marcas de Especificidad se puede generalizar a cualquier conjunto de palabras de entrada, sin afectar su funcionamiento. Nosotros lo aplicamos sobre el paradigma P_{Pk} previamente obtenido en el paso 2º.

La aplicación del algoritmo sobre el paradigma P_{Pk} lleva a la asignación de sentidos, uno o más, para X y los demás elementos del paradigma a él correspondiente, $P_{Pk.}$ Es decir, la desambiguación de X, como la de sus sustitutos, puede ser parcial.

La elección del algoritmo de la Marca de Especificidad Común es debida a sus características y a que se ajusta a nuestras necesidades desde varios puntos de vista: 1) usa *EWN* como fuente de conocimiento léxico; 2) toma como entrada y desambigua varios nombres a la vez y por lo tanto es idóneo para ser aplicado sobre un conjunto de nombres como los paradigmas asociados a la ocurrencia ambigua dentro de un patrón léxico-sintáctico; 3) hace un uso intensivo del conocimiento contenido en *EWN* ya que involucra simultáneamente los nombres de entrada en el proceso de DSA.

Ejemplo: La aplicación del algoritmo ME sobre el paradigma anterior, {*piano, violín, guitarra, solista, órgano, clarinete*}, lleva a las siguientes asignaciones de sentido: *piano_2, violín_2, guitarra_1, solista_1, órgano_4, clarinete_2.*

6.6 Estudio de caso: análisis de las limitaciones

Con el objetivo de identificar heurísticas de alta calidad, hemos estudiado y analizado manualmente el funcionamiento y la eficiencia de nuestra propuesta sobre el nombre *órgano*[207]. Hemos aplicado la estrategia sobre ocurrencias de este nombre en el corpus de entrenamiento de Senseval-2. Estas ocurrencias están etiquetadas con uno de los sentidos que *órgano* tiene en la fuente léxica de referencia en Senseval-2, *MiniDir*, sentidos que a su vez están en correspondencia con *synsets* de *WordNet*. Nuestro método asigna sentidos en términos de *synsets* de *WordNet*. Hemos podido así contrastar la asignación de sentidos obtenida por nuestro método con los sentidos correctos de Senseval-2. Los casos investigados corresponden a un número de veinte patrones léxico-sintácticos, de diferente estructura, que aparecen en la tabla 6.2.

Estructura morfosintáctica	Ejemplo
[**N** , N]	*órgano, célula*
[**N** CONJ N]	*órgano y orquesta* *órganos y tejidos* *órgano e institución*
[**N** ADJ]	*órgano administrativo* *órgano oficial* *órgano consultivo* *órgano judicial* *órgano eléctrico* *órgano sensorial* *órgano cutáneo*
[ADJ **N**]	*supremo órgano*
[**N** PREP N]	*órgano de dirección* *órgano del gobierno* *órgano del cuerpo* *órgano del PSOE* *órgano del Estado* *órgano del Ejecutivo*
[N PREP **N**]	*obras para órgano* *tráfico de órganos*

Tabla 6.2. Patrones lexico-sintácticos estudiados

[207] Es uno de los nombres sobre los cuales se ha organizado el ejercicio Senseval-2 para el español (apartado 5.2.2.).

Como corpus de búsqueda, hemos usado *CREA* y *LEXESP*[208].

Limitaciones. Hemos analizado los resultados obtenidos en cada uno de los tres pasos explicados en la sección anterior. Esta aplicación nos ha permitido identificar las limitaciones del método en su forma embrionaria y nos ha llevado a diseñar mejoras del método, bajo diversos aspectos. Enumeramos a continuación los problemas observados y en el siguiente apartado describimos las diferentes soluciones que proponemos con el objeto de subsanarlos.

Así, en la identificación de patrones léxico-sintácticos (paso 1º) se ha observado que hay "ruido" en la extracción de patrones léxico-sintácticos si sólo se usa el criterio estructural, ya que se obtienen secuencias que no corresponden a relaciones sintácticas. Damos a continuación algunos casos: [*UEFA*-N , *masa*-N], [*vida*-N *al*-PREP *arte*-N], [*arte*-N *a_través_del*-PREP *recorrido*-N]. Por tanto, el criterio estructural se ha mostrado insuficiente para esta operación.

En la misma identificación de patrones léxico-sintácticos, los patrones morfosintácticos considerados ([N1, N2]; [N ADJ]; [N VPART]; [N1 PREP N2]; [N1 CONJ* N2]) aseguran una cobertura limitada sobre la casuística de los textos, ya que quedan sin identificar muchos patrones léxico-sintácticos que incorporan las ocurrencias por desambiguar.

La delimitación del paradigma asociado a la ocurrencia ambigua dentro de un patrón dado (paso 2º) se encuentra perjudicada por la escasez de datos en la búsqueda dentro del corpus. El corpus *LEXESP* es demasiado pequeño para poder ofrecer datos suficientes en la aplicación del método.

El algoritmo de ME utilizado en el paso 3º está basado en la agrupación mayoritaria en la jerarquía de *EWN* de los sentidos próximos de las palabras de entrada (aquí el paradigma obtenido en el paso 2º). Esto implica una desviación en la asignación de sentido con respecto a esta agrupación mayoritaria. Así, el algoritmo no logrará identificar sentidos que estén fuera de esta agrupación en la red de *EWN*.

Por esto hay que decir que el nivel de desambiguación con ayuda del algoritmo ME está directamente relacionado con el grado de proximidad entre las diferentes palabras de entrada (aquí, el paradigma obtenido en el paso 2º). La heterogeneidad del paradigma asociado a la ocurrencia ambigua dentro del patrón empeora la calidad de la desambiguación; por lo tanto, hace falta asegurar un paradigma con un alto grado de homogeneidad.

Otra limitación del mismo algoritmo es el uso exclusivo de las relaciones de tipo hipo/hiperónimo entre las relaciones léxico-semánticas ofrecidas por *EWN*. Esta explotación parcial de la información relacional disponible en *EWN* determina que el rendimiento del algoritmo de ME sea limitado en la identificación de relaciones entre las palabras de entrada y por lo tanto en la asignación de su sentido.

Por otra parte, se ha observado que, en los patrones con dos nombres ([N1, N2]; [N1, N2]; [N1 CONJ* N2]; [N1 CONJ* N2]; [N1 PREP N2]; [N1 PREP N2]), el otro nombre del patrón es, a veces, muy informativo sobre el sentido de la ocurrencia ambigua y por consiguiente se debería considerar en el proceso de desambiguación.

6.7 Desarrollo del método

Hemos buscado soluciones concretas para superar las limitaciones del método básico identificadas en el estudio de caso, soluciones en parte sugeridas por los ejemplos analizados.

Así, en la operación previa de delimitación de los patrones léxico-sintácticos, hemos procedido de dos formas. Por una parte, mejorar la cobertura de esta operación y para ello hemos definido esquemas de búsqueda para la identificación de realizaciones discontinuas de los patrones. Por otra parte, mejorar la fiabilidad de esta operación y para ello hemos delineado posibles restricciones o filtros sobre las secuencias candidatas a ser patrones, eliminando las no idóneas.

Respecto al problema de escasez de datos, hemos optado por utilizar un corpus lo más amplio posible: *EFE* (75 millones de palabras)[209].

Para reducir o compensar la heterogeneidad del paradigma asociado a la palabra por desambiguar dentro del patrón léxico-sintáctico, hemos pensado en explotar intensivamente las restricciones mutuas entre las dos palabras de contenido léxico de un patrón léxico-sintáctico en el corpus. De esta manera se puede usar más información implícita del corpus asociada a la ocurrencia ambigua integrada en un

[208] Ver el apartado 7.1. para datos sobre estos corpus.
[209] Cf. el capítulo 7 de la experimentación. Presentamos más detalles sobre el corpus en el apartado 7.1.

patrón. A la vez, hemos considerado oportuno tratar los diferentes patrones que contienen una misma ocurrencia ambigua.

Así se minimiza también el problema de la escasez de datos que afecta a nuestro método como método basado en la búsqueda en el corpus. La consideración de ambas posiciones de contenido léxico en un patrón o de diferentes patrones correspondientes a la misma ocurrencia ambigua aumenta las probabilidades de hallar información paradigmática relacionada con la palabra por desambiguar.

Con el propósito de explotar intensivamente el conocimiento de tipo relacional contenido en *EWN*, hemos diseñado otro algoritmo, la *Prueba de Conmutabilidad*[210]. Este nuevo algoritmo usa la adaptación de *EWN*

dia.dSmes coig de(As n er dvAs tmoo en lasigneración dsmieenies a sla palabdes 5la) Tj 0 Tc -0.12 Tw ( ) T0.4 -12.6 TD -0.0126 Tc

Detallamos a continuación la modalidad de utilizar cada tipo de restricción para la identificación de patrones léxico-sintácticos.

a) *Estructura morfosintáctica.* Hemos mantenido los tipos básicos, predefinidos en el método embrionario, de patrones morfosintácticos en que intervienen los nombres (apartado 6.4.3.). Sin embargo, estos patrones pueden tener realizaciones discontinuas en los textos. Para cubrir tales casos, es decir, para mejorar la cobertura en la identificación de patrones léxico-sintácticos, hemos predefinido los siguientes *esquemas de búsqueda* en el corpus, que nos permitan la identificación de los patrones léxico-sintácticos considerados incluso cuando aparecen de forma discontinua en el texto:

Esquemas de búsqueda
[N_1 (((ADV) ADV) ADJ/VPART), (PREP) (DET) (((ADV) ADV) ADJ/VPART) N_2]
[N_1 (((ADV) ADV) ADJ/VPART) CONJ* (DET) (((ADV) ADV) ADJ/VPART) N_2]
[N ((ADV) ADV) ADJ/VPART (CONJ* ((ADV) ADV) ADJ/VPART)]
[N_1 (((ADV) ADV) ADJ/VPART) PREP (DET) (((ADV) ADV) ADJ/VPART) N_2]

Tabla 6.3. Esquemas de búsqueda

donde las unidades entre paréntesis son opcionales y las separadas por una barra son alternativas para una posición del esquema[214].

Las secuencias identificadas en el corpus en base de estos esquemas predefinidos se descomponen en patrones léxico-sintácticos simples. Por ejemplo, el esquema:

[N ADJ_1 CONJ* ADJ_2]

se descompone en los dos patrones simples:

[N ADJ_1]

y

[N ADJ_2].

Así, la secuencia:

[*partido-partido*-N *tumultuoso-tumultuoso*-ADJ *y-y*-CONJ *enrabietado-enrabietado*-ADJ]

se descompone en los patrones:

[*partido*-N *tumultuoso*-ADJ]

y

[*partido*-N *enrabietado*-ADJ].

Ofrecemos en el anexo 2 la lista de reglas que hemos definido para la descomposición de las secuencias halladas en base de los esquemas de búsqueda. La lista no cubre toda la casuística de las realizaciones de los patrones básicos. Un caso notable, por ejemplo, es el esquema:

[N_1 ADJ PREP N_2],

que se puede descomponer en:

[N_1 ADJ]

[N_1 PREP N_2],

o bien en:

[N_1 ADJ]

[ADJ PREP N_2].

Así, la secuencia:

[*autoridades-autoridad*-N *sanitarias-sanitario*-ADJ *de-de*-PREP *muchos-mucho*-DET *países-país*-N]

se puede descomponer en los patrones:

[*autoridad*-N *sanitario*-ADJ]

y

[*autoridad*-N *de*-PREP *país*-N].

En cambio, la secuencia:

[*autoridades-autoridad*-N *encargadas-encargado*-ADJ *de-de*-PREP *la-la*-DET *preservación-preservación*-N]

[214] Ver los esquemas detallados en el anexo 1.

se debe descomponer en los patrones:

[*autoridad*-N *encargado*-ADJ]

y

[*encargado*-ADJ *de*-PREP *preservación*-N][215].

La opción no es trivial y tampoco sigue una regla, dependiendo más bien de los lemas implicados. Su descomposición no se puede resolver a nivel morfosintáctico sino a nivel de lemas y por lo tanto supone un estudio más profundo, que se aleja del propósito fundamental de nuestro trabajo. Hemos preferido no considerar estos casos en vez de tratarlos de manera superficial.

En el caso de identificar patrones léxico-sintácticos que contienen una palabra dada N_0, se particularizan los patrones básicos generales previamente definidos para N_0 (tabla 6.4.).

Patrones básicos para N_0	
[N_0, N]	[N, N_0]
[N_0 CONJ* N]	[N CONJ* N_0]
[N_0 ADJ]	[ADJ N_0]
[N_0 VPART]	[VPART N_0]
[N_0 PREP N]	[N PREP N_0]

Tabla 6.4. Patrones básicos particulares para un nombre dado

Igualmente, se particularizan los esquemas definidas genéricamente para la búsqueda en el corpus de realizaciones discontinuas de los patrones de la palabra fijada N_0:

Esquemas morfosintácticas de búsqueda para N_0
[N_0 (((ADV) ADV) ADJ/VPART), (PREP) (DET) (((ADV) ADV) ADJ/VPART) N]
[N (((ADV) ADV) ADJ/VPART), (PREP) (DET) (((ADV) ADV) ADJ/VPART) N_0]
[N_0 (((ADV) ADV) ADJ/VPART) CONJ* (DET) (((ADV) ADV) ADJ/VPART) N]
[N (((ADV) ADV) ADJ/VPART) CONJ* (DET) (((ADV) ADV) ADJ/VPART) N_0]
[N_0 ((ADV) ADV) ADJ/VPART (CONJ* ((ADV) ADV) ADJ/VPART)]
[N_0 (((ADV) ADV) ADJ/VPART) PREP (DET) (((ADV) ADV) ADJ/VPART) N]
[N (((ADV) ADV) ADJ/VPART) PREP (DET) (((ADV) ADV) ADJ/VPART) N_0]

Tabla 6.5. Esquemas de búsqueda particulares para un nombre dado

donde los paréntesis indican elementos opcionales y la barra alternativas para una misma posición.

De manera correspondiente, se particularizan para N_0 la lista de reglas que descomponen las secuencias halladas mediante los esquemas de búsqueda previos en patrones léxico-sintácticos básicos. Ofrecemos en el anexo 2 la lista de estas reglas particulares de descomposición.

Para la eliminación del ruido en la extracción de patrones léxico-sintácticos en base al criterio estructural, o sea para alcanzar una alta fiabilidad en la extracción de patrones léxico-sintácticos, hemos impuesto restricciones sobre los elementos funcionales. Hemos realizado un estudio empírico de dimensiones reducidas, sobre la extracción de patrones para varios nombres, exclusivamente en base del criterio considerado. Así, primero hemos limitado las conjunciones a las dos más frecuentes, de tipo coordinativo, *y*, *o*, con sus variante eufónicas, *e*, *u*: CONJ* = {*y*, *o*, *e*, *u*}. La selección de las preposiciones, en cambio, ha planteado dificultades, debido a que la selección es muy dependiente del nombre estudiado y, por lo tanto, difícilmente automatizable. Las observaciones sobre el corpus han revelado que la preposición *de* suele ser la preposición dominante[216] y a la vez más fiable para la identificación correcta de patrones léxico-sintácticos. Esta constatación nos ha llevado en un primer momento a restringir las preposiciones a *de*. Sin embargo, el análisis de los patrones léxico-sintácticos extraídos exclusivamente con la preposición *de* ha puesto de relieve que incluso con esta restricción

[215] Ambos ejemplos son del corpus de prueba del Senseval-2 para el español.

[216] De hecho, se confirma un resultado consagrado en los estudios sobre la frecuencia de las unidades léxicas: la preposición *de* es la palabra más frecuente en las lenguas románicas occidentales (Sala, M. (coord.), *Vocabularul reprezentativ al limbilor romanice*, Bucuresti, Editura Stiintifica si Enciclopedica, 1988).

sobre la preposición hay ruido en la extracción de patrones léxico-sintácticos. Este estudio nos demuestra la insuficiencia del criterio estructural incluso con restricciones añadidas y ha puesto de manifiesto la necesidad de recurrir al criterio de frecuencia en el corpus de los patrones a nivel de lema.

b) *Frecuencia.* El cálculo de la frecuencia de los patrones potenciales en el corpus de búsqueda toma en consideración también las apariciones discontinuas de los patrones. Es decir, se buscan en el corpus apariciones de los patrones bajo la forma de los esquemas predefinidos, con y sin los elementos opcionales DET y ADJ.

Por ejemplo, si se calcula la frecuencia de aparición en el corpus del patrón formado por N_{10} $PREP_0$ N_{20}, la búsqueda en el corpus se realiza de la siguiente manera:

N_{10} (((ADV) ADV) ADJ/VPART) $PREP_0$ (DET) (((ADV) ADV) ADJ/VPART) N_{20}[217],

donde ADJ/VPART significa alternancia entre ADJ y VPART. En la búsqueda, los elementos N_{10}, $PREP_0$ y N_{20} son fijos a nivel de lema, mientras que los elementos DET y ADJ/VPART son variables a nivel de lema.

Así, en el cálculo de la frecuencia de aparición del patrón:

[*partido*-N *de*--PREP *derecha*-N]

en el corpus, se contarían también secuencias como:

[*partido-partido*-N *político-político*-ADJ *de-de*-PREP *extrema-extrema*-ADJ *derecha-derecha*-N].

En el anexo 3, ofrecemos la lista de los esquemas de búsqueda particulares que se utilizan para el cálculo de la frecuencia.

La propuesta presentada para la identificación de los patrones, combinando el criterio estructural con los filtros de frecuencia, permite que el proceso de extracción automático de los patrones sea independiente de un *chunker* o de un *parser*[218] y sea condicionada sólo de un etiquetador morfológico.

No hemos tratado la cuestión del nivel de los patrones (forma flexiva *vs.* lema *vs.* categoría sintáctica). Se puede operar con patrones léxico-sintácticos a varios niveles lingüísticos de sus constituyentes L1, R y L2: forma, lema o categoría morfosintáctica. Nuestra hipótesis es que cuanto más precisa sea la información morfológica asociada a las palabras, más restricciones o más información habrá sobre su significado en el patrón. Consideramos, por lo tanto, que idealmente se debería probar con el nivel más fino y, si no hay evidencia en el corpus a este nivel, subir al nivel de generalidad superior y trabajar en ese nivel, etc. Es decir operar con niveles ordenados de manera ascendente respecto a su grado de generalidad. Sin embargo, en el presente estudio hemos optado por operar en el nivel intermediario, el lema. Hemos estado motivados además, a favor de esta opción, por la sencillez en el proceso de delimitación de los patrones como por el problema de la escasez de datos que afectaría excesivamente a los patrones léxico-sintácticos a nivel de forma flexiva. El análisis del impacto en la utilización de los patrones léxico-sintácticos a diferentes niveles de generalidad no es de interés central dentro de nuestra tesis sino que será objeto de un estudio futuro.

6.7.2 Explotación de los patrones léxico-sintácticos: información asociada a la ocurrencia ambigua

La explotación de los patrones léxico-sintácticos que contienen una ocurrencia ambigua tiene como objetivo la obtención de información vinculada con la ocurrencia. Cuanto más amplia y diversificada sea, mejor, ya que se identificará mejor su sentido.

En el presente trabajo, partimos del análisis de los siguientes aspectos:

a) el tipo de información asociada a la ocurrencia ambigua, que está integrada en un patrón léxico-sintáctico;

[217] El despliegue de todas las secuencias correspondientes a este esquema sintético es similar al del punto a), descrito en el anexo 2.

[218] Ambos son herramientas para el análisis sintáctico, total y respectivamente parcial. En otras palabras, el *parser* realiza un análisis sintáctico a nivel profundo, hasta la construcción del árbol de la estructura sintáctica de la oración. En cambio, el *chunker* se limita al análisis sintáctico en un nivel superficial, delimitando solamente los segmentos que corresponden a sintagma no recursivos, o sea sintagmas sin incluir a su vez sintagmas.

b) la interacción, en el proceso de asignación de sentido, entre los patrones léxico-sintácticos de la ocurrencia y entre estos patrones y la oración de la ocurrencia.

Detallamos a continuación cada una de estas cuestiones.

a) Delimitamos dos tipos de información que se pueden asociar a una ocurrencia ambigua a partir de un patrón léxico-sintáctico que la incorpora:

- información paradigmática, que corresponde a palabras asociadas a la ocurrencia a lo largo de un eje paradigmático;
- información sintagmática, que corresponde a palabras asociadas a la ocurrencia a lo largo de un eje sintagmático.

Nuestra idea es que el sentido de una palabra en un contexto está determinado ante todo por la colaboración entre la información sintagmática y paradigmática asociada a ella a través de los patrones léxico-sintácticos en que participa. Ambos tipos de información son útiles para la asignación de sentido y en la presente tesis se explotarán igualmente en este proceso. La obtención de información vinculada a la ocurrencia ambigua supera los límites de la oración y por lo tanto es necesaria la explotación de un corpus, sobre el que se hace la búsqueda de manera transversal. Así, la información paradigmática que se extrae del corpus asociada a una ocurrencia ambigua consiste en el paradigma formado por los posibles sustitutos de la palabra ambigua dentro del patrón. La información sintagmática corresponde a las palabras que co-ocurren con el patrón. En nuestra investigación, consideramos exclusivamente los nombres que co-ocurren con el patrón.

Además, el estudio empírico ha revelado que, para la identificación de una ocurrencia ambigua, en algunos casos es útil aplicar el algoritmo de DSA sobre los dos nombres del patrón léxico-sintáctico (cuando se trata de un patrón con dos nombres). Así, un algoritmo de DSA basado en una fuente de tipo relacional como *EWN* permite identificar una eventual relación léxico-semántica entre las dos unidades de contenido léxico del patrón. Estas relaciones se pueden expresar tanto a través de una relación sintáctica de coordinación (por ejemplo, la hipo/hiperonimia: *órganos o partes*) como a través de una relación de dependencia (por ejemplo, la meronimia: *presidente del gobierno*).

Sin embargo, en la asignación del sentido en el proceso de DSA, tomaremos igualmente en consideración la información proporcionada por la oración entera. Nos limitamos, en el presente estudio, a considerar sólo los nombres de la oración.

b) El uso de diferentes informaciones asociadas a la ocurrencia ambigua, obtenidas a partir de los patrones léxico-sintácticos, plantea el problema de la interacción entre estas informaciones respecto a la aplicación del algoritmo de DSA. Identificamos dos posibilidades:

- En la primera, las restricciones mutuas entre las varias informaciones consideradas actúan al principio, previamente a la aplicación del algoritmo, como filtros recíprocos. En este caso, el proceso de DSA se realiza sobre la información así filtrada.
- En la segunda, las relaciones entre las varias informaciones se consideran al final del proceso de desambiguación. Aquí, se realiza la asignación del sentido usando autónomamente cada una de las informaciones consideradas y luego se combinan las asignaciones individuales de los sentidos.

Se trata de establecer si, para la desambiguación de una ocurrencia, en el momento de la aplicación de un algoritmo de DSA, se toma en consideración la información asociada a la ocurrencia ambigua respecto a:

1) cada patrón léxico-sintáctico en parte,
2) los varios patrones de la ocurrencia juntos, o bien
3) éstos y a toda la oración.

En otras palabras, los patrones de la ocurrencia ambigua entre ellos, por una parte, y los patrones con la oración, por otra parte, pueden interaccionar al nivel de la información que proporcionan como vinculada con la ocurrencia ambigua o bien al nivel de las propuestas que ofrecen para el sentido de la ocurrencia.

Estas posibilidades en la explotación de los patrones léxico-sintácticos en la asignación de sentido están determinadas por la adopción o no de las siguientes *restricciones*:

> *(R1) Se considera que la contribución de los patrones léxico-sintácticos al sentido de la ocurrencia ambigua es independiente del resto de la oración.*
>
> *(R2) Se supone que la influencia de cada uno de los patrones léxico-sintácticos sobre el significado de la ocurrencia ambigua es independiente una de la otra.*

En la presente tesis aplicamos la doble reducción R1 y R2 en la explotación de los patrones léxico-sintácticos, ya que consideramos que los patrones intervienen sobre el sentido de la palabra ambigua de manera independiente entre sí y respecto de la oración. El sistema de DSA que proponemos se construye con las heurísticas correspondientes a las restricciones R1 y R2 juntas. La opción está relacionada con la hipótesis de "tendencia hacia un sentido por patrón sintagmático" (en inglés, "*towards one sense per syntagmatic pattern*") y al propósito de obtener patrones etiquetados con sentidos, reutilizables en futuras asignaciones de sentido. Nos respaldan parcialmente los experimentos con humanos de Miller y Charles (1991)[219]. Sin embargo, en la experimentación (presentada aquí en el capítulo 7), analizamos la idoneidad de las restricciones R1 y R2 que hemos adoptado, eliminándolas progresivamente en el orden R2, R1, y confrontando los resultados que se obtienen en la desambiguación, con o sin ellas.

Las consideraciones previas delimitan los posibles tipos de información, asociados a la ocurrencia ambigua, que utilizamos para la asignación de sentido. Definimos un conjunto para cada una de las informaciones:

- el otro nombre del patrón o el par de nombres del patrón (PAT_k);
- la información paradigmática (PAR_k);
- la información sintagmática ($SINT_k$);
- la oración (OR).

Nombramos con X_0 la ocurrencia ambigua y con $S_k = [X_0\text{-}R_k\text{-}Y_k]$, sus patrones léxico-sintácticos locales. Los índices k asociados a los conjuntos hacen referencia al patrón S_k respecto del cual se obtiene la información. Presentamos, a continuación, la obtención de cada conjunto, que se explicará para un patrón $[X_0\text{-}R_0\text{-}Y_0]$ fijado.

- PAT_k es el par de nombres del patrón. El conjunto se define sólo para los patrones con dos nombres, de tipo: [N1 CONJ N2], [N1, N2] o bien [N1 PREP N2]. Su obtención es trivial.

- PAR_k es la información paradigmática o sea los posibles sustitutos de la palabra dentro del patrón. El paradigma se obtiene a través de la búsqueda en el corpus, manteniendo fijo el patrón léxico-sintáctico a nivel de lema y categoría morfosintáctica, y dejando libre, a nivel de lema, sólo la posición de la ocurrencia ambigua dentro del patrón. Más detalladamente, se mantienen fijos los demás elementos R_0 e Y_0 del patrón $[X_0\text{-}R_0\text{-}Y_0]$ (menos los elementos opcionales DET, ADV y ADJ[220]), en términos de lemas y de categorías morfosintácticas, y se deja variable la posición de la ocurrencia ambigua X_0, dentro su categoría morfosintáctica (aquí, N). Se buscan en el corpus, que debe estar previamente procesado a nivel morfológico, los posibles nombres X que pueden aparecer en el patrón, en la posición de la ocurrencia ambigua X_0. La búsqueda en el corpus se debe realizar de manera similar a las realizadas en la identificación de los patrones léxico-sintácticos, teniendo en cuenta las realizaciones discontinuas del patrón en el corpus.

Por ejemplo, si se debe desambiguar X_0 en el patrón:

$[X_0\ PREP_0\ N_0]$,

[219] Cf. apartados 6.4.2.2., 6.4.2.4.

[220] Hace excepción el patrón N ADJ, en el cual el elemento ADJ es fijo y obligatorio.

se mantienen fijos los elementos $PREP_0$ y N_0 a nivel de lema y se buscan los sustitutos X de X_0 (con la categoría morfosintáctica predeterminada, la de X_0) en todas las posibles realizaciones en el corpus del patrón:

[X ((ADV) ADJ) $PREP_0$ (DET) ((ADV) ADJ) N_0].

Los posibles elementos X_0 se identificarán igualmente a nivel de lema. Estos nombres, junto con X_0, forman el conjunto PAR_{k0} de tipo paradigmático.

El conjunto de estos sustitutos contiene "ruido" que altera la asignación de sentido, o sea palabras que no están relacionadas con el sentido correcto de la palabra dentro del patrón (cf. apartado 6.6.). Por lo tanto, hemos diseñado diferentes modalidades de obtener subconjuntos más homogéneos y/o más relevantes para el sentido de la palabra ambigua dentro del patrón. Hemos delimitado así los siguientes conjuntos de información paradigmática como variantes del conjunto PAR_k:

- PAR_{1k} es el paradigma correspondiente a la posición de la palabra ambigua dentro de un patrón léxico-sintáctico considerado en su totalidad.

- PAR_{2k} es el subparadigma formado por los sustitutos de la palabra ambigua dentro del patrón con una frecuencia por encima de un umbral U preestablecido. En concreto, en la experimentación (capítulo 7) hemos probado el valor 5 para U.

Por otra parte, consideramos que, dentro de un patrón léxico-sintáctico, las palabras que pueden ocupar una posición de contenido léxico y las palabras que pueden ocupar la otra posición de contenido léxico se imponen mutuamente restricciones semánticas. Partiendo con esta hipótesis, de los sustitutos del nombre por desambiguar en el marco del patrón léxico-sintáctico dado, se guardan aquellos que comparten con éste diferentes palabras en la otra posición de contenido léxico del patrón. De esta manera, dentro del paradigma correspondiente a la palabra ambigua se identifican palabras que estén lo más relacionadas con esta. Hemos explotado estas restricciones mutuas entre las posiciones de contenido léxico de un patrón de dos maneras: por una parte, como un simple filtro sobre el paradigma de sustitutos de la palabra ambigua (conjunto PAR_{3k}); por otra parte, para intentar delimitar, dentro del paradigma, los subparadigmas correspondientes a posibles zonas conceptuales distintas y así diferentes sentidos de la palabra ambigua en el marco del patrón (conjuntos PAR_{4k}).

- PAR_{3k} está formado por los sustitutos de la palabra ambigua dentro del patrón, filtrados por la condición de que compartan con la palabra ambigua más de un elemento en la otra posición de contenido léxico del patrón. Para la obtención de este conjunto, se siguen los pasos que presentamos a continuación:

-Se buscan todos los sustitutos X_i de X_0 dentro del patrón (es decir los nombres X_i que pueden aparecer en la posición de X_0).

-Para X_0 y para cada uno de estos X_i, se buscan todos los sustitutos de Y_0 dentro del patrón (o sea las palabras de la misma categoría morfosintáctica con Y_0 que pueden aparecer en su posición en el patrón). Notamos con $\{Y_{0j}\}$ los sustitutos de Y_0 correspondientes a X_0 y con $\{Y_{ij}$, con i constante dentro de cada conjunto$\}$ los sustitutos de Y_0 correspondientes a X_i.

-Hacemos la intersección de $\{Y_{0j}\}$ con cada uno de los conjuntos $\{Y_{ij}\}$ y anotamos con Q_{0i} las intersecciones.

-Guardamos sólo los sustitutos X_i correspondiente a los cuales hemos obtenido un conjunto Q_{0i} de cardinalidad p_i superior a 1; así X_0 y X_i compartirán más palabras en la posición de Y_0. Los sustitutos guardados formarán el conjunto PAR_{3k} asociado a la ocurrencia ambigua X_0 en el patrón [X_0-R_0-Y_0].

- PAR_{4k} remite a los subconjuntos de sustitutos de la palabra ambigua determinados por dos elementos comunes en la otra posición del patrón. Los subconjuntos se obtienen siguiendo las operaciones presentadas a continuación.

-Se buscan todos los sustitutos Y_j de Y_0 dentro del patrón (es decir las palabras de la misma categoría morfosintáctica que Y_0 que pueden aparecer en la posición de Y_0).

-Para Y_0 y para cada uno de estos Y_j, se buscan todos los sustitutos de X_0 dentro del patrón (o sea los nombres que pueden aparecer en su posición en el patrón). Notamos con $\{X_{i0}\}$ los

sustitutos de X_0 correspondientes a Y_0 y con {X_{ij}, con j constante dentro de cada conjunto} los sustitutos de X_0 correspondientes a Y_j.

-Hacemos la intersección de {X_{i0}} con cada uno de los conjuntos {X_{ij}}. Anotamos P_{0j} a cada una de las intersecciones.

-Guardamos los P_{0j} con cardinalidad superior a 1 (los de cardinalidad 1 están formados exclusivamente por X_0). Se obtienen así subparadigmas para la posición de X_0 en el patrón. Cada conjunto P_{0j} será un conjunto de tipo PAR_{4k} para X_0 dentro del patrón.

En el paso 3º del método, se aplicará el algoritmo sobre cada conjunto P_{0j}. Se guardan todos los sentidos que se obtiene de esta manera. Si se obtienen así más sentidos, se considerará que la palabra ambigua puede tener más sentidos dentro del patrón: es decir que constituye una excepción de la tendencia hacia un sentido por patrón léxico-sintáctico (la hipótesis "*towards one sense per lexico-syntactic pattern*"). Además, el procedimiento nos permite una generalización en la anotación del sentido, de manera que esta operación se realice no sólo para el patrón de partida sino para un grupo de patrones.

Si se obtiene el mismo sentido *s* para dos conjuntos distintos P_{0j} y $P_{0j'}$, entonces se unifican los dos conjuntos y X_0 tendrá el sentido *s* en cualquier patrón de tipo [A-R_0-B], donde A es la reunión de los conjuntos P_{0j} para los cuales X_0 tiene el sentido *s*, mientras que B es la reunión de Y_0 y de sus sustitutos correspondientes a los conjuntos P_{0j}.

Esta última modalidad de formar agrupaciones dentro del paradigma puede ayudar a identificar si hay más sentidos para la ocurrencia ambigua dentro del patrón.

Ejemplificamos la obtención de los conjuntos de tipo PAR_{4k}, para la desambiguación del nombre *órgano* dentro del patrón *órgano del gobierno*. En el caso de los conjuntos $PAR3_k$, se procede de manera similar, siguiendo los pasos previamente descritos.

-Notamos los elementos del patrón de partida: [*órgano del gobierno*] = X_0 – R - Y_0.

-Buscamos los nombres que pueden sustituir a *gobierno* en el patrón, o sea los N en *órgano*_de_N:

Y_j = {*partido, cuerpo, animal, administración, …*}

-Para *gobierno* y para cada uno de los sustitutos Y_j de *gobierno*, buscamos los nombres X_{i0} y respectivamente X_{ij} (manteniendo j constante en cada caso) que pueden sustituir a *órgano* en el patrón, o sea:

N_*de_gobierno*: {*órgano, jefe, presidente, declaración, política, …*} = {X_{i0}}

N_*de_partido*: {*órgano, jefe, presidente, periódico, política, …*} = {X_{ij}, j constante}

N_*de_cuerpo*: {*órgano, forma, temperatura, color, corazón, …*} = {X_{ij}, j constante}

N_*de_animal*: {*órgano, corazón, mano, temperatura, …*} = {X_{ij}, j constante}

N_*de_administración*: {*órgano, política, jefe, declaración, …*} = {X_{ij}, j constante}

-Para *gobierno* y para cada uno de sus sustitutos Y_j, hacemos la intersección entre los conjuntos {X_{i0}} y {X_{ij}, j constante} correspondientes, obteniendo los conjuntos P_{0j}:

N_*de_gobierno* ∩ N_*de_partido* = {*órgano, jefe, presidente, política, …*}

N_*de_gobierno* ∩ N_*de_cuerpo* = {*órgano*}

N_*de_gobierno* ∩ N_*de_animal* = {*órgano*}

N_*de_gobierno* ∩ N_*de_administración* = {*órgano, jefe, declaración, política, …*}

-De los conjuntos así obtenidos, se guardan sólo los que tienen cardinalidad superior a uno:

N_*de_gobierno* ∩ N_*de_partido* = {*órgano, jefe, presidente, política, …*}

N_*de_gobierno* ∩ N_*de_administración* = {*órgano, jefe, declaración, política, …*}

Estos dos conjuntos serán los conjuntos de tipo PAR_{4k} sobre los que se aplicará uno de los dos algoritmos de DSA (Prueba de Conmutabilidad o Marcas de Especificidad).

-Aun más, si sobre estos conjuntos se obtiene un mismo sentido, los conjuntos se pueden unificar, lo que lleva a una generalización del patrón de partida a un grupo de patrones para los cuales queda válida la asignación de sentido para la palabra objetivo (aquí, *órgano*).

En el proceso de DSA, se utilizará, como información paradigmática (conjunto PAR_k), uno de los conjuntos PAR_{1k}, PAR_{2k}, PAR_{3k}, PAR_{4k}. La variante más útil para la asignación de sentido se elegirá a través de la experimentación[221].

[221] Nosotros hemos probado parte de estas variantes en la experimentación (capítulo 7).

- $SINT_k$ está constituido por las palabras que coaparecen frecuentemente con un patrón fijado.
Hemos definido este conjunto para disminuir el desvío que pueda haber en el uso del conjunto PAR, debido a que este puede contener palabras no relacionadas con la ocurrencia ambigua y su sentido. La frecuencia hace de filtro, de manera que se guardan las palabras cuya co-aparición con el patrón no es casual y por lo tanto es significativa. Asumimos que, debido a la hipótesis "*towards one sense per lexico-syntactic pattern*", estas palabras tendrán relevancia, a través del patrón, también para el sentido de la palabra ambigua dentro del patrón. En la experimentación (capítulo 7), hemos probado dos valores del umbral V de frecuencia mínima para las palabras co-ocurrentes con los patrones: 5 y 10. La obtención de las palabras co-ocurrentes con un patrón dado se hace con la ayuda de los esquemas de búsqueda asociadas al patrón, que corresponden a sus posibles realizaciones discontinuas en el texto, de manera similar al cálculo de la frecuencia de los patrones en el corpus.

Así, por ejemplo, partiendo con el patrón [*autoridad*-N *civil*-ADJ], se usa el esquema de búsqueda [*autoridad*-N ((ADV) ADV) ADJ/VPART (CONJ* ((ADV) ADV) *civil*-ADJ)] para hallar ocurrencias del patrón en el corpus. De las oraciones encontradas, se extraen los nombres con su frecuencia. Así, entre los nombres más frecuentes que co-aparecen con el patrón de partida en el corpus EFE son los siguientes: {*presidente, militar, jefe, gobierno, estados, policía, misión, ciudad, país, guardia, acuerdo, ministro, capital, alcalde, seguridad, palacio*}.

- *OR* es el conjunto de todos los nombres de la oración en que aparece la ocurrencia ambigua. Su obtención es trivial.

Hemos resumido, en la tabla 6.4, los conjuntos de las posibles explotaciones de los patrones léxico-sintácticos locales de una palabra para la obtención de información asociada a ella. Para la asignación de sentido, sobre estos conjuntos se aplicará un algoritmo de DSA.

Restricciones	Conjunto	Caracterización/Descripción	
R1 + R2	PAT_k	El patrón: el par de nombres de un patrón	
	PAR_k	Información paradigmática	PAR_{1k}: El paradigma entero de sustitutos de la palabra ambigua dentro del patrón
			PAR_{2k}: Los sustitutos de la palabra ambigua dentro del patrón con frecuencia superior a U
			PAR_{3k}: Los sustitutos de la palabra ambigua filtrados por la condición de que compartan con la palabra ambigua varios elementos en la otra posición del patrón
			PAR_{4k}: Subparadigmas de sustitutos de la palabra ambigua dentro del patrón, determinados por dos elementos comunes en la otra posición del patrón
	$SINT_k$	Información sintagmática: los nombres frecuentemente coocurrentes con el patrón	
	OR	La oración: los nombres	

Tabla 6.4. Los conjuntos de información asociados a una ocurrencia ambigua que se usan en el proceso de DSA

6.7.3 Asignación de sentido: Prueba de Conmutabilidad

Dedicamos este apartado a la presentación de las modificaciones que hemos aportado al tercer paso de nuestra estrategia respecto a la variante básica. Nos detenemos en la presentación de los algoritmos que usamos para la asignación de sentido.

Nos hemos guiado en el diseño de alternativas para este tercer paso según las dificultades observadas en el estudio de caso, aquí presentadas en el apartado 6.6. Hemos elaborado un algoritmo, denominado *Prueba de Conmutabilidad*, que explote la adaptación de *EWN* bajo la forma de *Discriminadores de Sentido* (introducidos en el apartado 6.4.2.3). Además, hemos diseñado variantes de este algoritmo, que utilizan de diferente manera la fuente léxica *Discriminadores de Sentido*.

En la base del algoritmo *Prueba de Conmutabilidad* se halla la hipótesis de que dos palabras que pueden conmutar en un contexto dado son afines semánticamente. En términos de la adaptación que hemos desarrollado a partir de *EWN* (apartado 6.4.2.3.), esto significa que si una ocurrencia ambigua puede ser sustituida en sus patrones léxico-sintácticos por un discriminador de sentido, entonces se le puede asignar el sentido correspondiente al respectivo discriminador de sentido.

Por ejemplo, si en el patrón *obras para órgano* se quiere desambiguar *órgano*, se observa en el corpus que éste puede sustituirse por nombres como *violín*, *guitarra*, *piano*, etc. Estos nombres son discriminadores del sentido *órgano_4* 'instrumento musical' y por lo tanto *órgano* puede tener el sentido 4 en la ocurrencia dada. Para optimizar el cálculo, preferimos realizar un proceso equivalente al descrito previamente. Así, primero extraemos del corpus los sustitutos de la palabra ambigua dentro del patrón léxico-sintáctico y luego hacemos la intersección del paradigma obtenido con el conjunto de discriminadores correspondientes con cada uno de los sentidos de la palabra.

Como hemos procedido también en el caso del algoritmo *Marca de Especificidad*, generalizamos el algoritmo *Prueba de Conmutabilidad*, para que se aplique sobre cualquier conjunto de palabras de entrada. Para una ocurrencia dada del nombre X, y un conjunto de palabras asociado a ella a partir de sus patrones léxico-sintácticos, el algoritmo

Sentido tal como los hemos extraído en el apartado 6.4.2.3. Una posibilidad para minimizar el problema es ampliar los conjuntos de discriminadores. Proponemos la siguiente ampliación: para cada discriminador existente en la adaptación estricta, se extraen los hipónimos, y luego se eliminan los eventuales elementos comunes a otros conjuntos de Discriminadores, para preservar la condición de disjuntividad. Llamamos esta nueva fuente "Discriminadores de Sentidos extendidos", y la variante correspondiente del algoritmo, "Prueba de Conmutabilidad extendida".

Sintetizamos los diferentes algoritmos en la tabla 6.5.

Algoritmo de base	Variante
Marca de Especificidad	-
Prueba de Conmutabilidad	Prueba de Conmutabilidad básica
	Prueba de Conmutabilidad restringida
	Prueba de Conmutabilidad ampliada

Tabla 6.5. Algoritmos de DSA

6.8 Sistema de DSA

Arquitectura del sistema de DSA. Describimos a continuación el método tal como lo hemos desarrollado en el apartado precedente. La diversificación de la información asociada a la ocurrencia ambigua en base de sus patrones y también la diversificación de los algoritmos determina varias heurísticas de DSA. Una heurística es resultado de la combinación entre un conjunto de información asociado a la ocurrencia por desambiguar y un algoritmo de DSA que se aplica sobre este conjunto.

De acuerdo con nuestro objetivo de obtener una alta fiabilidad en la asignación de los sentidos, utilizaremos las diferentes heurísticas que se configuran. El uso de diferentes heurísticas en el proceso de DSA impone un cuarto paso en el método, con el objetivo de combinar las heurísticas individuales.

Nuestro método tendrá, por lo tanto, los siguientes pasos:

Paso 1º. Identificación de los patrones léxico-sintácticos de la ocurrencia ambigua.

Paso 2º. Obtención de información asociada a la ocurrencia ambigua a partir de los patrones y de la oración.

Paso 3º. Aplicación de los algoritmos de DSA sobre los conjuntos obtenidos en el paso 2º. (Lo que equivale a la implementación de las heurísticas individuales.)

Paso 4º. Combinación de las heurísticas individuales.

Presentamos en la figura 6.4. la arquitectura del sistema de DSA construido en base a este método, con un módulo para cada uno de los pasos del método. En la arquitectura hemos representado también la fuente de información que accede cada módulo. Subrayamos que este sistema corresponde a la adopción de las dos restricciones R1 y R2. Sin embargo, hemos implementado también los sistemas alternativos, sólo para la restricción R1 y sin ninguna de las dos restricciones. Estos últimos dos sistemas de DSA se presentarán y evaluarán separadamente en los experimentos del apartado 7.2.

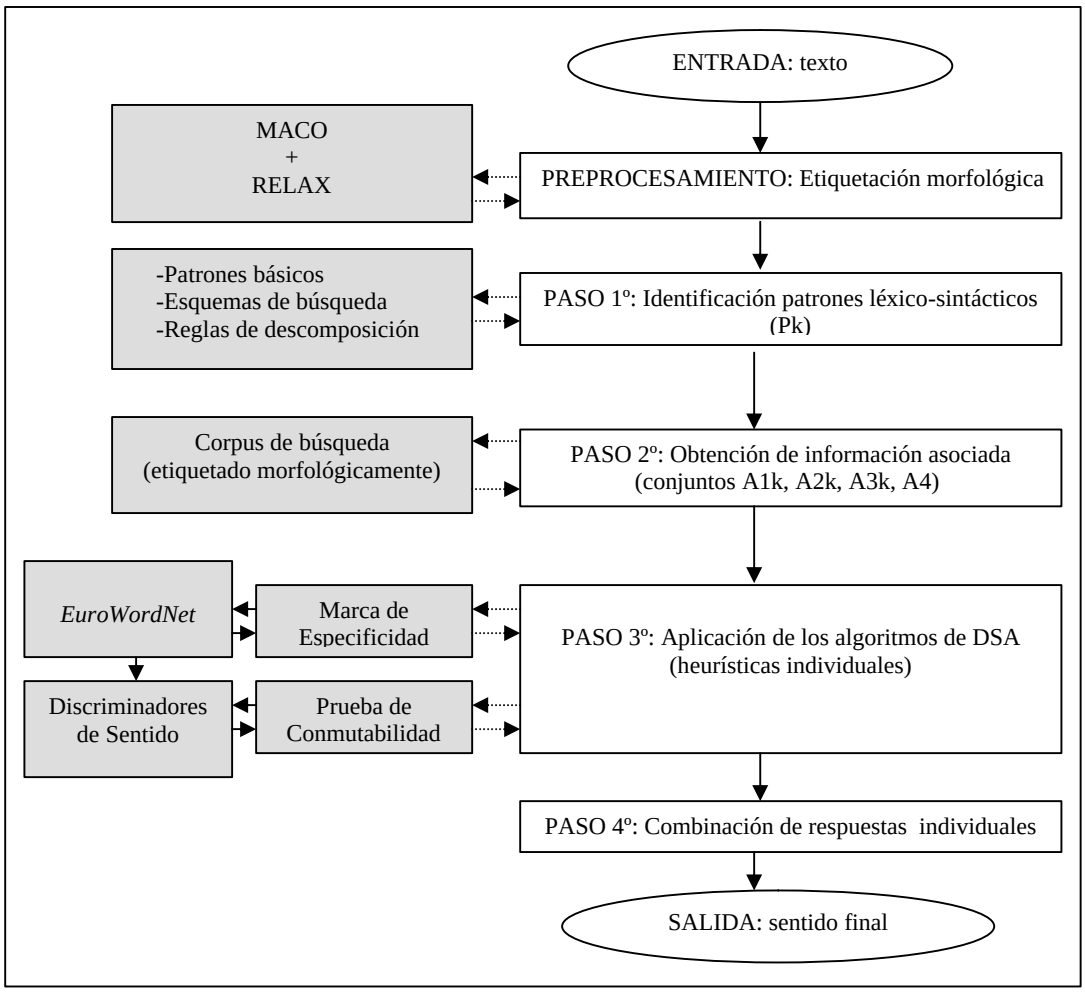

Figura 6.4. Arquitectura del sistema de DSA

Heurísticas. Detallamos a continuación las heurísticas individuales que hemos usado y su combinación. Así, las heurísticas que hemos obtenido aplicando las reducciones R1 y R2 (apartado 6.7.2) son las combinaciones entre uno de los algoritmos ME y PC (una de sus variantes) y uno de los conjuntos PAT, PAR_k (en una de las variantes PAR_{1k} – PAR_{4k}), $SINT_k$ y OR:

- *ME_PAT*$_k$: se aplica la Marca de Especificidad sobre el par de nombres del patrón
- *PC_PAT*$_k$: se aplica la Prueba de Conmutabilidad (en una de sus variantes) sobre el otro nombre del patrón
- *ME_PAR*$_k$: se aplica la Marca de Especificidad sobre el conjunto de información paradigmática asociada a la ocurrencia ambigua dentro del patrón (en una de las variantes PAR_{1k}, PAR_{2k}, PAR_{3k}, PAR_{4k})
- *ME_PAR*$_k$: se aplica la Marca de Especificidad sobre el conjunto información sintagmática asociada a la ocurrencia ambigua dentro del patrón
- *PC_SINT*$_k$: se aplica la Prueba de Conmutabilidad (en una de sus variantes) sobre el conjunto de información sintagmática asociada a la ocurrencia ambigua dentro del patrón.
- *PC_SINT*$_k$: se aplica la Prueba de Conmutabilidad (en una de sus variantes) sobre el conjunto de información sintagmática asociada a la ocurrencia ambigua dentro del patrón;
- *ME_OR*: se aplica la Marca de Especificidad sobre el conjunto de palabras (nombres) de la oración;
- *PC_OR*: se aplica la Prueba de Conmutabilidad (en una de sus variantes) sobre el conjunto de palabras (nombres) de la oración.

La aplicación de estas heurísticas sigue el funcionamiento de los dos algoritmos, ME y PC, previamente descritos.

Mostramos un resumen de las heurísticas en la tabla 6.6., según los dos factores que se utilizan en una heurística: el algoritmo y el conjunto de palabras sobre que se aplica el algoritmo.

Información / (conjunto) Algoritmo	*El patrón*	*Información paradigmática*	*Información sintagmática*	*Oración*
	PAT_k	PAR_k	$SINT_k$	OR
ME	ME_PAT_k	ME_PAR_k	ME_SINT_k	ME_OR
PC	PC_PAT_k	PC_PAR_k	PC_SINT_k	PC_OR

Tabla 6.6. Heurísticas de DSA utilizadas

Combinación de las heurísticas. La combinación de las heurísticas usadas debe ser conforme con nuestro enfoque a la DSA: la desambiguación de la ocurrencia ambigua se hace en primer lugar en relación con sus patrones léxico-sintácticos. Las hipótesis sobre la que se fundamenta el método, recordamos, son la dependencia del sentido de una palabra en el contexto principalmente de sus relaciones sintácticas y la "tendencia hacia un único sentido por patrón léxico-sintáctico" ("*towards one sense per lexico-syntactic pattern*"), respectivamente. Por consiguiente, organizamos las heurísticas en dos categorías: heurísticas relacionadas con los patrones (y eventualmente en cierta medida también con la oración) y heurísticas relacionadas con la oración (tabla 6.7).

Grupo	Heurísticas
I: Heurísticas relacionadas con los patrones	ME_PAT_k PC_PAT_k
	ME_PAR_k PC_PAR_k
	ME_SINT_k PC_SINT_k
II: Heurísticas relacionadas con la oración	ME_OR PC_OR

Tabla 6.7. Tipología de las heurísticas de DSA utilizadas

A partir de esta organización de las heurísticas, consideramos oportuna una modalidad progresiva de combinar las heurísticas dentro del sistema de DSA (ver figura 6.5.):

a) por una parte, se combinan las heurísticas del grupo I. Para ello, se implementan todas las heurísticas para cada uno de los patrones y luego se combinan sus resultados para obtener una propuesta de sentido por parte de cada patrón; al final se juntan las propuestas de sentido de todos los patrones;
b) por otra parte, se combinan las heurísticas del grupo II, es decir, las dos heurísticas relacionadas con la oración. Para ello, se implementan ambas heurísticas y luego se combinan sus resultados para obtener una propuesta de sentido a partir de la oración;
c) finalmente, se combinan las heurísticas del grupo I con heurísticas del grupo II. Para ello, se combinan las propuestas de sentidos obtenidas en base al patrón con las propuestas de sentidos obtenidas en base a la oración. A partir de esta combinación, se decide el sentido final.

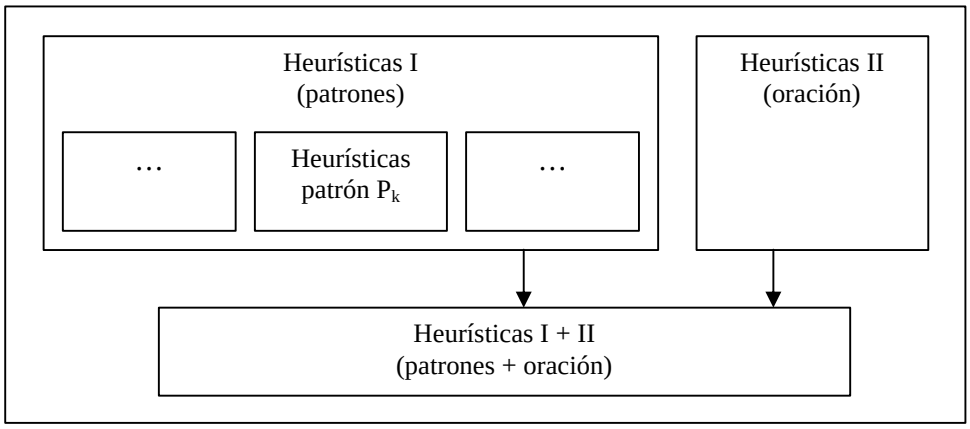

Figura 6.5. Combinación de las heurísticas

La modalidad efectiva de proceder a cada combinación se ha establecido en la experimentación (capítulo 7).

Ilustramos la aplicación del método propuesto (en una variante particular) sobre la ocurrencia número 75 del nombre *órgano* en el corpus de prueba de Senseval-2:

> *Un informe del <head>órgano</head> de gobierno de los jueces advierte de que no pretende una modificación legislativa, sino proponer soluciones a problemas del nuevo Código, pues, señala, "provoca la comisión de delitos, tiene penas desproporcionadas y ha generado una situación penitenciaria que alcanzará cotas insostenibles".*

PASO 1.Identificación de los patrones léxico-sintácticos de la ocurrencia ambigua
En este paso y en el siguiente, particularizamos para *órgano* los patrones básicos predefinidos y los esquemas de búsqueda.

1a. Utilizando estos esquemas particulares, hallamos en el ejemplo la secuencia:
[*informe*-N *de*-PREP *órgano*-N *de*-PREP *gobierno*-N].

1b. De esta secuencia, extraemos dos patrones básicos:
P1=[*informe*-N *de*-PREP *órgano*-N]
y
P2=[*órgano*-N *de*-PREP *gobierno*-N].

PASO 2. Extracción de información asociada a la ocurrencia ambigua
Este paso se divide en los siguientes:

2a. Extracción de información paradigmática a partir del corpus
Para extraer el paradigma correspondiente a la posición de *órgano* en cada uno de los dos patrones léxico-sintácticos previamente identificados, P_1 y P_2 respectivamente, procedemos de la manera que se describe a continuación. Con la ayuda de los esquemas de búsqueda particularizados, buscamos en el corpus posibles nombres en la posición de X en cualquier realización de cada uno de los dos patrones. Obtenemos dos conjuntos, correspondientes a P_1 y P_2. Si optamos, como variante de la información paradigmática, guardar sólo los 20 sustitutos más frecuentes dentro del patrón, obtenemos los siguientes dos conjuntos:
PAR_1: {*gestión, comisión, policía, prensa, servicio, organización, experto, coyuntura, organismo, ponencia, autoridad, auditoría, agencia, perito, intervención, observador, conclusión, situación, fiscalización, grupo, emisora, candidatura, ...*}

PAR_2: {*jefe, programa, año, órgano, formación, equipo, partido, coalición, representante, acción, miembro, cambio, comité, parte, alianza, período, pacto, funcionario, crisis, responsabilidad, mes, alternativa, acuerdo, ...*}

2b. Extracción de información sintagmática a partir del corpus
Buscamos en el corpus las oraciones que contienen el patrón P_1 y, separadamente, las oraciones que contienen el patrón P_2. De las oraciones halladas en cada caso, extraemos sólo los nombres. Si, como variante de la información sintagmática, establecimos 10 como umbral mínimo de frecuencia de coocurrencia con el patrón de partida, obtenemos los siguientes dos conjuntos correspondientes a los patrones P_1 y P_2:
$SINT_1$ = {*tráfico, sugerencia, producto, observación, mención, medida, justicia, estupefaciente, estilo, desvío, consumo, Junta, anteproyecto, Departamento_de_Justicia, ...*}
$SINT_2$ = {*juez, magistrado, presidente, CGPJ, EFE, fuente, acuerdo, miembro, reunión, poder, comisión, ciudad, Senado, representante, vicepresidente, texto, función, forma,...*}

2c. Extracción de información sintagmática de la oración
Los nombres de la oración son los siguientes:
OR = {*órgano, informe, gobierno, juez, modificación, solución, código, comisión, delito, penas, situación, cota*}

PASO 3. Aplicación de los algoritmos de DSA sobre la información asociada a la ocurrencia ambigua
La aplicación de las heurísticas individuales nos lleva a las propuestas de sentidos que enumeramos a continuación, para el patrón P_1, para el patrón P_2 y para la oración respectivamente[222].
Para el patrón P_1:

ME_PAR_1: s_4, s_5
ME_SINT_1: -
PC_PAR_1: s_1 (2), s_2 (8), s_3 (5), s_4 (2), s_5 (7)
PC_SINT_1: s_1 (0), s_2 (3), s_3 (3), s_4 (0), s_5 (3)

Para el patrón P_2:

ME_PAR_2: s_2
ME_SINT_2: -
PC_PAR_2: s_1 (0), s_2 (9), s_3 (3), s_4 (5), s_5 (5)
PC_SINT_2: -

Para la oración:

ME_OR: s_1, s_2
PC_OR: s_1 (0), s_2 (1), s_3 (0), s_4 (0), s_5 (0)

Cuando hay más propuestas de sentido, en el caso de las heurísticas con *Prueba de Conmutabilidad*, se elige el sentido con más discriminadores de sentido.

En la tabla 6.8., sintetizamos las propuestas de sentido para *órgano* de parte de todas las heurísticas usadas:

P_1				P_2				Oración	
ME_PAR_1	ME_SINT_1	PC_PAR_1	PC_SINT_1	ME_PAR_2	ME_SINT_2	PC_PAR_2	PC_SINT_2	ME_OR	PC_OR
S_4 s_5	-	s_2	s_2	s_2	-	s_2	-	s_1 s_2	s_2

2 2 9 .

PASO 4. La asignación final de sentido

Ejemplificamos la modalidad de combinar las propuestas de sentido por parte de las heurísticas individuales tomando una decisión sobre el sentido final en cuanto a las opciones particulares relacionadas con:

1) la combinación de las propuestas por parte de las heurísticas asociadas a un patrón (grupo I),
2) la combinación de las propuestas por parte de las heurísticas asociadas a la oración (grupo II) y
3) la combinación de las propuestas por parte de las heurísticas asociadas a los patrones y a la oración (grupos I y II).

Así, para cada patrón, elegimos el sentido más votado. Hacemos la intersección de las diferentes propuestas del sentido de los diferentes patrones. En cuanto a las heurísticas relacionadas a una oración, si hay respuestas de ambas heurísticas, hacemos la intersección con los sentidos propuestos. Si la intersección es nula, elegimos la respuesta de la heurística PC_PC o, si ésta no tiene resultados, de la heurística ME_OR. Para la asignación del sentido final, aplicamos primero las heurísticas del grupo I y luego las del grupo II. Si no hay propuesta de parte de las heurísticas del grupo I, cogemos las respuestas del grupo II. En caso de paridad entre dos o más sentidos, elegimos la propuesta de la heurística más precisa (de las seis) entre las que tienen respuesta.

En este caso, obtenemos el sentido 2 de *EWN* de parte de ambos patrones y de parte de la oración, por lo tanto se asigna este sentido a la ocurrencia de *órgano*, sentido que corresponde al sentido 3 de referencia en Senseval-3.

7 Experimentación

En el capítulo anterior hemos perfilado el método de DSA en sus rasgos definitorios. Sin embargo, hemos dejado abiertas varias cuestiones, debido a la necesidad de una investigación empírica previa a cualquier decisión. Presentamos en este capítulo una serie de experimentos destinados a ultimar el método y además a estudiar algunas implicaciones sobre el proceso de DSA y sobre los elementos que contribuyen a perfilar el sentido de una palabra en el contexto. Ante todo, debemos resaltar la cantidad de parámetros involucrados en nuestra propuesta y por lo tanto el amplio volumen de la experimentación requerida para un estudio profundo del método. En estas circunstancias, la experimentación que presentamos a continuación no ha podido cubrir toda la problemática que nuestra aproximación plantea a la DSA. Nuestro principal interés en el desarrollo de esta investigación ha sido estimar el potencial de nuestra propuesta, bajo diferentes aspectos. En otras palabras, nos hemos centrado en identificar los puntos clave en torno a los cuales se articula el método, es decir los puntos de flexibilidad, en cada uno de los cuales se pueden abrir líneas de experimentación más detallada. Por lo tanto, la investigación realizada tiene un valor prototípico, destinada a orientar una investigación necesaria de más alto calado alrededor de la estrategia de DSA propuesta. Las conclusiones actuales necesitan una comprobación a gran escala para que se vuelvan en conocimiento y contribución real sobre la cuestión de la DSA.

Abrimos la experimentación con una presentación de los recursos utilizados (apartado 7.1.). Los experimentos han sido orientados inicialmente hacia la comprobación de los elementos básicos de nuestra propuesta (apartado 7.2.), y luego hemos afinado el método en sus parámetros (apartado 7.3). Presentamos también la evaluación del método en el marco del ejercicio *Senseval*-3 (apartado 7.4.).

7.1 Entorno experimental

Guiados por los principios del apartado 6.3., hemos usado pocos recursos externos en nuestro método y especialmente recursos disponibles, sin necesidad de un procesamiento. Como herramienta de preprocesamiento, hemos necesitado sólo un etiquetador morfológico. La información de referencia sobre los sentidos de las palabras por desambiguar nos ha sido proporcionada por *EuroWordNet*, que hemos explotado en su forma original y en nuestra adaptación, a través de los *Discriminadores de Sentido*. El método que proponemos explota intensivamente la información proporcionada por los corpus. En nuestra experimentación, hemos usado dos corpus: *LEXESP* y *EFE*. Un último elemento que hemos usado en la experimentación ha sido el muestrario sobre el cual hemos probado el sistema y evaluado los resultados: el corpus de prueba de Senseval-2, excepto en la evaluación que se ha desarrollado en el marco de Senseval-3 (apartado 7.4). Detallamos a continuación cada uno de los elementos mencionados que intervienen en la experimentación.

Analizador y desambiguador morfológico. Nuestro método opera al nivel de lema y de categoría morfosintáctica, por lo tanto necesita un preprocesamiento de la información textual con la cual trabaja: el contexto oracional mismo en que aparece la ocurrencia por desambiguar y respectivamente el corpus de búsqueda. Para el análisis textual al nivel morfológico, utilizamos una combinación de dos herramientas: un analizador morfológico, MACO (Civit, 2003; Carmona *et al*., 1998) y un desambiguador morfológico (RELAX, Atserias *et al*., 1998). En esta última variante, el desambiguador alcanza una precisión alta, de un 97,4%.

Fuentes léxicas estructuradas. Como fuente léxica de referencia para los sentidos de la palabra, hemos optado por la componente española de *EuroWordNet*, en dos variantes: la variante relacionada a *WN 1.5* y la variante *WEI*[223]. Las razones de la selección de *EWN* fueron presentadas en el apartado 6.4.2.3. A la vez, hemos implementado un recurso léxico alternativo, los Discriminadores de Sentido, derivados del mismo *WordNet* español. Hemos desarrollado este recurso sólo para las palabras investigadas, del muestrario de evaluación, sin embargo el método para su obtención es relativamente sencillo y es aplicable a cualquier nombre de *EuroWordNet*. En el anexo 3, listamos los *Discriminadores de Sentido* (básicos) adquiridos para las palabras que hacen objeto de nuestro análisis.

Corpus de búsqueda. En nuestra investigación hemos hecho uso de tres corpus: *LEXESP*, *EFE* y *CREA*. Si hemos desarrollado el estudio de caso sobre el corpus *LEXESP* y tangencialmente sobre el corpus *CREA*, en la experimentación hemos explotado casi exclusivamente el corpus *EFE*.

El corpus *LexEsp*[224] (Sebastián *et al*., 2000) es un corpus de cinco millones y medio de palabras del español estándar escrito contemporáneo (1978-1995) tanto de España como de América Latina, que recoge fuentes variadas: prensa, literatura, etc.

El *corpus EFE* pertenece a la base de datos EfeData, en que se recoge, se organiza y se mantiene accesible la información difundida por la agencia EFE a partir de 1982, actualizada diariamente. El corpus que hemos utilizado contiene más de 70 millones de palabras.

El *Corpus de referencia del español actual (CREA)* de la Real Academia Española es un banco de datos del español contemporáneo, con textos de diversa procedencia. Como corpus de referencia, intenta proporcionar información exhaustiva acerca de todas las variedades relevantes del español desde 1975 hasta la actualidad, de todos los países de habla hispana. Con 140 millones de registros en octubre de 2003, está previsto que se llegue a 160 millones, a finales de 2004. El corpus se puede consultar en la página web: www.rae.es.

Muestrario de evaluación. Hemos desarrollado los experimentos en las condiciones de concurso la segunda y tercera edición española de Senseval. La excepción es la evaluación dentro del ejercicio Senseval-3 (apartado 7.4.), que se ha realizado sobre el corpus de prueba preparado para esta edición. Las pruebas se han focalizado sobre las palabras por desambiguación en estas competiciones. La selección de este muestrario ha sido motivada por varios factores: la escasez de corpus en español etiquetados con sentidos, para la evaluación de los resultados; la utilidad de una evaluación relativa de los resultados obtenidos, a través de la comparación con otros sistemas de DSA.

Recordamos, en la tabla 7.1., los datos para los dos corpus de Senseval-2 y de Senseval-3, limitados a los nombres.

	Número ocurrencias			
	Senseval-2		Senseval-3	
Palabra	Corpus de entrenamiento	Corpus de prueba	Corpus de entrenamiento	Corpus de prueba
autoridad	87	68	268	132
bomba	75	74	-	-
canal	114	82	262	131
circuito	73	98	261	132
corazón	98	94	123	62
corona	78	80	124	64
gracia	98	122	72	38
grano	55	44	117	61
hermano	77	114	128	66
masa	89	82	172	85
naturaleza	110	112	258	128
operación	94	94	134	66

[223] Web *EuroWordNet* Interface (version 0.2). La fuente está disponible a la página web: http://nipadio.lsi.upc.es/wei1.html.

[224] *Léxico Informatizado del español.*

órgano	130	162	263	131
partido	101	114	133	66
pasaje	70	82	220	111
programa	94	94	267	133
tabla	77	82	130	64
arte	-	-	251	121
banda	-	-	230	114
columna	-	-	129	64
letra	-	-	226	114
mina	-	-	134	66
TOTAL	1520	799	3902	1949

Tabla 7.1. El muestrario de evaluación

Así, en Senseval-2, se ha usado *MiniDir*, mientras nosotros hemos usado la variante *1.5* de *EuroWordNet*. Ofrecemos en la tabla 7.2. el grado de cobertura de los sentidos de las dos fuentes léxicas mediante la correspondencia entre sentidos.

SENSEVAL-2: 17 nombres	Minidir		*EuroWordNet*	
	Número	Porcentaje	Número	Porcentaje
Total sentidos	53	100%	83	100%
Sentidos en correspondencia	47	88,67%	65	78,31%

Tabla 7.2. Grado de cobertura de la correspondencia
entre los sentidos de Senseval-2 y de *EuroWordNet* español

En Senseval-3, se ha usado *Minidir-2*, mientras nosotros hemos usado la variante *WEI* de *EuroWordNet*. Ofrecemos en la tabla 7.3 el grado de cobertura de los sentidos de las dos fuentes léxicas mediante la correspondencia entre sentidos.

SENSEVAL-3: 21 nombres	Minidir-2		*EuroWordNet* (WEI)	
	Número	Porcentaje	Número	Porcentaje
Total sentidos	107	100%	147	100%
Sentidos en correspondencia	80	74,76%	121	82,31%

Tabla 7.3. Grado de cobertura de la correspondencia
entre los sentidos de Senseval-3 y *WEI español*

Por lo tanto, en el corpus de prueba de Senseval-2 - sobre el que hemos desarrollado casi todos los experimentos iniciales y para el refinamiento del método -, sólo parte de las ocurrencias por desambiguar tienen asignaciones de sentido traducibles en términos de *EuroWordNet - WN1.5*. Esta información se muestra en la tabla 7.4.

Ocurrencias	Número	Porcentaje
Todas las ocurrencias del corpus de prueba de Senseval-2	799	100%
Ocurrencias con sentido asignable en términos de *EWN*	688	86,10%

Tabla 7.4. Grado de cobertura de las ocurrencias de Senseval-2
con sentidos de *EuroWordNet* español

Creemos que esta correspondencia parcial entre los inventarios de sentidos que se usan por nuestro sistema afecta negativamente la calidad del proceso de desambiguación en términos de cobertura, ya

que se podrá evaluar sólo el 86,10% de las respuestas. La precisión relativa, en cambio, no se ve afectada por estas limitaciones.

Debido a estos problemas y a nuestro interés centrado en la precisión, prescindimos, en la presentación de los resultados obtenidos en la experimentación, de los valores para la cobertura y la precisión absoluta. Nos limitamos a los valores para la precisión relativa, con sólo unas pocas excepciones.

7.2 Experimentos de control

Objetivo: El primer grupo de experimentos se han orientado hacia la comprobación de algunos elementos fundamentales de nuestro enfoque, como:

- la evaluación de la extracción de los patrones;
- la hipótesis de la "tendencia hacia un sentido por patrón léxico-sintáctico";
- la utilidad de la información paradigmática para la DSA;
- la utilidad de los Discriminadores de Sentido en la DSA y del algoritmo de la Prueba de Conmutabilidad;
- la idoneidad de la aplicación del algoritmo de la Marca de Especificidad sobre el paradigma de sustitutos de la palabra ambigua dentro del patrón;
- la dependencia del corpus utilizado;
- la idoneidad de las restricciones R1 y R2.

Parámetros. Con el propósito de justificar empíricamente nuestra propuesta, los primeros experimentos se han realizado sobre dos conjuntos de control para la información asociada a la ocurrencia ambigua. Así, como información paradigmática, hemos considerado la reunión de los conjuntos de sustitutos de la palabra ambigua en sus diferentes patrones léxico-sintácticos: $S_1 = \bigcup_k PAR_k$, mientras que para la información sintagmática hemos usado los nombres de la oración: $S_2 =$ OR.

Punto de partida (de referencia). Un método nuevo justifica su existencia en la medida en que aporta alguna mejora a los métodos ya existentes. Debido a que nosotros aprovechamos el algoritmo de la *Marca de Especificidad* (ME) previamente utilizado, tomamos este algoritmo, en su forma ordinaria, como punto de referencia en estos experimentos iniciales. Así, la *Marca de Especificidad*, complementada por ocho heurísticas (Montoyo, 2002), obtiene los siguientes resultados en las condiciones de Senseval-2:

Algoritmo	Conjunto información	Precisión relativa	Precisión absoluta	Cobertura
ME	OR	56,6%	43,5%	76,0%
			49,95%	**88,26%**

Tabla 7.5. Resultados del algoritmo *Marca de Especificidad* sobre la oración (en Senseval-2)

Organización. Estructuramos la experimentación de control de la manera sintetizada en la tabla 7.6.

Cuestión estudiada	Parámetros	Experimento
Utilidad de la información paradigmática para la DSA		1
Utilidad de los Discriminadores de Sentido para la DSA		2
Modalidad usada para la identificación de patrones	Estructura	3a
	Filtro	3b
Calidad de la información asociada a la ocurrencia	Filtro	4
Hipótesis "*towards one sense per lexico-syntactic pattern*"		5
Hipótesis sobre el uso de los patrones léxico-sintácticos para la DSA	Restricción R1	6a
	Restricción R2	6b

Tabla 7.6. Experimentos de control

A continuación, detallamos cada uno de estos experimentos iniciales.

EXPERIMENTO 1
Objetivo. Nos proponemos ante todo comprobar la utilidad de la información paradigmática para la DSA. Para este experimento, hemos usado dos patrones básicos:

[N ADJ]
[N PREP N]

y los siguientes esquemas de búsqueda:

[N ADJ (CONJ* ADJ)]
[N (ADJ) PREP* (DET) (ADJ) N],

con restricciones sobre el elemento relacional: CONJ* = {*y, o*}, PREP* = {*de*}. La aplicación del algoritmo Marcas de Especificidad lleva a los resultados de la tabla 7.7.

Corpus	Algoritmo	Conjunto información	Precisión relativa	Precisión absoluta	Cobertura
LEXESP (5,5 millones palabras)	ME	$\bigcup_k \text{PAR}_k$	45,7%	7,5%	16,4%
				8,5%	**18,64%**

Tabla 7.7. Resultados obtenidos usando la información paradigmática (con el algoritmo Marca de Especificidad)

Conclusión. Los resultados son positivos: la información paradigmática sí es útil para la DSA.

EXPERIMENTO 2
Objetivo. El propósito de este experimento es comprobar la utilidad del algoritmo de los Discriminadores de Sentido (DS) y de la Prueba de Conmutabilidad (PC) para la DSA. Respecto a la prueba anterior, hemos ampliado los patrones básicos usados:

[N, N]
[N CONJ* N]
[N PREP N]
[N ADJ]
[N PART]
[ADJ N]
[PART N]

y también los esquemas de búsqueda:

[N1 ((ADV) ADJ/VPART) , (DET) N2]
[N1 , (DET) ((ADV) ADJ/VPART) N2]
[N1 ((ADV) ADJ/VPART) CONJ* (DET) N2]
[N1 CONJ* (DET) ((ADV) ADJ/VPART) N2]
[N1 PREP (DET1) ((ADV2) ADJ2/VPART2) N2]
[N ((ADV) ADJ/VPART)]
[N ((ADV2) ADJ2/VPART2 CONJ*) (ADV1) ADJ1/VPART1)]
[ADJ1/VPART1 (CONJ* (ADV2) ADJ2/VPART2) N],

con restricciones sólo sobre las conjunciones (conjunciones coordinativas): CONJ* = {*y, o, e, u*}. Hemos desarrollado tres pruebas, en las que hemos aplicado el algoritmo Prueba de Conmutabilidad respectivamente sobre el conjunto S_1, sobre S_2, y sobre ambos. Presentamos los resultados en la tabla 7.8.

	Corpus	Algoritmo	Conjunto información	Precisión relativa	Precisión absoluta	Cobertura
EXPERIMENTO 2ª	EFE (>70millones palabras)	PC	$\cup$ PAR_k	54,1%	11,6%	21,4%
					13,15%	**24,32%**
EXPERIMENTO 2b			OR	59,6%	4,7%	7,9%
					5,35%	**8,97%**
EXPERIMENTO 2c			$\cup$ PAR_k + OR	56,1%	15,2%	27,1%
					17,28%	**30,8%**

Tabla 7.8. Resultados del algoritmo Prueba de Conmutabilidad
(sobre información paradigmática y sintagmática)

Conclusión. El experimento demuestra que la Prueba de Conmutabilidad es viable como algoritmo de DSA. La confrontación de las pruebas 2a y 2b sugiere que la Prueba de Conmutabilidad usa con precisión similar los dos tipos de información, paradigmática y sintagmática, pero que obtiene mayor cobertura con la información paradigmática. Además, la diferencia de precisión entre las pruebas 2b y 2c indica que la información paradigmática mejora los resultados.

EXPERIMENTO 3
Objetivos. Este experimento está orientado hacia la comprobación de la metodología usada para la identificación de patrones con respecto al criterio estructural (experimento 3a) y al uso de un filtro de frecuencia (experimento 3b). Nos motiva en esta prueba el hecho de que el nivel de DSA, en nuestro enfoque, es altamente dependiente de la identificación de patrones tanto en términos de cantidad como de calidad. El experimento se ha desarrollado en las mismas condiciones que el experimento 2.

EXPERIMENTO 3a
Específicamente, analizamos primero la cobertura de las ocurrencias del muestrario de evaluación por parte de los patrones identificados y el grado de explotación de los patrones en la desambiguación. Recogemos en la tabla 7.9. los resultados de este análisis.

	Ocurrencias del corpus de prueba de Senseval-2	
Total ocurrencias	799	100%
Ocurrencias con patrones identificados	565	70,71%
		82,12%
Ocurrencias con respuestas	168	21,02%
		24,41%
Ocurrencias con respuestas correctas	91	11,38%
		13,22%

Tabla 7.9. Cobertura de las ocurrencias con patrones y grado de explotación de los patrones para la DSA

EXPERIMENTO 3b
Segundo, estudiamos el impacto que el uso de un filtro de frecuencia ($f \geq 2$) tiene en la identificación de los patrones.

Patrones identificados	640	
Patrones repetidos (corpus EFE + Senseval-2 prueba)	163	25,4%

Patrones repetidos (corpus EFE + Senseval-2 prueba)	163	
Patrones buenos	157	96,3%

Tabla 7.10. Cobertura y calidad de los patrones repetidos en el corpus

EXPERIMENTO 4
Objetivo: Nos proponemos en este experimento identificar modalidades para mejorar la calidad de la información que se obtiene del corpus para una ocurrencia ambigua. Específicamente, comprobamos el impacto que la frecuencia tiene como filtro sobre la calidad de la información asociada a la ocurrencia ambigua. Así, aplicamos un filtro de frecuencia mínimo 2 sobre los sustitutos de la palabra dentro del patrón. Usamos los mismos patrones básicos y esquemas de búsqueda que en los experimentos 2 y 3.

Corpus	Algoritmo	Conjunto información	Precisión relativa	Precisión absoluta	Cobertura
EFE (>70millones)	PC	S_1 (filtro $f \geq 2$)	61,7%	7,0%	11,3%

Tabla 7.11. Resultados obtenidos con información filtrada ($f \geq 2$)

Conclusión. La comparación entre los experimentos 4 y 2a indica que el filtro de frecuencia sobre la información paradigmática mejora la precisión de la desambiguación.

EXPERIMENTO 5
Objetivo: En este experimento comprobamos la idoneidad de nuestra hipótesis sobre la tendencia hacia un sentido dentro de un patrón léxico-sintáctico ("*towards one sense per lexico-syntactic pattern*").
Hemos comprobado en qué medida la hipótesis se verifica para los patrones identificados en el corpus de prueba de *Senseval-2*. Para ello, se han buscado los patrones que se repiten y se han confrontado las asignaciones de sentido en las diferentes ocurrencias de los patrones. Presentamos los datos de estas comprobaciones en la tabla 7.12.

Patrones repetidos (Senseval-2 prueba)	46	100%
Patrones con un único sentido	45	97,8%

Tabla 7.12. Patrones repetidos en el corpus de prueba de *Senseval-2*

Conclusión. La integración en patrones léxico-sintácticos es adecuada como primer paso hacia la reducción de la ambigüedad de una palabra y por lo tanto para la DSA.

EXPERIMENTO 6
Objetivo: En este experimento centramos nuestra atención en la modalidad de usar los patrones léxico-sintácticos para la DSA. En concreto, analizamos la idoneidad de las dos reducciones que hemos asumido en la construcción de nuestro sistema de DSA (cf. apartado 6.7.): la independencia entre los patrones léxico-sintácticos en la asignación de sentido (R1), la independencia entre los patrones léxico-sintácticos por una parte y la oración por otra en la asignación de sentido (R2). Hemos investigado las dos restricciones en el orden R2 (experimento 6a), R1 (experimento 6b).

EXPERIMENTO 6a
Para comprobar la validez de la hipótesis de independencia entre los patrones léxico-sintácticos en la asignación de sentido contrastamos los resultados que se obtienen en la desambiguación con y sin esta hipótesis. La eliminación de la restricción R2 significa la interacción de la información asociada a la ocurrencia ambigua, que se obtiene considerando la ocurrencia integrada en cada uno de los patrones. Definimos los siguientes conjuntos de palabras para cubrir esta interacción: PAR_INTERS, SINT_INTERS. La introducción de los conjuntos se hace para una ocurrencia ambigua X_0 que participa en los patrones léxico-sintácticos locales S_k: X_0-R_k-Y_k.

- *PAR_INTERS*: los posibles sustitutos comunes de la ocurrencia ambigua dentro de todos los patrones locales identificados. El conjunto se obtiene como intersección entre los paradigmas

extraídos del corpus para la posición de la ocurrencia ambigua X en los diferentes patrones léxico-sintácticos locales, paradigmas considerados enteros (es decir los conjuntos PAR_k).

- *SINT_INTERS*: las palabras co-ocurrentes con todos los patrones identificados para la ocurrencia. Para determinar el conjunto, se interseccionan los conjuntos $SINT_k$ de los diferentes patrones de la ocurrencia ambigua.

La información que caracteriza la ocurrencia ambigua en este caso es la reunida en la tabla 7.13.

Restricciones	Conjunto	Caracterización/Descripción
Sin R2, sólo R1	PAR_INTERS	La intersección de los paradigmas asociados a la palabra ambigua respecto de cada patrón en parte
	SINT_INTERS	La intersección de la información sintagmática obtenida para los patrones individuales
	OR	Los nombres de la oración

Tabla 7.13. La información asociada a la ocurrencia ambigua si se elimina la restricción R2

De aquí, las heurísticas que se obtienen si se renuncia a la restricción R2, guardando sólo R1, son las de la tabla 7.14. a continuación.

		Información (conjuntos correspondientes)		
		Patrones léxico-sintácticos interactivos entre ellos, independientes de la oración		*Oración*
		Información paradigmática	*Información sintagmática*	
		PAR_INTERS	SINT_INTERS	OR
Algoritmo	ME	ME_PAR_INTERS	ME_SINT_INTERS	ME_OR
	PC	PC_PAR_INTERS	PC_SINT_INTERS	PC_OR

Tabla 7.14. Las heurísticas usadas si se elimina la restricción R1

Nosotros hemos comprobado la validez de R2 sólo mediante la evaluación del algoritmo Prueba de Conmutabilidad aplicado sobre la información paradigmática que corresponde a la adopción o no de la restricción R2. O sea hemos comparado las heurísticas PC_PAR y PC_PAR_INTERS respectivamente.

Restricciones	Heurística	Cobertura	Precisión relativa	Precisión absoluta
R1 + R2	PC_PAR	24,32%	54,10%	13,15%
R1	PC_PAR_INTERS	15,11%	44,23%	6,68%

Tabla 7.15. Evaluación de la Prueba de Conmutabilidad si se elimina la restricción R1

Los resultados son favorables a la heurística PC_PAR_INTERS que se obtiene cuando se asume la restricción R2, lo que nos confirma en la elección que hemos operado en la construcción de nuestro sistema de DSA.

EXPERIMENTO 6b

En esta prueba hemos comprobado la hipótesis sobre la independencia entre los patrones y la oración en la asignación de sentido, o sea la restricción R1. La eliminación, además de R2, de la restricción R1 corresponde a la interacción de la información asociada a la ocurrencia ambigua en base de todos sus patrones léxico-sintácticos con la información ofrecida por toda la oración. En este caso, se usarán los siguientes conjuntos de palabras asociados a una ocurrencia ambigua X_0 que participa en los patrones léxico-sintácticos locales S_k: X_0-R_k-Y_k: PAR_INTERS', SINT_INTERS'.

- *PAR_INTERS'* corresponde a los posibles sustitutos de la ocurrencia ambigua (dentro de todos sus patrones locales) que aparecen en la oración. Para obtener este conjunto, se hace la intersección entre el subconjunto común a los paradigmas extraídos del corpus para la posición de la ocurrencia ambigua X_0 en los diferentes patrones léxico-sintácticos locales (el conjunto PAR_INTERS) y el conjunto de nombres de la oración (el conjunto OR).
- *SINT_INTERS'* consiste en las palabras coocurrentes con el patrón que aparecen también en el contexto de la ocurrencia por desambiguar. La obtención del conjunto es trivial, como intersección entre los conjuntos SINT_INTERS y OR

En la tabla 7.16., sintetizamos la información que se obtiene si se eliminan ambas restricciones R1 y R2.

Restricciones	Conjunto	Caracterización
Sin R1 ni R2	PAR_INTERS'	La intersección de los paradigmas asociados a la palabra ambigua respecto de cada patrón en parte y el conjunto de nombres en la oración
	SINT_INTERS'	La intersección de la información sintagmática obtenida para los patrones individuales y la oración

Tabla 7.16. La información asociada a la ocurrencia ambigua
si se eliminan ambas restricciones R1 y R2

Las heurísticas correspondientes que se obtienen si se renuncia a ambas restricciones R1 y R2 se recogen en la tabla 7.17.

Información (conjunto) / Algoritmo	*Patrones léxico-sintácticos interactivos entre ellos y con la oración*	
	Información paradigmática	*Información sintagmática*
	PAR_INTERS'	SINT_INTERS'
ME	ME_ PAR_INTERS'	ME_ SINT_INTERS'
PC	PC_ PAR_INTERS'	PC_ SINT_INTERS'

Tabla 7.17. Las heurísticas de DSA si no se aceptan las restricciones R1 y R2

Para la evaluación de la idoneidad de la reducción R1, hemos procedido como en el caso de la restricción R2. Es decir, evaluamos la heurística que consiste en la aplicación del algoritmo Prueba de Conmutabilidad sobre la información paradigmática correspondiente a la adopción o no de la restricción R1. Para ello, hemos comparado las heurísticas PC_PAR_INTERS y PC_PAR_INTERS' respectivamente. Los resultados obtenidos se presentan en la tabla 7.18.

Restricciones	Heurísticas individuales relacionadas con los patrones y la oración	Cobertura	Precisión relativa	Precisión absoluta
Sólo R1	PC_ PAR_INTERS	15,11%	44,23%	6,68%
Sin R1 ni R2	PC_PAR_INTERS'	0,7%	0%	0%

Tabla 7.18. Evaluación de las heurísticas usadas si se eliminan ambas restricciones, R1 y R2

La evaluación indica netamente que la reducción R2 es justificada ya que prácticamente no se logra desarrollar el proceso de desambiguación. La evidencia nos respalda en asumir la restricción R2 en nuestro método.

Conclusión. En este experimento, hemos podido comprobar que las dos reducciones asumidas en nuestra estrategia de desambiguación son idóneas. La evidencia empírica es bien clara en cuanto a la validez de la restricción R1 (la independencia entre los patrones léxico-sintácticos y la oración en la asignación del sentido) y menos firme respecto a la validez de la restricción R2 (la independencia entre los patrones léxico-sintácticos en la asignación del sentido). Aunque en la presente tesis asumimos esta

última restricción, creemos oportuna una investigación más amplia al respecto, como por ejemplo su comprobación para todas las heurísticas relacionadas con los patrones (ME_PAR, PC_PAR, ME_SINT, PC_SINT).

7.3 Experimentos para el refinamiento del método

Objetivo principal. En el primer grupo de experimentos hemos verificado elementos básicos de nuestro método, obteniendo evidencias a favor de las decisiones sobre la construcción de nuestra propuesta. A continuación, presentamos una serie de pruebas que hemos desarrollado para refinar el método, con el propósito de alcanzar una alta fiabilidad en la asignación de sentidos. En concreto, la experimentación se ha orientado hacia los siguientes objetivos:

- evaluación de las heurísticas individuales.
- establecer los parámetros del método:
    - extracción patrones;
    - filtrado de los conjuntos de palabras asociadas a la ocurrencia ambigua;
    - combinación de las heurísticas relacionadas a los patrones;
    - combinación de las heurísticas relacionadas a la oración;
    - combinación de las heurísticas de ambos grupos;
    - aplicación óptima de la PC.

En la tabla 7.19. presentamos los experimentos de este segundo grupo, con sus características.

Cuestión estudiada	Parámetros		Experimento
Identificación de patrones léxico-sintácticos	Estructura		7a
	Filtros	Frecuencia (F1)	7b
		Número de substitutos en el marco del patrón (F2)	7 c
		F1 + F2	7 d
Heurísticas individuales Combinación de heurísticas	Cálculo de las respuestas de las heurísticas individuales		8a
	Combinación de las heurísticas relacionadas con los patrones		8b
	Combinación de las heurísticas relacionadas con la oración		8c
	Combinación de las heurísticas relacionadas con los patrones y con la oración		8d
Modalidad de uso de los Discriminadores de Sentido	Número DS	Respuestas individuales (unidades) + combinación posterior	9a
		Respuestas individuales (porcentajes) + combinación posterior	9b
		Combinación directa	9c
	Frecuencia DS	Respuestas individuales (unidades) + combinación posterior	9d
		Respuestas individuales (porcentajes) + combinación posterior	9e
		Combinación directa	9f

Tabla 7.19. Experimentos para el refinamiento del método

EXPERIMENTO 7

Nos hemos detenido en este experimento en la mejora del proceso previo a la desambiguación, o sea la identificación de los patrones léxico-sintácticos en que una ocurrencia ambigua participa. Como hemos resaltado previamente (apartado 6.6.1.), esta operación es fundamental ante todo para la cobertura del método, pero también condiciona la calidad de la asignación de sentidos. La delimitación de falsos patrones puede introducir "ruido", llevando a asignaciones incorrectas de sentido. Nuestra opción ha sido renunciar a tratar los casos difíciles, como las secuencias [N ADJ/VPART PREP N] (cf. apartado 6.7.1.). En otras palabras, hemos sacrificado la cobertura a favor de la precisión. Los criterios investigados para la delimitación de los patrones son los ya utilizados en precedencia (estructura y

frecuencia) y además hemos analizado un criterio más, el número de sustitutos que la palabra ambigua pueda tener dentro del patrón.

EXPERIMENTO 7a
En este experimento, nos hemos propuesto evaluar la calidad de la identificación de los patrones en base del criterio estructural, o sea mediante el uso de los patrones básicos y de los esquemas de búsqueda. Hemos procedido a dos ampliaciones sucesivas, tanto de los patrones básicos como de los esquemas de búsqueda.

Así, inicialmente hemos usado sólo los dos patrones básicos:

[N ADJ]
[N PREP N]

y los esquemas de búsqueda:

[N ADJ (CONJ* ADJ)]
[N (ADJ) PREP* (DET) (ADJ) N],

con restricciones sobre el elemento relacional: CONJ* = {*y, o*}, PREP* = {*de*}. En este caso, se obtienen 545 patrones.

Posteriormente, hemos ampliado los patrones a los siguientes (cf. apartado 6.7.1.):

[N_1, N_2]	[N_1, N_2]
[N_1 CONJ* N_2]	[N_1 CONJ* N_2]
[N ADJ]	[ADJ N]
[N VPART]	[VPART N]
[N_1 PREP N_2]	[N_1 PREP N_2]

mientras que los esquemas de búsqueda se han ampliado a los siguientes:

[N1 ((ADV) ADJ/VPART) , (DET) N2]
[N1 , (DET) ((ADV) ADJ/VPART) N2]
[N1 ((ADV) ADJ/VPART) CONJ* (DET) N2]
[N1 CONJ* (DET) ((ADV) ADJ/VPART) N2]
[N1 PREP (DET1) ((ADV2) ADJ2/VPART2) N2]
[N ((ADV) ADJ/VPART)]
[N ((ADV2) ADJ2/VPART2 CONJ*) (ADV1) ADJ1/VPART1)]
[ADJ1/VPART1 (CONJ* (ADV2) ADJ2/VPART2) N],

donde las unidades entre paréntesis son opcionales y las separadas por una barra son alternativas para una posición. Con esta ampliación, se identifican 640 patrones en el corpus de prueba de Senseval-2. Finalmente, hemos ampliado los esquemas de búsqueda incorporando las posibles secuencias con dos adverbios y dos determinantes:

[N1 (((ADV) ADV) ADJ/VPART) , (DET (DET)) N2]
[N1 , (DET (DET)) (((ADV) ADV) ADJ/VPART) N2]
[N1 (((ADV) ADV) ADJ/VPART) CONJ* (DET (DET)) N2]
[N1 CONJ* (DET (DET)) (((ADV) ADV) ADJ/VPART) N2]
[N1 PREP (DET (DET)) (((ADV) ADV) ADJ/VPART) N2]
[N (((ADV) ADV) ADJ1/VPART1 (CONJ* (((ADV) ADV) ADJ2/VPART2))]
[ADJ/VPART (CONJ* ((ADV) ADV) ADJ/VPART) N]
[N (ADV) ADJ (ADV) ADJ/VPART].

Esta última ampliación nos permite extraer 803 patrones en el mismo muestrario. Como se puede constatar, las ampliaciones operadas implican un incremento sustancial del número de patrones, de 545 a 640 y luego a 803. La supervisión manual del proceso indica que el incremento no perjudica la calidad en la delimitación de los patrones.

EXPERIMENTO 7b
Sobre los patrones identificados en la segunda y la tercera serie de las tres descritas previamente en el experimento 7a, hemos aplicado un filtro de frecuencia mínima 2 en un corpus. Para los 640 patrones adquiridos en la segunda serie, hemos contado los que se repiten en el corpus *LEXESP*, mientras que de los 803 patrones de la tercera serie hemos buscado los que se repiten en el corpus *EFE*. Los resultados se presentan en la tabla 7.20.

Patrones identificados	640	
Patrones filtrados (frecuencia ≥ 2 en *LEXESP*)	163	25,40%

Patrones identificados	803	
Patrones filtrados (frecuencia ≥ 2 en EFE)	335	41,71%

Tabla 7.20. Filtro de frecuencia (f ≥ 2) sobre patrones

Como era previsible, la dimensión del corpus usado para identificar los patrones repetidos influye drásticamente en el porcentaje de patrones que se repiten. Para los últimos 803 patrones, hemos comprobado además cómo repercute el filtro sobre la cobertura de las ocurrencias de Senseval-2 (tabla 7.21.).

	Patrones	Ocurrencias	Cobertura absoluta de las ocurrencias
Total identificados	803	750	750 / 799 = 93,86%
Filtrados F1 (frecuencia ≥ 2 en EFE)	335 335 / 803 = 41,71%	307 307 / 750 = 40,93%	307 / 799 = 38,42%

Tabla 7.21. Filtro de frecuencia (f ≥ 2) sobre los patrones. Incidencia sobre la cobertura de las ocurrencias

Por otra parte, sobre los patrones repetidos de la segunda serie (163 sobre 640), hemos evaluado manualmente la corrección de los patrones filtrados. La evaluación confirma que la condición de la repetición en el corpus aporta una mejora sustancial en la delimitación de los patrones, permitiendo una muy buena precisión en la extracción de patrones léxico-sintácticos (tabla 7.22.).

Patrones filtrados (frecuencia ≥ 2 en *LEXESP*)	163	
Patrones correctos	157	96,83%

Tabla 7.22. Evaluación de la calidad de los patrones repetidos en el corpus

Sin embargo, el análisis manual muestra que la aplicación del filtro de frecuencia, en el sentido de guardar exclusivamente los patrones que se repiten en el corpus, no es de por sí suficiente para la identificación de los patrones. Por una parte, algunos patrones son débiles semánticamente, es decir no delimitan lo suficiente el significado de la palabra ambigua. Por lo tanto, una solución sería imponer otro filtro más sobre los patrones. Por otra parte, hay numerosos patrones que aparecen una sola vez en el corpus y sin embargo corresponden a una relación sintáctica. En otras palabras, el filtro de frecuencia es demasiado estricto por escasez de datos; vemos necesario un filtro complementario para los patrones con frecuencia uno, que suavice el filtro actualmente usado. La primera posibilidad se ha estudiado en el experimento 7c y la segunda es objeto de una investigación en curso (cf. capítulo 8).

EXPERIMENTO 7c
Para identificar patrones que realmente son informativos sobre la palabra ambigua hemos impuesto la condición de que no haya más de 1000 sustitutos para la palabra ambigua dentro del patrón. Hemos considerado que el número de palabras que pueden ocupar la posición de la palabra ambigua dentro del patrón es un índice del grado de restricción que el patrón impone sobre una palabra que ocupe aquella posición, incluso la palabra por desambiguar. En la tabla 7.23., presentamos los resultados de este filtro, en términos de cobertura del total de patrones identificados y cobertura de las ocurrencias de prueba en Senseval-2.

Patrones		Ocurrencias correspondientes	Cobertura absoluta de las ocurrencias
Total identificados	803	750	750 / 799 = 93,86%
Filtrados F2 (≤1000 sustitutos en EFE)	367 367 / 803 = 45,70%	316 316 / 750 = 42,13%	316 / 799 = 39,42%

Tabla 7.23. Filtro de número de sustitutos (N ≤ 1000) sobre los patrones. Incidencia sobre la cobertura de las ocurrencias

El análisis manual indica que hay muchos que aportan información útil sobre la palabra ambigua, como por ejemplo, [*órgano*-N *de*-PREP *gobierno*-N] u [*órgano*-N *administrativo*-ADJ]. La restricción no es, por la tanto, adecuada.

EXPERIMENTO 7d
En este experimento, hemos aplicado ambos filtros F1 (f ≥ 2) y F2 (N ≤ 1000) sobre los 803 patrones obtenidos en la última serie. Hemos obtenido 300 patrones, 205 disjuntos. Recogemos en la tabla 7.24 la reducción respecto al total de patrones identificados, en términos de patrones y de ocurrencias.

	Patrones	Ocurrencias correspondientes	Cobertura absoluta de las ocurrencias
Total identificados	803	750	750 / 799 = 93,86%
Filtrados F1 +F2	300 300 / 803 = 37,35%	267 267 / 750 = 35,60%	267 / 799 = 33,41%

Tabla 7.24. Filtro de número de sustitutos (N ≤ 1000) sobre los patrones. Incidencia sobre la cobertura de las ocurrencias

Como se puede observar, la combinación entre ambos filtros es excesivamente fuerte y reduce drásticamente el número de patrones (de 803 a 300). Se eliminan muchos patrones que son correctos y altamente informativos sobre el significado de la palabra ambigua.

Sobre la operación de delimitación de los patrones léxico-sintácticos de una ocurrencia queda una cuestión abierta (cf. capítulo 8). *Hemos seguido nuestra investigación con los 300 patrones obtenidos en base a los primeros dos filtros:* frecuencia mínima dos para los patrones y número máximo de mil para los sustitutos de la palabra ambigua dentro del patrón. Este conjunto relativamente limitado nos ha permitido un estudio más a fondo de los diferentes aspectos de nuestro método. En los futuros estudios renunciaremos a este filtro combinado.

EXPERIMENTO 8
Objetivo: Este experimento investiga las heurísticas desde dos perspectivas: la modalidad usada para su evaluación (prueba 7a) y la modalidad usada para su combinación (pruebas 7b-d). Se estudian las heurísticas introducidas en el capítulo 6:

- las heurísticas relacionadas a los patrones (grupo I): PC_PAR, ME_PAR, PC_SINT, ME_SINT;
- las heurísticas relacionadas a la oración (grupo II): PC_OR, ME_OR

Parámetros: En este experimento se usan, como conjuntos de información, PAT, SINT y OR. Los primeros dos están filtrados (PAT a las primeras 20 palabras más frecuentes y SINT a las primeras 10 palabras más frecuentes), mientras que OR no está filtrado.

EXPERIMENTO 8a
Para la evaluación de las heurísticas, hemos delimitado tres modalidades de contar las respuestas correctas de las heurísticas:

1. Se aceptan como respuesta por parte de una heurística uno o dos sentidos. Si hay más sentidos propuestos, se considera que no hay respuesta.
2. Se aceptan como respuesta por parte de una heurística hasta el 50% del número de sentidos de la palabra. Es decir, se acepta: un sentido para las palabras con dos o tres sentidos; dos sentidos para las palabras con cuatro o cinco sentidos; tres sentidos para las palabras con seis o siete sentidos. Si hay más sentidos propuestos, se considera que no hay respuesta.
3. Se aceptan como respuesta por parte de una heurística todos los sentidos propuestos.

En la tabla 7.25 registramos los resultados obtenidos en estas tres evaluaciones.

Heurística	Evaluación 1: 1 o 2 sentidos		Evaluación 2: ≤ 50% del número de sentidos de la palabra		Evaluación 3: todos los sentidos	
	Precisión	Clasificación	Precisión	Clasificación	Precisión	Clasificación
ME_PAR	41,47%	3º	35,45%	3º	43,68%	4º
ME_SINT	38,09%	4º	32,14%	4º	44,13%	3º
PC_PAR	58,92%	2º	58,92%	2º	65,03%	2º
PC_SINT	80%	1º	68,33%	1º	70,58%	1º
ME_OR	-	-	28,93%	2º	46,76%	2º
PC_OR	-	-	62,74%	1º	46,82%	1º

Tabla 7.25. Heurísticas individuales
(tres evaluaciones)

La comparación de las tres evaluaciones nos lleva a las conclusiones que presentamos a continuación.

- Las evaluaciones proponen prácticamente la misma clasificación de las heurísticas individuales, dentro del grupo de las heurísticas relacionadas con los patrones y entre las heurísticas relacionadas con la oración.
- De las tres evaluaciones, la segunda es la más estricta y la tercera, la menos estricta, según reflejan los resultados.
- En todas las evaluaciones, el algoritmo Prueba de Conmutabilidad obtiene mejor precisión relativa que la Marca de Especificidad.
- El experimento permite otras conclusiones que confirman observaciones anteriores. La comparación entre las heurísticas que usan el conjunto de palabras frecuentemente co-ocurrentes con el patrón (SINT) y las heurísticas que usan las palabras de la oración (OR) indica que se obtienen mejores resultados con el primer conjunto usando cualquiera de los dos algoritmos. Se confirma, por lo tanto, nuestra opción de usar el conjunto de palabras coocurrentes con el patrón. Por otra parte, la comparación de los algoritmos indica que la Prueba de Conmutabilidad obtiene mejor precisión que la Marca de Especificidad.

EXPERIMENTO 8b
En la evaluación de las heurísticas relacionadas con los patrones, hemos seguido primero la modalidad de considerar las respuestas de las heurísticas individuales que corresponde a la *evaluación 2* definida

en el experimento 8a. Es decir, se aceptan las respuestas por parte de una heurística que contiene hasta el 50% del número de sentidos de la palabra por desambiguar.

Para la combinación de las respuestas que proceden de heurísticas relacionadas con un patrón, hemos definido las modalidades A, B, C, tal como se detalla a continuación. Mencionamos que se pueden tener una, dos, tres o cuatro respuestas por parte de las cuatro heurísticas relacionadas con un patrón.

Combinación A. Las respuestas se consideran como votantes iguales y se trabaja con una selección estricta del sentido, de tal manera que el sentido mayoritario en 2/2 respuestas, 3/3 respuestas, 4/4 o 3/4 respuestas. En otras palabras, se toma el 75% como límite inferior de votación para la selección de un sentido.

Combinación B. Las respuestas se consideran como votantes iguales y se trabaja con una selección relajada del sentido, de la manera siguiente:

- Si no hay ninguna respuesta o sólo una por parte de las heurísticas, entonces no se considera ninguna asignación de sentido asociada a los patrones.
- Si hay dos respuestas, entonces se considera el sentido mayoritario en estas respuestas.
- Cuando hay tres respuestas, el sentido correcto será:
  a) el sentido mayoritario en las tres respuestas;
  b) el sentido mayoritario en dos respuestas y también entre los sentidos primeros, con voto igual, de la tercera respuesta (lo que significa que el límite inferior de votación para la selección de un sentido es en este caso de 62,5%);
  c) el sentido mayoritario en una respuesta y también entre los sentidos primeros con voto igual en las dos otras respuestas (el límite inferior de votación para la selección de un sentido es en este caso del 55,55%).
- Cuando hay cuatro respuestas, el sentido correcto será:
  a) el sentido mayoritario en cuatro respuestas;
  b) el sentido mayoritario en tres respuestas;
  c) el sentido mayoritario en dos respuestas y también entre los sentidos primeros con voto igual de una tercera respuesta (es decir, el límite inferior de votación para la selección de un sentido es en este caso de 58,33%).

Combinación C. Se elige el sentido según el voto obtenido en orden decreciente de la precisión de las cuatro heurísticas, es decir:

1º PC_SINT;
2º PC_PAR;
3º ME_PAR;
4º ME_SINT.

Como resultado de la combinación de las heurísticas según cualquiera de las modalidades previamente definidas se obtienen patrones en los que la palabra focalizada tiene asignado un sentido. En otras palabras, se adquieren patrones etiquetados con sentidos para la palabra de partida. Resumimos en la tabla 7.26. los resultados obtenidos en estas combinaciones. La evaluación de la asignación de sentido es manual.

	Patrones etiquetados	Patrones con etiquetado correcto	Precisión
Combinación A	30	29	96,66%
Combinación B	62	52	83,87%
Combinación C	318	169	53,14%

Tabla 7.26. Combinación de las heurísticas relacionadas a los patrones (evaluación 2)

Según los resultados, la modalidad más precisa es la combinación A. En este caso, se desambiguan 30 ocurrencias de patrones, de las cuales 29 son correctamente resueltas, o sea se obtiene una precisión de 96,66%. En término de patrones disjuntos, se trata de 27 patrones desambiguados, de los cuales 26 correctamente, lo que significa una precisión de 96,29%. Enumeramos estos patrones etiquetados con alta precisión en la tabla 7.27.

Núm.	Patrón	Palabra por desambiguar	Sentido asignado	Sentido correcto
1	*autoridad social*	*autoridad*	1	1
2	*autoridad soviética*	*autoridad*	1	1
3	*autoridad teológica*	*autoridad*	4	4
4	*opinión de autoridad*	*autoridad*	1	1
5	*grano en cara*	*grano*	3	2
6	*naturaleza de función*	*naturaleza*	5	5
7	*Éxito de operación*	*operación*	6	6
8	*informe de órgano*	*órgano*	2	2
9	*partido burgués*	*partido*	1	1
10	*partido vencedor*	*partido*	1	1
11	*partido socialista*	*partido*	1	1
12	*partido de derecha*	*partido*	1	1
13	*partido vuelta*	*partido*	2	2
14	*partido firmante*	*partido*	1	1
15	*partido abertzale*	*partido*	1	1
16	*partido parlamentario*	*partido*	1	1
17	*partido de ida*	*partido*	2	2
18	*partido diferente*	*partido*	1	1
19	*Sigla de partido*	*partido*	1	1
20	*Financiación de partido*	*partido*	1	1
21	*Líder de partido*	*partido*	1	1
22	*personalidad de partido*	*partido*	1	1
23	*dirigente de partido* (4 ocurrencias)	*partido*	1	1
24	*Seno de partido*	*partido*	1	1
25	*Sigla de partido*	*partido*	1	1
26	*racha de partido*	*partido*	2	2
27	*situación de partido*	*partido*	2	2

Tabla 7.27. Patrones etiquetados con una precisión de 96%

Hemos evaluado la combinación de las heurísticas relacionadas a los patrones también en relación con las ocurrencias. Los resultados difieren ligeramente con respecto a la evaluación anterior realizada con respecto a los patrones, debido a que una misma ocurrencia puede tener más patrones con respuestas. En este caso, se toma como respuesta el sentido más votado entre las propuestas por parte de los diferentes patrones; si hay igualdad no se toma ninguna decisión y no se da una respuesta (tabla 7.28.).

Heurística	Precisión relativa	Precisión absoluta	Cobertura
ME_PAR	35,45%	15,40%	43,45%
ME_SINT	32,14%	7,84%	24,41%
PC_PAR	58,92%	14,38%	24,41%
PC_SINT	68,33%	11,91%	17,44%
Combinación A	92,59%	3,63%	3,92%
Combinación B	82,45%	6,81%	8,28%
Combinación C	52,50%	21,36%	40,69%

Tabla 7.28. Heurísticas relacionadas a los patrones y sus combinaciones
(evaluación respecto a las ocurrencias)

Conclusión. El experimento 8b indica que se pueden obtener patrones etiquetados con alta fiabilidad, del 96% aproximadamente, para un patrón sobre diez de los patrones identificados (30 de 300).

EXPERIMENTO 8b'
Hemos comprobado la combinación de las heurísticas relacionadas con los patrones también en el contexto en que se considera como respuesta de una heurística el sentido mayoritario y, en caso de igualdad entre varios sentidos en la primera posición, se guardan todos (*evaluación 3*).

Para la combinación de las respuestas de las diferentes heurísticas, hemos probado esta vez las siguientes modalidades:
Combinación D. Se elige el sentido *único* propuesto por al menos el 50% de las heurísticas y que se halla también entre los sentidos propuestos por otra heurística.
Combinación E. Se elige el sentido *único* propuesto por al menos el 50% de las heurísticas.
Combinación F. Se elige el sentido *único* propuesto por al menos el 75% de las heurísticas.
Combinación G. Se elige el sentido *único* propuesto por el 100% de las heurísticas.
La evaluación de las combinaciones lleva a los resultados que recogemos en la tabla 7.29.

Combinación	Precisión
Combinación D	70,37%
Combinación E	72,72%
Combinación F	75,86%
Combinación G	82,29%

Tabla 7.29. Combinación de las heurísticas relacionadas a los patrones (evaluación 3)

Conclusión (experimento 8b')**.** Las diferentes combinaciones llevan a una precisión media, insuficiente frente a nuestro objetivo de alcanzar una precisión próxima al 100%. Aunque en el experimento *8a* las heurísticas individuales habían obtenido precisión superior en la evaluación 3 respecto a la evaluación 2, su combinación en las modalidades con un único sentido aquí adoptadas hacen bajar los resultados de las combinaciones respecto a las combinaciones con la evaluación 2 previamente presentadas. Parece, por lo tanto, que la consideración exclusivamente de un sentido único mayoritario no es adecuada. La alternativa que resulta de aquí es la consideración de todos los sentidos de una heurística y no sólo del sentido o los sentidos mayoritarios.

EXPERIMENTO 8c
Para la combinación de las heurísticas relacionadas con la oración, aplicamos las dos heurísticas en el orden de la precisión medida en el experimento *8a*: primero PC_OR, luego ME_OR. Presentamos la evaluación en la tabla 7.30.

Heurísticas	Total respuestas validables	Respuestas validables correctas	Precisión
ME_OR	667	193	28,93%
PC_OR	51	32	62,74%
PC_OR+ME_OR	667	211	31,63%

Tabla 7.30. Combinación de las heurísticas relacionadas a la oración

Conclusión. Aunque la cobertura es baja, la heurística PC_OR que consiste en la aplicación de la Prueba de Conmutabilidad al conjunto de nombres de la oración mejora la desambiguación obtenida mediante la heurística ME_OR en que se aplica la Marca de Especificidad al mismo conjunto.

EXPERIMENTO 8d
Objetivo: Analizamos en este experimento la combinación entre el grupo I de heurísticas relacionadas con los patrones, es decir, con los patrones etiquetados, y el grupo II de heurísticas relacionadas con la oración. Presentamos los resultados de la evaluación en la tabla 7.31.

Heurísticas	Precisión relativa	Precisión absoluta	Cobertura
I (combinación A)	92,59%	3,63%	3,92%
I (combinación B)	82,45%	6,81%	8,28%
I (combinación C)	52,50%	21,36%	40,69%
II	31,63%	30,66%	96,94%
I (combinación A) + II	33,28%	32,26%	96,94%
I (combinación B) + II	35,02%	34,01%	97,09%
I (combinación C) + II	40,92%	39,97%	97,67%

Tabla 7.31. Combinación entre las heurísticas relacionadas con los patrones y las heurísticas relacionadas con la oración

Conclusión. La principal evidencia de este experimento revela que los patrones etiquetados mejoran los resultados que se obtienen usando exclusivamente la oración. La mejora varía según el grupo de patrones etiquetados que se use, o sea los que corresponden a la combinación A, B o C. El número de estos patrones es inversamente proporcional con la precisión del etiquetado con sentidos (cf. tabla 7.26.). La evaluación indica que la mejora es superior cuando se usan muchos patrones de precisión inferior (los de la combinación C) respecto a cuando se usan pocos patrones pero de precisión alta (los de la combinación A). La causa reside en que los patrones etiquetados obtenidos, incluso los del grupo C, superan en precisión relativa el grupo II de heurísticas relacionadas a la oración.

EXPERIMENTO 9
Objetivo: Esta experimentación está orientada hacia la identificación de una modalidad óptima para explotar los Discriminadores de Sentido (DS) para la DSA mediante el algoritmo Prueba de Conmutabilidad. Con este propósito, procedemos a un estudio detallado del uso de los Discriminadores de Sentido en la asignación de sentido, bajo varios aspectos que detallamos a continuación como parámetros a, b y c.

a) En los experimentos precedentes, los Discriminadores de Sentido se han considerado sólo como marcas de posibles asignaciones de sentido. Es decir, la simple *aparición* de algún discriminador dentro del patrón en la posición de la palabra ambigua señala la posibilidad de que la palabra tenga el sentido al que corresponde el discriminador. Sin embargo, la Prueba de Conmutabilidad se puede aplicar también considerando los DS a otros dos "niveles":

- en cuanto al *número* de discriminadores para cada sentido que pueden sustituir la palabra en el patrón;
- en cuanto a la *frecuencia* de los discriminadores para cada sentido que pueden sustituir la palabra en el patrón.

En este experimento nos centramos en el uso de los DS de estas dos maneras.

b) Por otra parte, interesa la modalidad en que se calcula la respuesta de una heurística en base a la votación que los sentidos reciben por parte de los DS (bajo una de las dos formas previamente descritas: como número o como frecuencia).
A lo largo de todo este experimento, la heurística elige como respuesta el sentido con mayoría de votos, que denominaremos sentido mayoritario. En caso de igualdad entre varios sentidos, se guardan todos, sin restricciones sobre el número que una heurística propone en el momento en que se evalúa su respuesta. Es decir, seguimos la modalidad que corresponde a la *evaluación 3* introducida en el experimento 8. Hemos investigado dos posibilidades de tomar en cuenta los votos por parte de los DS para los sentidos:

- *Unidades*. Se considera cada discriminador hallado para un sentido como una unidad en la votación que este sentido recibe dentro de la aplicación de la heurística.
- *Porcentaje.* Los votos de los DS para cada sentido se expresan en porcentajes sobre el total de votos que han recibido todos los sentidos. Se guardan sólo los sentidos que superan un umbral establecido de porcentajes.

c) Finalmente, estudiamos la combinación de las heurísticas individuales para la asignación del sentido final para la palabra dentro del patrón. Hemos utilizado dos estrategias de combinación:

- respuestas individuales, con propuestas de sentido por parte de cada heurística, y la *combinación posterior* de estas respuestas;
- *combinación directa*, o sea sumar los votos de todas las heurísticas y sobre esta suma considerar los votos que ha recibido cada uno de los sentidos.

Para cada una de estas dos estrategias de combinación se han analizado diferentes formas variadas. Aunque en el presente experimento nuestra atención se dirige hacia la Prueba de Conmutabilidad, también estudiamos su combinación con la Marca de Especificidad. Así, se utilizarán cuatro heurísticas para la desambiguación de la ocurrencia ambigua dentro del patrón: PC_PAR, PC_SINT, ME_PAR, ME_SINT. En la tabla 7.33., sintetizamos la organización de este experimento; en ella se muestran los parámetros anteriormente mencionados en relación con el uso de los DS para la asignación de sentidos.

Parámetros		Experimento
Número DS	Respuestas individuales (unidades) + combinación posterior	9a
	Respuestas individuales (porcentajes) + combinación posterior	9b
	Combinación directa	9c
Frecuencia DS	Respuestas individuales (unidades) + combinación posterior	9d
	Respuestas individuales (porcentajes) + combinación posterior	9e
	Combinación directa	9f

Tabla 7.33. Experimentos sobre el uso de los Discriminadores de Sentido

Parámetros: En este experimento hemos operado sobre los conjuntos PAR y SINT no filtrados. En este experimento se ha cambiado la variante de *EWN*, ya que se utilizará el WEI español y no la variante vinculada a *WN 1.5*.

EXPERIMENTO 9a

En este experimento se consideran las respuestas individuales en forma de unidades y su posterior combinación, es decir se cuenta el número de discriminadores que cada sentido recibe de una heurística. Cada discriminador hallado para un sentido se considera como un voto a favor del sentido. Se considera como respuesta sólo el primer sentido, o sea el sentido para el que se han encontrado más discriminadores; si hay más con un número igual de votos (de discriminadores), se toman en consideración todos los primeros sentidos iguales.

La evaluación de las heurísticas individuales lleva a los resultados de la tabla 7.34.

Heurística	Precisión relativa
ME_PAR	44,17%
ME_SINT	45,81%
PC_PAR	52,67%
PC_SINT	53,17%

Tabla 7.34. Evaluación de las heurísticas individuales
(número de DS, votos como unidades)

Para la combinación de las respuestas de las diferentes heurísticas, probamos las siguientes variantes:

Combinación A. Se elige el sentido propuesto por al menos el 75% de las heurísticas y como mínimo por dos heurísticas.

Combinación B. Se elige el sentido propuesto por el 100% de las heurísticas y como mínimo por dos heurísticas.
Combinación C. Se elige el sentido que es respuesta única de al menos el 75% de las heurísticas.
Combinación D. Se elige el sentido que es respuesta única del 100% de las heurísticas, con la condición de que haya respuestas de cómo mínimo dos heurísticas.
Combinación E. Se consideran sólo las heurísticas PC_par y PC_sint y se hace la intersección de sus respuestas.
Combinación F. Se consideran sólo las heurísticas PC_par y PC_sint y se hace la intersección de sus respuestas, considerando sólo las respuestas únicas.

Presentamos la evaluación de estas combinaciones en la tabla 7.35.

Combinación	Precisión
Combinación A	48,71%
Combinación B	53,33%
Combinación C	55%
Combinación D	53,12%
Combinación E	53,65%

Tabla 7.35. Evaluación de las la combinación posterior de heurísticas
(heurísticas basadas en número de DS, votos como unidades)

EXPERIMENTO 9b
Para el cálculo de la respuesta de una heurística, se suman los discriminadores hallados para todos los sentidos y se calcula el porcentaje de votación para cada sentido. Como respuesta de una heurística, se consideran aquellos sentidos que han obtenido un porcentaje de votación superior a un umbral mínimo preestablecido. Hemos probado las dos variantes que presentamos a continuación.

Respuesta 1. Se considera como respuesta el sentido que ha recibido al menos el 75% de los votos.
Respuesta 2. Se considera como respuesta el sentido que ha recibido al menos el 75% de los votos y además, si tiene el 100%, que se haya encontrado más de un discriminador para el sentido.

Los resultados obtenidos en este experimento se muestran en la tabla 7.36.

Heurística		Precisión
ME_PAR		38,57%
ME_SINT		42,18%
PC_PAR	Respuesta 1	56,60%
	Respuesta 2	58,82%
PC_SINT	Respuesta 1	34,28%
	Respuesta 2	44,44%

Tabla 7.36. Evaluación de las heurísticas individuales
(número de DS, votos como porcentajes)

Para la combinación de las heurísticas individuales, hemos usado dos fórmulas:

Combinación A. Se guarda el sentido que recibe al menos el 75% de los votos de las heurísticas con respuestas
Combinación B. Se guarda el sentido que recibe al menos el 75% de los votos de las heurísticas con respuestas y en las heurísticas con PC que se haya encontrado más de un discriminador para el sentido.

Presentamos la evaluación de ambas combinaciones en la tabla 7.37.

Combinación	Precisión
Combinación A	62,22%
Combinación B	58,33%

Tabla 7.37. Evaluación de las la combinación posterior de heurísticas (heurísticas basadas en número de DS, votos como porcentajes)

EXPERIMENTO 9c
En esta prueba, comprobamos la combinación directa de los votos por parte de todas las heurísticas. Para ello, se suman los DS identificados como sustitutos de la palabra ambigua para cada sentido y luego se elige el sentido final según una de las modalidades siguientes:

Respuesta 1. Se considera el sentido mayoritario en la suma de votos.
Respuesta 2. Sobre esta suma, se impone la condición de que el sentido tenga al menos el 75% de los votos.
Respuesta 3. Se impone la condición de que el sentido tenga al menos el 75% de los votos y la condición suplementaria de que se tiene que haber encontrado para él más de un discriminador.

La tabla 7.38. muestra los resultados obtenidos en la evaluación de estas modalidades de combinación directa de las heurísticas.

Respuesta	Precisión
Respuesta 1	42,78%
Respuesta 2	48,78%
Respuesta 3	61,29%

Tabla 7.38. Evaluación de las la combinación directa de heurísticas (heurísticas basadas en número de DS, votos como porcentajes)

Conclusión. Los resultados obtenido en los experimentos 9a, 9b, 9c son bajos, por lo tanto la explotación de los DS obtener el sentido de cada palabra no es adecuada.

Una posible solución es considerar la *frecuencia* de los discriminadores, es decir la cantidad de información que aporta cada discriminador a los sentidos. Para ello, los conjuntos de palabras PAR y SINT asociados a la palabra ambigua dentro del patrón se extraerán junto con la frecuencia de las palabras. El número de votos que recibe un sentido por parte de una heurística será la suma de las frecuencias que tienen los discriminadores hallados para el sentido dentro de los conjuntos de palabras asociadas a la palabra ambigua dentro del patrón (PAR y SINT). Hemos experimentado esta variante en 9d, 9e y 9f para la explotación de los DS.

EXPERIMENTO 9d
De manera similar a como se ha procedido en la prueba 9a, las respuestas individuales se calculan bajo la forma de unidades. El número de votos que cada sentido recibe es igual a la suma de las frecuencias de los DS hallados para el sentido en los conjuntos PAR y SINT. Se elige como sentido propuesto por

una heurística el que más votos tiene, o sea aquí para el cual se ha encontrado la suma de frecuencias más alta.

La evaluación de las heurísticas en estas condiciones es la que presentamos en la tabla 7.39.

Heurística	Precisión relativa
ME_PAR	41,45%
ME_SINT	43,33%
PC_PAR	41,05%
PC_SINT	43,10%

Tabla 7.39. Evaluación de las heurísticas individuales
(número de DS, votos como unidades)

Su combinación se ha hecho de tres maneras:

Combinación A. Se guarda el sentido que recibe al menos el 75% de los votos de las heurísticas con respuestas
Combinación B. Se guarda el sentido que recibe el 100% de los votos de las heurísticas con respuestas y además es el sentido único propuesto por las heurísticas.
Combinación C. Consiste en la combinación B a la cual se le añade la condición de que el sentido sea propuesto por todas las cuatro heurísticas.

Los resultados de la evaluación se recogen en la tabla 7.40.

Selección	Precisión
Combinación A	51,68%
Combinación B	55,55%
Combinación C	80%

Tabla 7.40. Evaluación de las la combinación posterior de heurísticas
(heurísticas basadas en frecuencia de DS, votos como unidades)

EXPERIMENTO 9e
Se consideran como respuesta de una heurística los sentidos que han obtenido un porcentaje de la votación por encima de un umbral mínimo preestablecido. Para calcular los votos que tiene un sentido, se calcula el porcentaje que representa la suma de las frecuencias que tienen los discriminadores hallados para este sentido sobre la suma de las frecuencias que tienen los discriminadores hallados para todos los sentidos. Posteriormente, se guardan como propuesta de una heurística los sentidos que superan el 75% de los votos.

En este caso, las heurísticas tienen la precisión relativa que se muestra en la tabla 7.41.

Heurística	Precisión relativa
ME_PAR	38,19%
ME_SINT	41,53%
PC_PAR	45,45%
PC_SINT	44,26%

Tabla 7.41. Evaluación de las heurísticas individuales
(frecuencia de DS, votos como porcentajes)

Para la combinación de las respuestas individuales, se han probado las dos variantes que enumeramos a continuación.

Combinación A. Se elige el sentido que recibe el 100% de los votos de las heurísticas con respuestas, si hay al menos dos heurísticas con respuestas.
Combinación B. Se guarda el sentido que recibe el 100% de los votos y además todas las cuatro heurísticas proponen el sentido.

Los resultados de la evaluación se recogen en la tabla 7.42.

Respuesta 8. Se considera el sentido que tiene al menos 95% votos y frecuencia de los discriminadores correspondientes hallados de al menos 2.
Respuesta 9. Se considera el sentido que tiene al menos 66,6% votos y frecuencia de los discriminadores correspondientes hallados de al menos 5.
Respuesta 10. Se considera el sentido que tiene al menos 75% votos y frecuencia de los discriminadores correspondientes hallados de al menos 5.
Respuesta 11. Se considera el sentido que tiene al menos 90% votos y frecuencia de los discriminadores correspondientes hallados de al menos 5.
Respuesta 12. Se considera el sentido que tiene al menos 95% votos y frecuencia de los discriminadores correspondientes hallados de al menos 5.
Respuesta 13. Se considera el sentido que tiene al menos 75% votos y frecuencia de los discriminadores correspondientes hallados de al menos 10.
Respuesta 14. Se considera el sentido que tiene al menos 95% votos y frecuencia de los discriminadores correspondientes hallados de al menos 10.

La tabla 7.44. muestra los resultados obtenidos en la evaluación de estas modalidades de combinación directa de las heurísticas. Hemos resaltado los casos en que se ha obtenido una precisión más alta en la desambiguación.

Variante	Total de respuestas	Respuestas correctas	Precisión
Respuesta 1	139	78	56,11%
Respuesta 2	125	71	56,80%
Respuesta 3	105	62	59,04%
Respuesta 4	93	56	60,21%
Respuesta 5	52	36	69,23%
Respuesta 6	37	32	84,48%
Respuesta 7	33	23	69,69%
Respuesta 8	24	21	87,5%
Respuesta 9	104	63	60,57%
Respuesta 10	-	-	50,43%
Respuesta 11	-	-	88,57%
Respuesta 12	18	17	94,44%
Respuesta 13	-	-	49,72%
Respuesta 14	18	17	94,44%

Tabla 7.44. Evaluación de las la combinación directa de heurísticas
(heurísticas basadas en frecuencia de DS)

En esta prueba hemos estudiado dos parámetros de la aplicación de los DS en esta variante (o sea, la combinación directa de los votos a nivel de frecuencia de los DS) para elegir un sentido como respuesta: el porcentaje mínimo de votos y la frecuencia de los discriminadores correspondientes hallados. Para un mejor análisis de la interacción de estos dos parámetros en la asignación de sentidos, en la tabla 7.45., sintetizamos los valores que hemos usado en la presente prueba, con los resultados obtenidos en cada caso.

		Frecuencia de los discriminadores			
		-	≥ 2	≥ 5	≥ 10
Porcentaje de votos	≥ 66,6%	56,11%	56,80%	60,57%	-
	≥ 75%	59,04%	60,21%	-	-
	≥ 90%	69,23%	84,48%	88,57%	-
	≥ 95%	69,69%	87,50%	94,44%	94,44%

Tabla 7.45. Evaluación de las la combinación directa de heurísticas
(heurísticas basadas en frecuencia de DS). Valores de los parámetros usados

Conclusiones de la prueba.
1. Se puede obtener desambiguación de alta calidad, cerca del 100%, usando sólo un patrón y las cuatro heurísticas correspondientes al patrón: ME_PAR, ME_SINT, PC_PAR, PC_SINT, cuando ambos conjuntos (paradigmático y sintagmático) no son filtrados.
2. La comparación de las trece modalidades aquí probadas revela una serie de aspectos relacionados con el uso de los DS, según se detalla a continuación.

- Un análisis contrastivo de las variantes 1, 3, 5, y 7 indica que el aumento del umbral mínimo de votación para la selección de un sentido como respuesta mejora la precisión, pero no suficientemente. Es decir, sólo el criterio del porcentaje de la votación no es suficiente (la primera columna en la tabla 7.45.).
- La comparación de las variantes 1 y 2, 3 y 2, 5 y 6, 7 y 8 y finalmente 9, 10, 11, 12, 13 y 14 muestra que el aumento del umbral mínimo de la frecuencia mejora la precisión sobre todo a partir de un nivel mínimo de votación alto. O sea no tiene mucho efecto cuando la elección del sentido se hace por encima del 66,6% o del 75% de la votación, pero sí cuando se consideran sentidos con al menos el 90% o el 95%. Lo que parece indicar que la frecuencia de los discriminadores es útil para la selección del sentido, pero como criterio complementario al del porcentaje de votos en la votación que los sentidos reciben.

En el presente experimento, la mejor variante de las probadas para la selección de un sentido en base de las cuatro heurísticas relacionadas a los patrones es la votación con un mínimo de 95% y con una frecuencia de los discriminadores de al menos 5 o 10 (se debe establecer en futuros experimentos), o sea las variantes 12 y 14.
En la variante 14, hemos obtenido 14 patrones etiquetados, de los cuales 13 son correctos, o sea con una precisión de 92,85%. Esto corresponde a 18 ocurrencias de patrones, de las cuales 17 están correctamente resueltas, es decir se alcanza una precisión de 94,44%. La tabla 7.46 muestra los resultados obtenidos de los 14 patrones anotados.

Núm.	Patrón	Palabra por desambiguar	Sentido asignado	Sentido correcto
1	*medida de gracia* (2 ocurrencias)	*gracia*	4	4
2	*Masa obrera*	*masa*	3	3
3	*fuerza de naturaleza*	*naturaleza*	2	2
4	*treintena de partido*	*partido*	3	2
5	*partido comunista* (2 ocurrencias)	*partido*	1	1
6	*partido burgués*	*partido*	1	1
7	*partido socialista*	*partido*	1	1
8	*partido firmante*	*partido*	1	1
9	*partido abertzale*	*partido*	1	1
10	*partido parlamentario*	*partido*	1	1
11	*personalidad de partido*	*partido*	1	1
12	*presidente de partido*	*partido*	1	1
13	*seno de partido*	*partido*	1	1
14	*dirigente de partido* (2 ocurrencias)	*partido*	1	1

Tabla 7.46. Grupo 3 de patrones etiquetados (precisión 92,85%)

Análisis. El patrón *treintena de partido*, etiquetado con un sentido erróneo, contiene un nombre con un valor de numeral colectivo (un cuantitativo). Se pone de manifiesto aquí la necesidad de imponer más restricciones en la extracción de los patrones o por lo menos en su explotación para la DSA. Entre los candidatos a ser eliminados son los patrones que contienen un cuantitativo, por ser éstos nada restrictivos sobre el otro nombre de contenido léxico del patrón léxico-sintáctico.

Síntesis de los experimentos 8 y 9. En la tabla 7.47. reunimos los tres grupos de patrones anotados obtenidos en los experimentos 8 y 9, junto con la frecuencia de estos patrones en el corpus EFE. Recordamos antes las condiciones en que se han obtenido cada uno de los tres conjuntos de patrones etiquetados:

- El grupo 1 de patrones se ha adquirido sobre los conjuntos PAR y SINT filtrados (las primeras veinte palabras y las primeras diez palabras, respectivamente), aceptando como máximo la mitad de los sentidos como respuesta de una heurística y eligiendo el sentido propuesto por al menos el 75% de las heurísticas.
- Los grupos 2 y 3 de patrones se han obtenido usando conjuntos PAR y SINT no filtrados, sin restricciones sobre el número de sentidos que la respuesta de una heurística pueda contener y usando los DS al nivel de frecuencia. En el caso del grupo 2, se ha elegido el sentido propuesto por todas las cuatro heurísticas (por lo tanto, también con el 100% de los votos) mediante la combinación posterior de las propuestas individuales. Los patrones del grupo 3 son resultado de la combinación directa de los votos de las heurísticas y corresponden a la elección del sentido con un mínimo de 95% de los votos y a una frecuencia de los DS de mínimo 10.

Núm.	Patrón	Grupo 1	Grupo 2	Grupo 3	Núm. ocurrencias en el corpus EFE
1	*opinión de autoridad*	+			23
2	*autoridad civil*	+			269
3	*autoridad social*	+			269
4	*autoridad soviética*	+			25
5	*autoridad teológica*	+			2
6	*grano en cara*	+			2
7	*naturaleza de función*	+			2
8	*éxito de operación*	+			43
9	*informe de órgano*	+			2
10	*partido burgués*	+		+	6
11	*partido vencedor*	+			18
12	*partido socialista*	+		+	225
13	*partido de derecha*	+			269
14	*partido de vuelta*	+			803
15	*partido firmante*	+	+	+	43
16	*partido abertzale*	+	+	+	6
17	*partido parlamentario*	+	+	+	56
18	*partido de ida*	+			849
19	*partido diferente*	+			17
20	*sigla de partido*	+			6
21	*financiación de partido*	+	+		184
22	*finanzas de partido*		+		29
23	*líder de partido*	+	+		263
24	*personalidad de partido*	+		+	2
25	*seno de partido*	+	+	+	33
26	*dirigente de partido*	+	+	+	225
27	*racha de partido*	+			23
28	*situación de partido*	+			20
29	*presidente de partido*	+		+	114
30	*partido comunista*			+	237
31	*partido ganado*		+		73
32	*treintena de partido*			+	23
33	*medida de gracia*			+	99
34	*masa obrera*			+	6
35	*fuerza de naturaleza*			+	14
	TOTAL				4173

Tabla 7.47. Patrones etiquetados obtenidos en la experimentación

Análisis. A continuación, exponemos algunas observaciones sobre los patrones etiquetados obtenidos en la experimentación.

1. En total, hemos obtenido un conjunto de 35 patrones con alta fiabilidad, superior al 90% (entre 92% y 100%), sobre los 300 patrones de partida. Esto significa que hemos etiquetado uno de cada diez patrones con que hemos trabajado.
2. Los patrones aparecen 4173 veces en el corpus EFE. La frecuencia total de las veintiuna palabras de Senseval-2 es de 196.611, lo que significa que con estos 35 patrones hemos cubierto el 2,12% de todas las apariciones de estas palabras en el corpus EFE (de más de 70 millones de palabras). Si nos limitamos a las palabras que participan en estos patrones (*autoridad, gracia, grano, masa, naturaleza, operación, órgano, partido*), los patrones etiquetados cubren el 2,82% de las 147,881 ocurrencias que estas palabras tienen en EFE. Por fin, para el nombre *partido*, los 23 patrones que lo contienen corresponden a 3715 apariciones suyas en EFE, o sea el 4,40% de todas sus 84305 ocurrencias. Estas cifras demuestran el potencial de nuestro método y ofrecen una imagen sobre la cobertura real del etiquetado resultante.
3. La comparación de los tres conjuntos de patrones etiquetados muestra que hay patrones en común, lo que significa que hay convergencia o paralelismo entre las variantes del método que se han probado a lo largo de la experimentación.
4. La confrontación de las condiciones que han propiciado la obtención de patrones de alta fiabilidad lleva a la observación de que es necesaria alguna información de tipo estadístico que se combine con la información lingüística para potenciar el rendimiento del método. Consideramos el uso de la frecuencia de los DS como la variante más idónea para alcanzar una alta fiabilidad en la asignación de sentido.

7.4 Evaluación parcial en el marco de Senseval-3

Nuestra participación en la tercera edición del ejercicio Senseval ha sido exclusivamente para nombres, dentro de un sistema de DSA que cubría igualmente los adjetivos y los verbos, a través de un método estadístico basado en Máxima Entropía, desarrollado en la Universidad de Alicante (Suárez y Palomar, 2004). Para la desambiguación de los nombres, hemos usado una variante simplificada de nuestro método, con la combinación de sólo dos heurísticas, ambas relacionadas con los patrones: PC_PAR_{2k} y PC_SINT_k. Detallamos a continuación las decisiones que hemos tomado en cada paso del método.

En el paso 1º, para la identificación de los patrones léxico-sintácticos, hemos aplicado el criterio estructural junto con un filtro de frecuencia de mínimo 5.

En el paso 2º, para la adquisición de información asociada a la ocurrencia ambigua, hemos usado EFE como corpus de búsqueda. De éste, hemos extraído información paradigmática bajo la forma del conjunto PAR_{2k} e información sintagmática bajo la forma del conjunto $SINT_k$. En ambos casos, el umbral de frecuencia mínima se ha fijado en 5.

Para la definición de las heurísticas, en el paso 3º, hemos usado el algoritmo de la Prueba de Conmutabilidad en su variante básica. Combinando este algoritmo con cada uno de los conjuntos PAR_{2k} y $SINT_k$, hemos obtenido dos heurísticas de DSA.

La combinación de las heurísticas individuales, en el paso 4º, la hemos hecho de la siguiente manera: para cada patrón léxico-sintáctico, se hace la intersección de las propuestas de sentido de las dos heurísticas; luego se eligió el sentido más propuesto entre los diferentes patrones de la ocurrencia; finalmente, si quedaban más sentidos propuestos, se daba preferencia a uno según el orden de los sentidos en *EWN* (su número).

Como muestrario de evaluación de nuestro sistema en Senseval-3, se puede considerar el corpus de prueba para los veintiún nombres en concurso: *arte, autoridad, bomba, canal, circuito, columna, corazón, corona, gracia, grano, hermano, letra, masa, mina, naturaleza, operación, órgano, partido, pasaje, programa, tabla*. Hemos presentado en el apartado 7.1. los datos de este conjunto de prueba.

Ofrecemos en la tabla 7.48. los resultados oficiales para nuestro método sobre nombres:

Palabra	Ocurrencias por tratar	Total respuestas	Respuestas correctas	Cobertura	Precisión relativa	Precisión absoluta
arte	121	0	0	0%	0%	0%
autoridad	**132**	**38**	**35**	**28,79%**	**92,11%**	**26,52%**
banda	114	0	0	0%	0%	0%
canal	**131**	**21**	**21**	**16,03%**	**100%**	**16,03%**
circuito	**132**	**3**	**1**	**1,52%**	**50%**	**0,76%**
columna	64	0	0	0%	0%	0%
corazón	62	0	0	0%	0%	0%
corona	64	0	0	0%	0%	0%
gracia	38	0	0	0%	0%	0%
grano	**61**	**2**	**0**	**3,28%**	**0%**	**0**
hermano	66	0	0	0%	0%	0%
letra	114	0	0	0%	0%	0%
masa	85	0	0	0%	0%	0%
mina	66	0	0	0%	0%	0%
naturaleza	128	0	0	0%	0%	0%
operación	66	0	0	0%	0%	0%
órgano	131	0	0	0%	0%	0%
partido	**66**	**17**	**14**	**25,76%**	**82,35%**	**21,21%**
pasaje	111	0	0	0%	0%	0%
programa	**133**	**26**	**23**	**19,55%**	**88,46%**	**17,29%**
tabla	64	0	0	0%	0%	0%
TOTAL	1949	106	94	5,44%	88,68%	4,82%

Tabla 7.48. Los resultados obtenidos en Senseval-3 (sólo nombres)

Cabe precisar que la implementación de las dos heurísticas ha sido sólo parcial, sobre trece palabras de las veintiuna en concurso. No se han tratado las palabras señaladas en color en la tabla 7.48. En la tabla 7.49., calculamos la precisión y la cobertura del sistema para las palabras tratadas.

Palabra	Ocurrencias por tratar	Total respuestas	Respuestas correctas	Cobertura	Precisión relativa	Precisión absoluta
autoridad	**132**	**38**	**35**	**28,79%**	**92,11%**	**26,52%**
canal	**131**	**21**	**21**	**16,03%**	**100%**	**16,03%**
circuito	**132**	**3**	**1**	**1,52%**	**50%**	**0,76%**
corona	64	0	0	0%	0%	0%
gracia	38	0	0	0%	0%	0%
grano	**61**	**2**	**0**	**3,28%**	**0%**	**0**
hermano	66	0	0	0%	0%	0%
masa	85	0	0	0%	0%	0%
naturaleza	128	0	0	0%	0%	0%
partido	**66**	**17**	**14**	**25,76%**	**82,35%**	**21,21%**
pasaje	111	0	0	0%	0%	0%
programa	**133**	**26**	**23**	**19,55%**	**88,46%**	**17,29%**
tabla	64	0	0	0%	0%	0%
TOTAL	1211	106	94	8,75%	88,68%	7,76%

Tabla 7.49. Los resultados obtenidos en Senseval-3, calculados sobre las palabras tratadas

De hecho, la variante del sistema presentada en Senseval-3 ha contestado para ocurrencias de las seis palabras siguientes: *autoridad, canal, circuito, grano, partido, programa.*
Para las 655 ocurrencias de las palabras que han tenido respuesta (es decir para las cuales no hemos obtenido solamente "0", o sea: *autoridad, canal, circuito, grano, partido, programa*), el cálculo de los resultados es el siguiente:

Cobertura = 106 / 655 = 16,18%
Precisión absoluta = 94 / 655 = 14,35%

Nos detenemos brevemente en el análisis de los resultados. Estimamos que la calidad de la desambiguación ha sido limitada, especialmente en la cobertura, por los siguientes factores:

a) el uso de un conjunto parcial de patrones básicos y de esquemas de búsqueda, que cubre sólo parcialmente la casuística de las relaciones sintácticas que el nombre puede tener, limitándose al marco de un sintagma nominal y no tratando relaciones nombre-verbo;

b) el grado de cobertura de la correspondencia entre los sentidos de referencia usados en Senseval-3 y los sentidos del WEI español (74,76% y 82,31% respectivamente, cf. tabla 7.3.);

c) la restricción excesiva para la frecuencia de los patrones y de los sustitutos del conjunto PAR (en ambos casos mínimo cinco)[225];

d) el uso de sólo dos heurísticas relacionadas a los patrones (PC_PAR y PC_SINT).

La evaluación pone de manifiesto, por una parte, la necesidad de relajar el filtro de frecuencia para los patrones y, por otra parte, la necesidad de considerar más heurísticas en el sistema de DSA.

Respecto de la precisión relativa, el sistema ha alcanzado la mejor precisión de los sistemas participantes para el español: el 88,86%. Alcanzamos, así, nuestro objetivo de obtener una desambiguación de alta fiabilidad. Se confirma, implícitamente, que nuestro enfoque a la DSA es válido y que permite una buena calidad de la desambiguación.

[225] Así, hay 80 patrones que tienen respuestas de sentido para el conjunto SINT (o sea la heurística PC_SINT), pero no para el conjunto PAR (o sea la heurística PC_PAR).

8 Investigación en curso y futura

El método de DSA que proponemos en la presente tesis supone un enfoque nuevo a la asignación de sentidos y por lo tanto plantea toda una serie de cuestiones, de orden teórico (lingüístico) como práctico (desde el proceso mismo de desambiguación). Inevitablemente, una primera investigación, como la que hemos introducido en los capítulos anteriores (6 y 7), no puede cubrir, en su totalidad, la problemática de la propuesta. A pesar del estudio más profundo de ciertos aspectos del enfoque, la investigación desarrollada corresponde ante todo a un planteamiento general. Se han operado simplificaciones, teóricas y prácticas, sin que se tomen en cuenta parámetros que permitirían una estrategia de desambiguación diferenciada, matizada. En este capítulo, señalamos algunas limitaciones y aspectos abiertos no resueltos del presente estudio y esbozamos las vías que se exploran actualmente o se proyecta investigar para superar estas dificultades.

Nuestras preocupaciones actuales se orientan según los siguientes *objetivos* interrelacionados:

- mejorar los parámetros del sistema de DSA propuesto (apartado 8.1.),
- explotar el método para analizar aspectos lingüísticos del proceso de desambiguación (apartado 8.2.);
- explotar el método en aplicaciones del PLN (apartado 8.3.);
- investigar implicaciones sobre el desarrollo de la DSA (apartado 8.4.).

Estos objetivos corresponden a *líneas de estudio abiertas* al margen de nuestra propuesta, que detallamos a continuación.

8.1 Refinamiento y ampliación del método

Desde una perspectiva estricta de la DSA, interesa la mejora del método en los principales atributos: precisión y cobertura. En la línea que hemos seguido en toda nuestra investigación, nuestra atención se centra en las posibilidades de incorporar más conocimiento lingüístico en el método. Empezaremos, por lo tanto, con unas consideraciones de orden lingüístico sobre aspectos de nuestra propuesta susceptibles de refinamiento (subapartado 8.1.1.). Presentaremos, luego, los desarrollos efectivos destinados a mejorar cada uno de los pasos del método de DSA (sección 8.1.2.).

8.1.1. En el método que hemos presentado usamos sólo parte de los resultados teóricos de la semántica léxica y composicional que hemos sintetizado en el capítulo 2. A continuación, enumeramos algunos de estos resultados y apuntamos otros tantos igualmente aprovechables para nuestra estrategia de DSA.

Así, hemos asumido que el Principio de la Composicionalidad es siempre aplicable, sin considerar sus límites. De momento no utilizamos herramientas para la identificación de locuciones (multipalabras) en el texto, previamente a la aplicación del método de desambiguación que proponemos. Esta opción no es intrínseca a nuestro sistema de DSA, sino que es una opción de circunstancia, fácilmente corregible a través de un módulo previo de identificación de multipalabras en *WordNet*.

En la forma en que hemos usado un patrón local X-R-Y de una ocurrencia ambigua X para su desambiguación, no hemos considerado las eventuales relaciones sintácticas de la segunda palabra Y, de contenido léxico, con otras palabras de la oración. Sin embargo, estas últimas relaciones se muestran altamente útiles debido a que a menudo la polisemia de Y hace que se introduzca mucho "ruido" en la extracción de información relacionada con X dentro del patrón. El estudio sobre el corpus demuestra que ampliando el patrón con el especificador o modificador de Y se supera esta limitación.

Por otra parte, nos proponemos tener en cuenta las restricciones semánticas recíprocas entre los elementos de un sintagma. Es decir, explotar las restricciones recíprocas entre las dos palabras de contenido léxico del patrón, tal como las hemos presentado en el apartado 6.7.2.

Sumando las observaciones previas, consideramos que se deben explotar de manera más intensiva los patrones léxico-sintácticos para la desambiguación: en su interior, en cuanto a las restricciones recíprocas entre las dos palabras de contenido léxico de un patrón, y fuera del patrón, en cuanto a las restricciones que reciben ambas palabras de un patrón fijado por parte de otros patrones que las contienen. Para dar cuenta de estas múltiples interdependencias (intra- e interrelaciones) que, en nuestra opinión, intervienen en la construcción del significado de una palabra y, por lo tanto, en su desambiguación, delimitamos una serie de modalidades para la explotación de los patrones léxico-sintácticos en la DSA, tal como se describen a continuación.

- *modalidad SIMPLE*, en que se opera sobre una sola posición del patrón: L1 o L2.
- *modalidad DOBLE*, en que se opera sobre ambas posiciones: L1 y L2;
- *modalidad MIXTA*, en que se consideran los varios patrones en que participa una ocurrencia ambigua;
- *modalidad CÍCLICA*, en que se consideran todas las palabras léxicas de una oración, y para cada una todos los patrones sintagmáticos en que participa;
- *modalidad INTEGRATIVA*, en que se integra el patrón sintagmático en la oración, y se usa toda la oración como “bolsa de palabras”.

Hemos implementado ya las modalidades SIMPLE, MIXTA e INTEGRATIVA y estamos desarrollando la modalidad DOBLE, mediante el conjunto PAR" (apartado 6.7.2.). Una variante simplificada de la modalidad CÍCLICA, fácil de implementar, consiste en la propuesta comentada previamente, de incorporar dentro del patrón léxico-sintáctico el modificador de la otra palabra de contenido léxico

Volviendo a la interdependencia entre las dos palabras léxicas de un patrón léxico-sintáctico, nuestro método se vería beneficiado si aprovechara los estudios teóricos sobre los modos en que se combinan los significados de las palabras de un sintagma, según factores como el tipo de relación sintáctica (coordinación *vs.* modificación) y el tipo semántico de las unidades léxicas implicadas. Consideramos también necesario distinguir entre la desambiguación de un núcleo de un sintagma y la desambiguación de un modificador.

En la identificación de patrones actualmente se opera al nivel del lema. Al considerar el lema y no la forma flexiva de las palabras, hemos operado una reducción: hemos supuesto que la forma gramatical no aporta diferencia de significado. En otras palabras, focalizamos el segundo paso del proceso de composición semántica, o sea la construcción del significado de las expresiones lingüísticas complejas a partir del significado de las expresiones simples, y hemos trascurrido la construcción del significado de la palabras en base de sus elementos componentes (cf. Löbner, 2002:11-16).

8.1.2. No hemos podido resolver de manera satisfactoria algunos pasos de nuestro método o implementar algunas variantes suyas. A continuación, detallamos las limitaciones presentes y algunas soluciones para su reducción.

En la *identificación de los patrones* queda un problema por resolver, ante todo en cuanto a la *cobertura*.

Para cubrir la casuística de los patrones léxico-sintácticos en que los nombres pueden participar, incorporaremos los patrones con verbos y los correspondientes esquemas de búsqueda:

[N V]
[V N]
[N PRON-REL V]
[V PREP N], etc.

En la misma línea de mejorar la cobertura, es necesario resolver el tratamiento de las secuencias - muy frecuentes - que corresponden al esquema:

[N ADJ/VPART PREP N],

esquema altamente problemático para la descomposición en patrones básicos (cf. apartado 6.7.2.). Actualmente las secuencias se usan sólo parcialmente, o sea se reducen al patrón:

[N ADJ/VPART].

Siempre dentro la categoría de los nombres, los *nombres propios* requieren un tratamiento particular, ya que actualmente tratamos indistintamente los nombres comunes y los propios. En concreto, proponemos el uso de un diccionario enciclopédico o de una herramienta para el reconocimiento de entidades nombradas, que permitan la individuación de la clase a la cual pertenecen las entidades denotadas por los nombres propios. Esta operación nos permitiría reducir los patrones con nombres propios a patrones generalizados que contienen el nombre común de su clase. Estamos hablando de una generalización del tipo:

[Nc1-PREP-Np2] → [Nc1-PREP-Nc2{Np2}],

como por ejemplo:

[*pasaje para París*] → [*pasaje para ciudad*{*París*}]
[*billete para Inglaterra*] → [*billete para país*{*Inglaterra*}]

De esta manera, los patrones con nombres propios se podrían reducir a patrones con nombres comunes y además se operarían generalizaciones como:

[{*pasaje, viaje, billete*} *para* {*ciudad, país*}].

La utilidad de esta generalización es obvia, permitiría cubrir una gran cantidad de casos en que están implicados los nombres propios.

Un imperativo del futuro inmediato es la implementación del método a las otras categorías gramaticales, adjetivo y verbo.

Pasando a la cuestión del *filtrado* de los patrones, hemos visto que queda una cuestión abierta, debido a que el requisito de la simple repetición de los patrones se ha mostrado insatisfactorio por ser demasiado restrictivo. Proponemos un criterio más, la ratio de asociación. Así, se guardarían los patrones cuyas palabras de contenido léxico tienen una ratio de asociación por encima de un límite inferior preestablecido[226]. El uso de este tipo de filtro encuentra respaldo en la teoría semántica, donde se considera que una relación gramática suele ser acompañada por una afinidad entre las palabras (Cruse, 2000a). Unas pruebas preliminares sobre el corpus de entrenamiento de Senseval-3 parecen indicar que este filtro es realmente útil. Además, es necesario estudiar si este filtro se debe usar como alternativa a la frecuencia de los patrones o bien en combinación con ésta.

En la *extracción de información* relacionada con la ocurrencia ambigua, una prioridad es la adquisición y el uso de los conjuntos PAT y PAR_{4k}, propuestos en el apartado 6.7.2.

Como complemento de los diferentes conjuntos de información relacionada a la ocurrencia ambigua dentro de un patrón (PAT, PAR, SINT, OR, cf. apartado 6.7.2.), proponemos la identificación de claves de desambiguación asociadas a los sentidos. En concreto, las claves para un sentido dado se obtienen a partir de la los patrones etiquetados con este sentido y corresponden a los siguientes elementos vinculados con cada uno de estos patrones: la otra palabra léxica del patrón, las palabras

[226] Hemos introducido la fórmula de la ratio de asociación en la nota 144.

filtradas coocurrentes con el patrón, los sustitutos de la palabra ambigua dentro del patrón (estos últimos de los paradigmas filtrados PAR_{3k} o PAR_{4k}). El conjunto de claves para un sentido dado se obtendrá como reunión de los conjuntos de claves obtenidas a partir de cada uno de los patrones etiquetados con este sentido.

Para el *filtrado* de la información asociada a la ocurrencia ambigua dentro de un patrón léxico-sintáctico, proponemos dos criterios más: la ratio de asociación y el dominio. La ratio de asociación se usa igual que para el filtrado de los patrones. La información de dominio a la cual nos referimos para el filtrado son las etiquetas de *WordNet Domains* (Magnini y Cavaglià, 2000, cf. apartado 3.1.1.3.). Precisamente, de un conjunto vinculado a la ocurrencia, se guardarán las palabras que tengan la misma etiqueta de dominio con la palabra por desambiguar.

Respecto a los algoritmos de DSA que usamos, proponemos implementar las variantes de la Prueba de Conmutabilidad que se describen a continuación.

Así, una primera variante es la *PC extendida*, que hemos introducido en el apartado 6.7.3. Hemos ya extraído los DS ampliados para las palabras del muestrario analizado en la presente tesis, es decir las palabras de concurso en Senseval-2. Está en curso la aplicación del algoritmo correspondiente, la Prueba de Conmutabilidad extendida, sobre los mismos conjuntos usados hasta ahora: PAR y SINT. Nuestro objetivo es la comparación de las dos variantes de la Prueba de Conmutabilidad: PC básica *vs.* PC extendida. En concreto, nos interesa estimar si la precisión disminuye al ampliar los conjuntos de discriminadores y evaluar la variante justa respecto a ambos parámetros de precisión y cobertura.

Otra extensión de la PC que nos proponemos implementar es lo que podemos llamar *Prueba de Conmutabilidad débil*. En esta variante, se considerarán también los *variants* de los *synsets* relacionados en *EuroWordNet* con varios sentidos de una misma palabra y que en la PC básica se han eliminado para obtener la disjunción entre los conjuntos de discriminadores de sentido. Específicamente, se considerará que si en los conjuntos palabras asociados a la ocurrencia ambigua hay "*variants* comunes" a dos o más sentidos, entonces la palabra ambigua puede tener todos estos sentidos en al patrón de partida. Esta extensión está destinada a mejorar la cobertura del método, debido a que para los sentidos muy próximos en la jerarquía de *EuroWordNet* se eliminan muchos *variants* de los *synsets* relacionados comunes y en consecuencia los conjuntos de discriminadores son muy reducidos. Obviamente, en este caso las heurísticas propondrán varios sentidos como respuesta. Nos interesa comparar la Prueba de Conmutabilidad débil con la Prueba de Conmutabilidad básica en cuanto a precisión y cobertura.

Con el propósito de desarrollar un análisis orientado igualmente hacia la mejora del sistema y hacia la interpretación lingüística, hemos establecido un sistema de etiquetas específicas para poner al lado de cada discriminador. Usamos tres marcas, para indicar los siguientes parámetros:

1) el tipo de relación que establece cada discriminador con el sentido de la palabra de partida (el *synset* de partida) al que corresponde;
2) la vinculación "directa" del discriminador con el sentido de partida, o sea su aparición entre los DS básicos, o bien "indirecta", como hipónimo o merónimo de un discriminador básico, entre los DS extendidos;
3) la distancia, calculada en números de arcos, con respecto al sentido (*synset*) de partida.

Llamaremos esta variante de nuestra adaptación de *WordNet Discriminadores de Sentidos con marcas*.

Prueba de Conmutabilidad con similitud. La Prueba de Conmutabilidad opera actualmente en base a la intersección entre, por una parte, los conjuntos de discriminadores de cada uno de los sentidos de la palabra por desambiguar y, por otra, los conjuntos de palabras vinculados a la ocurrencia ambigua dentro del patrón (estos últimos conjuntos siendo extraídos del corpus). Obviamente, el requisito de coincidencia perfecta entre los elementos de estos conjuntos limita la cobertura del método. Una extensión natural destinada a la mejora de la cobertura es la relajación de esta coincidencia. Más precisamente, se trata de usar una medida de similitud - todavía por precisar - para aplicar a los discriminadores de sentido y a las palabras vinculadas con la ocurrencia (estas últimas, de los conjuntos PAR, SINT, PAT). En este caso, se aceptará como respuesta afirmativa a favor de un sentido (para la palabra ambigua dentro del patrón) también la existencia de pares de palabras

formados por un discriminador de sentido (de los asociados a este sentido) y una palabra vinculada a la ocurrencia ambigua, pares cuya similitud supera un lindar preestablecido. Llamaremos esta variante del algoritmo *Prueba de Conmutabilidad con similitud*. La extensión del método que acabamos de mencionar va en la línea de trabajos similares realizados por Li *et al.* (1995) y Pedersen *et al.* (2003). Como en el caso de la Prueba de Conmutabilidad extendida, interesa aquí también confrontar las supuestas ganancias en cobertura y pérdidas en precisión.

Siguiendo una tendencia cada vez más extendida en la DSA, nos proponemos el *uso de Internet como corpus*. Como todo método que recurre a un corpus, nuestra estrategia de DSA depende altamente de la cantidad de datos que se maneja (cf. experimentos del apartado 7.2.). Nos referimos aquí al uso de Internet igualmente para el filtrado de frecuencia como para la extracción de información asociada a la ocurrencia ambigua dentro de un patrón léxico-sintáctico.

Estudio de los patrones sintagmáticos de base léxico-semántica para la DSA. En nuestra investigación, hemos delimitado desde el principio el objeto de estudio a los patrones léxico-sintácticos. Creemos, sin embargo, que es necesario explotar la combinación de los dos tipos de patrones sintagmáticos (sintácticos y léxico-semánticos) para la DSA. El uso de los patrones de base léxico-semántica es útil para cubrir parte de las relaciones de significado entre palabras situadas a distancia, que los patrones léxico-sintácticos, tal como se identifican ahora (mediante esquemas de búsqueda y no con un *parser*), no pueden cubrir.

Desde lejos, nuestra prioridad absoluta es la *aplicación a gran escala* para una *evaluación real* de la estrategia de DSA propuesta. En esta dirección, una posibilidad es aplicar una estrategia de autoevaluación del método. En concreto, se tomarán en el muestrario de evaluación ocurrencias que participan en dos o más patrones léxico-sintácticos y se confrontarán las asignaciones de sentido que se propone usando los diferentes patrones de una misma ocurrencia. Si estas propuestas son convergentes, o sea tienen sentidos en común, entonces se considerará que la ocurrencia se ha desambiguado correctamente.

8.2 Estudios de carácter teórico

Nuestra posición en la presente tesis, expresada repetidamente, es que debe haber una relación dialéctica entre la DSA y la semántica teórica. Si en la investigación realizada hasta el presente nos hemos centrado en la incorporación de conocimiento lingüístico en la tarea de asignación de sentidos, es necesario desarrollar, inversamente, estudios con relevancia teórica. Para profundizar las implicaciones en plan lingüístico de las diferentes opciones que se toman en la construcción y aplicación del sistema de DSA, tenemos proyectados experimentos orientados hacia el análisis lingüístico, tal como se describe a continuación.

Un aspecto de interés es la evaluación de la contribución que tienen los diferentes tipos de información relacional léxico-sintáctica de *EWN* en la asignación de sentido. Con este objetivo, se utilizarán los *Discriminadores de Sentidos con marcas*, en que los discriminadores aparecen junto con la relación léxico-semántica que establecen con el sentido de partida (cf. apartado 8.1).

A la vez, profundizaremos el análisis del juego complejo de la información paradigmática y sintagmática en la asignación de sentido, una cuestión que hasta ahora hemos tocado sólo tangencialmente. Interesa analizar la modalidad óptima de uso de cada uno de los dos tipos de información y comparar su utilidad para la identificación del sentido.

Nos hemos pronunciado en la presente tesis (apartado 6.4.) para una estrategia de DSA diferenciada según la categoría gramatical de la ocurrencia por desambiguar. Se impone, por lo tanto, un análisis del rendimiento del método para las diferentes categorías, con propósito de identificar aspectos idiosincráticos de cada una de ellas. El estudio, como tantos otros de los realizados o propuestos, tiene un doble carácter teórico y aplicado, ya que es una fuente para futuras mejoras del método.

Finalmente, una cuestión teórica importante sobre la cual puede repercutir nuestra investigación es la identificación de criterios objetivos para la discriminación de los sentidos. Nos situamos en la línea

de los que proponen criterios sintácticos en la delimitación de los sentidos (cf. apartado 2.2.). Así, creemos que el uso de patrones léxico-sintácticos es una modalidad fundamentada lingüísticamente para la obtención de agrupaciones de palabras (*clusters*) que contengan una palabra dada y que correspondan a sentidos de la palabra.

8.3 Aplicaciones en el marco del PLN

Nuestra propuesta permite ante todo la adquisición de conocimiento léxico, pero es igualmente útil para aplicaciones concretas del PLN, como la recuperación de información, la traducción automática, sistemas de pregunta-respuesta, etc. A continuación, nos centramos en la adquisición de conocimiento léxico, aprovechable dentro de la DSA misma (subapartado 8.3.1.) o bien para el enriquecimiento de las fuentes léxicas (subapartado 8.3.2.).

8.3.1. A través del método propuesto, se puede obtener conocimiento que, en un proceso de *feed-back*, nutre la misma tarea de DSA. Nos referimos aquí a patrones etiquetados a nivel de sentido, claves de sentido y ejemplos etiquetados con sentidos.

Así, como resultado de la aplicación de nuestro sistema de DSA, se obtienen patrones etiquetados a nivel de sentido, en términos de sentidos y de *synsets* de *EuroWordNet*. Por lo tanto, proyectamos la construcción de una base de datos de patrones anotados con sentidos, con el formato siguiente:

Palabra	Tipo de patrón (categorías sobre posiciones)					Ejemplo (nivel de lema)	Votos por sentido					Sentido más votado
	-2	-1	0	+1	+2		s_1 (02831270n)	s_2 (03650737n)	s_3 (05302115n)	s_4 (07977350n)	s_5 (02604665n)	
órgano	N	S	N			*informe de órgano*		100%				s_2
						…						
						…						
			N	A		*Órgano afectado*		12,5%	62,5%	25%		s_3
						…						
						…						

Figura 8.1. Formato de la base de patrones anotados con sentidos

Vemos útil esta base de patrones para la reutilización de las asignaciones de sentido obtenidas para llevar a cabo futuras desambiguaciones. Obviamente, esta reutilización se hace en base de la hipótesis asumida que hay una tendencia hacia un único sentido por patrón léxico-sintáctico. Como hemos mencionado a lo largo de la presente tesis, en algunos casos la desambiguación mediante el uso de patrones léxico-sintácticos es parcial, o sea se reduce pero no se resuelve por completo la ambigüedad, por lo tanto la explotación de los patrones etiquetados es más bien un pre-etiquetado a nivel de sentido.

Como consecuencia de la anotación de los patrones, hemos proyectado un proceso de adquisición de claves de desambiguación, es decir palabras o expresiones complejas asociadas a los sentidos. Hemos presentado la tipología de estas claves en el apartado 8.1.

Otra consecuencia de la aplicación del método es la generalización de los patrones léxico-sintácticos a clases de palabras. Los patrones generalizados contendrán los hiperónimos de palabras que ocupan cada una de las dos posiciones de contenido léxico dentro de un tipo de patrón fijado. Creemos que de esta manera se reduce el problema de la escasez de datos, en el sentido que se cubren casos no registrados en el corpus. Por otra parte, esta generalización representa una posibilidad de delimitar tipos de patrones léxico-sintácticos en que participan las clases de palabras y de aquí identificar casos de polisemia regular.

Además, debido a que no necesita ningún tipo de intervención humana, nuestro sistema de DSA se puede usar como método para el etiquetado semántico automático de ejemplos para palabras dadas. Estos ejemplos pueden constituir corpus de entrenamiento para los sistemas supervisados (Nica *et al*., 2003a).

Saliendo de la esfera de la DSA, nos proponemos comprobar la utilidad de la Prueba de Conmutabilidad (en forma *débil*, cf. apartado 8.1.) para la identificación de colocaciones respecto a las locuciones. En concreto, consideraremos que la existencia de sustitutos (próximos en *EWN*) de una palabra dentro del patrón indica que se trata de una colocación y no de una estructura fosilizada. La falta de sustitutos nos sugiere que se trata más probablemente de una locución, pero no nos podemos pronunciar de manera segura, sino con reserva debido a la escasez de datos. O sea, es posible que estos sustitutos no aparezcan en el corpus que estamos usando.

8.3.2. La adquisición de conocimiento que supone la aplicación de nuestro método encuentra una aplicación en el *enriquecimiento de las fuentes léxicas*, mediante las siguientes operaciones: incorporación o ampliación de información sintagmática, desambiguación de unidades léxicas (en las definiciones o en los ejemplos asociados a los sentidos), agrupación de los sentidos, establecimiento de correspondencia entre las fuentes léxicas.

Desde la perspectiva de la DSA, *EuroWordNet* tiene dos grandes limitaciones: la falta de información sintagmática asociada a los sentidos y la excesiva granularidad de los sentidos (cf. apartado 3.1.1.3.). Una aplicación inmediata de nuestra propuesta es el enriquecimiento de *EuroWordNet* con información sintagmática bajo la forma de patrones léxico-sintácticos etiquetados con sentidos (Nica *et al*., 2004c). Para la agrupación de los sentidos de *EuroWordNet*, proponemos la evidencia empírica que resulta de la desambiguación de las palabras dentro de sus diferentes patrones en un corpus amplio. Si a los patrones de una palabra se les asignan de manera sistemática los mismos sentidos diferentes, consideramos que los sentidos se deben unificar en uno solo. En otras palabras, los sentidos que aparecen sistemáticamente en los mismos contextos sintácticos no justifican su separación.

En nuestro trabajo (Nica, 2002b), hemos usado los patrones para la desambiguación de unidades léxicas de un diccionario en formato electrónico. Aunque se trataba de patrones léxico-semánticos de meronimia, el hecho de que la meronimia se exprese predilectamente mediante el patrón [N *de*-PREP N] hace que estos patrones sean a la vez de tipo sintáctico.

Además, probaremos nuestro método en la tarea de desambiguación de las glosas de *WordNet* (variante española) inaugurada en Senseval-3. Creemos que el método es idóneo para esta tarea debido a que trabaja sobre el contexto local reducido a patrones léxico-sintácticos.

Como consecuencia directa de la desambiguación de elementos de las fuentes léxicas proponemos el establecimiento de correspondencia entre estas fuentes. Precisamente, mediante la desambiguación de la palabra de entrada en los patrones que aparecen en los ejemplos asociados a los sentidos de la entrada. Debido a que la desambiguación se realiza en términos de *synsets* de *EuroWordNet*, el sentido correspondiente estará vinculado, a través de la desambiguación, a un *synset* de *EWN*. Siguiendo este procedimiento, se pueden poner en correspondencia cualquiera dos fuentes léxicas que tengan ejemplos asociados a los sentidos.

En la actividad de preparación de Senseval-3, hemos investigado la posibilidad de enriquecer el diccionario allí usado, MiniDir 2.1. En concreto, nos hemos centrado en la ampliación de la lista de colocaciones que el diccionario ofrecía, mediante el uso de los patrones y de la Prueba de Conmutabilidad. Los primeros estudios nos han traído resultados esperanzadores, aunque falta profundizar el procedimiento.

8.4 Nuevo desarrollo del proceso de DSA

Una de las líneas prioritarias para la investigación futura son las implicaciones de nuestra propuesta sobre el desarrollo de la DSA.

Primero, nos referimos aquí a la desambiguación "colateral" que obtenemos en la implementación de la estrategia de DSA, o sea la desambiguación de las palabras que aparecen en los conjuntos de información vinculada a la ocurrencia ambigua cuando se considera dentro del patrón: la otra palabra léxica de patrón, los sustitutos, las palabras frecuentemente coocurrentes con el patrón y los otros nombres de la oración. Además, recordamos la propuesta de generalización de los patrones individuales a clases de patrones (cf. apartado 8.1.). En otras palabras, la desambiguación adquiere un carácter *extensivo*, en el sentido que se parte de un patrón y se desambigua todo un conjunto de patrones. En las pruebas realizadas hasta e momento, no hemos podido estimar la calidad de esta desambiguación "colateral". Es necesario desarrollar una metodología para la evaluación "multidimensional" de la desambiguación obtenida. Creemos que la evaluación actual, ocurrencia por ocurrencia, desfavorece nuestro sistema.

Por otra parte, la creación de una base (o biblioteca) de datos de patrones etiquetados y de claves de sentido (cf. apartado 8.3.) entrena un cambio sustancial en el proceso de DSA. Así, nuestra visión futura para el desarrollo de la tarea de DSA es que ésta se debe hacer explotando la información almacenada en una base de datos de patrones etiquetados con sentidos y de claves asociadas a los sentidos antes que proceder a una nueva desambiguación.

En el presente capítulo, hemos querido resaltar el carácter abierto de la investigación alrededor de nuestra propuesta para la DSA. Como hemos descrito en detalle aquí, la estrategia es mejorable bajo múltiples aspectos. Esto indica la probabilidad elevada de que se superen las limitaciones de las implementaciones que hemos desarrollado hasta ahora. Creemos haber aportado argumentos a favor de la viabilidad de nuestra propuesta y de su impacto positivo sobre el proceso de DSA.

Conclusiones

1 Problemas abiertos en DSA

A lo largo del presente trabajo se ha presentado una visión de conjunto del campo de investigación de la desambiguación semántica automática. La DSA es un campo muy activo de la lingüística computacional, con un desarrollo destacable en las últimas dos décadas. Sin embargo, sigue siendo una de las mayores cuestiones no resueltas del PLN. A pesar de la gran variedad de enfoques y propuestas para solucionar la ambigüedad de las palabras a nivel de sentidos, los resultados están muy por debajo de los obtenidos para los demás niveles de análisis del lenguaje, lo que hace que la DSA se halle en un sensible desfase respecto de éstos. Concluimos la presentación con un resumen de los aspectos más relevantes de la problemática relacionada con la DSA y de las principales líneas de investigación actuales.

La dificultad de la tarea de DSA, un tema recurrente en toda la bibliografía, radica en la necesidad de disponer del conocimiento del mundo, del sentido común, etc., para una buena solución. La muy difícil obtención y representación de estos conocimientos explica la diversidad de enfoques y de información usada. Por otra parte, las limitaciones y los pobres resultados de cada una de las aproximaciones han llevado al uso de sistemas híbridos, que tienden a ser la nota dominante en la labor actual en la DSA.

El progreso realizado recientemente en la investigación de DSA y el rápido desarrollo de los métodos para solucionar el problema ha requerido centrar los esfuerzos en la síntesis, evaluación y comparación de los trabajos realizados para establecer el estado de la cuestión, y considerar los pasos futuros necesarios[227].

Como esperamos haber puesto de manifiesto, los problemas que todavía quedan por superar son tanto de orden teórico como computacional. Ide y Véronis (1998) identifican en la desambiguación semántica automática, por encima de los problemas metodológicos específicos de los distintos sistemas, tres grandes problemas abiertos: a) el papel del contexto, b) la división de sentidos, y c) la evaluación de la tarea de DSA. Hemos tratado estas cuestiones en apartados o capítulos específicos (3.3., 2.2. y 5, respectivamente).

En relación con la mencionada cuestión de la división de los sentidos, el análisis de los resultados últimamente obtenidos (véase, por ejemplo, Senseval-2) ha reafirmado la necesidad de distinciones de sentido claras y bien motivadas, útiles para la DSA. Senseval-3 lo ha demostrado para el español y el catalán en la tarea "*lexical sample*". Así, se da más relieve a la información sintagmática en la delimitación de los sentidos de una palabra (v. 2.2.). El hecho de que el conocimiento útil para la DSA se extraiga predominantemente de las fuentes léxicas existentes ha motivado la tendencia a construir lexicones computacionales o bases de conocimiento léxico adecuadas a las necesidades de la DSA. Una alternativa para solucionar la cuestión del inventario de sentidos usados en la DSA está representada, por ejemplo, en la construcción de redes semánticas dinámicas, a base del análisis de ejemplos, como es el caso de *MindNet* (v. 3.1.1.3.). Se intenta así evitar establecer divisiones arbitrarias entre los sentidos y tratar casos de sentidos nuevos o palabras desconocidas, para alcanzar una amplia cobertura.

[227] Recordamos aquí los números monográficos de *Computational Linguistics. Special Issue on Word Sense Disambiguation*, 24 (1), 1998[227], y de *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs***,** 34 (1-2), 2000 y de *Natural Language Engineering*, 8(4), 2002, o los *Proceedings of SENSEVAL-2. Second International Workshop on Evaluating Word Sense Disambiguation Systems*, ACL, 2001.

Según Màrquez (2002), actualmente se pueden ver como dominantes en la labor de DSA dos líneas de investigación:

> 1) las aproximaciones basadas en técnicas supervisadas de aprendizaje automático a partir de corpus etiquetados semánticamente;
> 2) las aproximaciones basadas en el uso de fuentes de conocimiento léxico-semántico preexistentes (DAM, corpus bilingües alineados, ontologías y taxonomías semánticas de tipo *WordNet*, etc.), en las cuales no se realiza ningún aprendizaje a partir de los ejemplos.

Muchos estudios y experimentos recientes se dirigen hacia las necesidades específicas de cada una de estas dos líneas de investigación.

Las tres ediciones de la competición Senseval han puesto de manifiesto la superioridad de los sistemas de DSA supervisados. Sus resultados sensiblemente mejores hacen que dichos sistemas se vean actualmente como la línea de investigación más fecunda en DSA. De manera implícita, su gran limitación, la dependencia de corpus etiquetados, ha estimulado la investigación para compensar esta carencia (v. 3.1.2.). Tal como ha demostrado la tarea española de Senseval-3, la inversión en la calidad de las fuentes léxicas estructuradas y de aquí en los corpus etiquetados en base a estas fuentes tiene un impacto positivo evidente sobre el nivel de la DSA supervisada. Vemos en esta mejora cualitativa de las fuentes léxicas que se usan para la preparación de la anotación de los corpus de entrenamiento un potencial motor de progreso en la DSA supervisada.

Por otra parte, la edición de Senseval-3 aporta una nueva luz sobre el "equilibrio de fuerzas" entre los métodos supervisados y no supervisados. En la tarea del inglés de inventario limitado, la DSA no supervisada ha registrado un salto cualitativo notable, de casi 25% más que en Senseval-2, reduciendo así la distancia respecto a los sistemas supervisados en un sólo 7%. Con lo cual, consideramos que la DSA no supervisada no ha agotado su potencial y es un terreno fértil, insuficientemente explorado, con la gran ventaja de la independencia de la intervención humana.

La investigación actual está focalizada (cf. Yarowsky, 2000b; Màrquez, 2002) en:

a) explotar otros recursos potenciales para la obtención automática de datos de entrenamiento etiquetados a nivel de sentidos, como corpus bilingües alineados;
b) diseñar métodos para la construcción de muestras de aprendizaje representativas (*sampling*);
c) mejorar la velocidad y la eficiencia de la anotación humana mediante algoritmos de aprendizaje que guíen de manera dinámica sesiones de etiquetado interactivo, en base a información necesaria y no provista por las fuentes léxicas actuales;
d) desarrollar algoritmos que puedan alimentarse mejor a partir del conocimiento léxico y ontológico presente en fuentes existentes, como diccionarios electrónicos en línea, *WordNet*, tesauros, u otros algoritmos mínimamente supervisados de conducción sobre corpus no anotados; el uso de técnicas semi-supervisadas permite reestimar iterativamente los parámetros estadísticos de un modelo sin necesidad de disponer de grandes cantidades de datos etiquetados (*bootstrapping*);
e) usar agrupaciones (*clustering*) y la inducción de sentidos para aplicaciones (p.ej., RI) que no requieran alineamiento de las particiones de sentidos obtenidas a un inventario de sentidos existente (métodos no supervisados).

Últimamente, la preocupación de construir un marco de referencia para la labor en el área hace que haya una intensa actividad de análisis, con el propósito de obtener generalizaciones, de los varios factores implicados en el proceso de DSA. Así, se han desarrollado diversos experimentos con carácter comparativo que han permitido contrastar técnicas de tipo distinto o de la misma clase, el potencial de las diferentes fuentes y de los diferentes atributos para la desambiguación, categorías de corpus, o han estudiado la colaboración de estos elementos en el desarrollo del proceso.

Respecto de la portabilidad y adaptación (*tuning*) de los métodos de DSA, dichos experimentos han comprobado que el género y el área temática de los corpus serían dos parámetros a tener en cuenta en los modelos de la DSA**.** La DSA es muy dependiente del dominio de aplicación, así para asegurar la transferencia de los sistemas es imprescindible la aplicación de algún procedimiento de adaptación a nuevos dominios.

El análisis de la contribución de los factores implicados en el proceso de DSA parece conceder un papel más importante a los atributos frente a los algoritmos, con lo cual los atributos deben ser lo más informativos posible. De hecho, es una opinión cada vez más aceptada en la comunidad de la LC la necesidad de integrar más conocimiento lingüístico en los sistemas de DSA (Manning y Schütze, 1999; Corazzari *et al.*, 2000). La DSA podría beneficiarse de representaciones más ricas de atributos que representen dependencias léxicas, sintácticas, pragmáticas y discursivas más finas. Para ello se requiere mejorar los algoritmos para extraer dicha información de corpus y otras fuentes de conocimiento disponibles. En consecuencia, el progreso futuro en DSA depende ampliamente del progreso paralelo en las demás tareas de análisis textual (Yarowsky, 2000b). Destacamos, a la vez, la preocupación para mejorar y desarrollar metodologías para evaluar la calidad de las fuentes léxicas.

A modo de conclusión, consideramos que se puede hablar hoy de una nueva fase, de madurez, en el área de la DSA, con una natural búsqueda de sistematización. Vemos en esto la nota dominante de los esfuerzos de evaluación, comparación e intentos de llegar a respuestas que nos orienten en la labor futura de DSA. En el polo opuesto, Kilgarriff (1997) lanzaba una mirada indiscreta e irónica hacia el futuro: probablemente los futuros lexicones (*application lexicons*) estarán más bien orientados hacia una aplicación específica (*application-driven*) que guiados y condicionados por los recursos disponibles (*resource-driven*), y por lo tanto la dificultad de la DSA disminuirá.

2 Contribución de la presente tesis

Estado de la cuestión en DSA. Hemos ofrecido una síntesis al día del estado de la cuestión y un análisis crítico del trabajo realizado hasta el momento: recopilación de las heurísticas utilizadas para DSA; análisis crítico de las fuentes de conocimiento utilizadas, tanto léxicas como textuales, y su adecuación para tareas de DSA; análisis crítico del conocimiento lingüístico utilizado por los sistemas de DSA. Las conclusiones se han presentado en el apartado anterior.

Nuestra aportación. El punto de partida de nuestra propuesta para la DSA es un análisis de los puntos críticos de la misma: la distancia entre la información provista por las fuentes estructuradas de conocimiento léxico y la información disponible en el contexto de una ocurrencia ambigua; la delimitación y el tratamiento insuficientes del contexto local; la dificultad de la adquisición de conocimiento útil para la DSA. En base de este análisis, proponemos soluciones para las limitaciones señaladas desde una perspectiva lingüística.

a) Esto nos lleva a la delineación de una estrategia diferente para la DSA. Partimos de la hipótesis de que el significado de una palabra en un contexto está determinado principalmente por las relaciones sintácticas que ella establece con las demás palabras de la oración. Para representar, identificar y tratar las relaciones sintácticas, introducimos el término de *patrón léxico-sintáctico*. Esto nos permite proceder a la desambiguación de una palabra ambigua integrándola en sus patrones léxico-sintácticos. Asumimos que la integración de una palabra en un patrón léxico-sintáctico reduce drásticamente su polisemia y que hay una "tendencia hacia un único sentido por patrón léxico-sintáctico". La integración en los patrones léxico-sintácticos no es de por sí siempre suficiente para la desambiguación de una palabra. Vemos, por lo tanto, el proceso de desambiguación de una palabra ambigua a través de una integración progresiva en el contexto: primero en sus patrones léxico-sintácticos y luego en toda la oración. En el proceso de desambiguación consideramos tanto la información obtenida a través de la integración de la palabra en sus patrones léxico-sintácticos como la información provista por la oración en su totalidad.

b) La integración de una palabra ambigua en un patrón léxico-sintáctico nos permite aprovechar la interacción que existe en el lenguaje entre los ejes sintagmático y paradigmático para la obtención de información relacionada con la ocurrencia ambigua, igualmente paradigmática y sintagmática. Proponemos, por lo tanto, una estrategia para la adquisición completamente automática de información relacionada con una ocurrencia ambigua, con el uso mínimo de unos esquemas sintácticos. Además, el hecho de que una ocurrencia ambigua puede participar en varios patrones léxico-sintácticos permite adquirir más información relacionada con ella. Desde esta perspectiva, nuestra propuesta es una solución parcial a los problemas de escasez de datos y de adquisición de conocimiento léxico.

A la vez, proponemos modalidades para el filtro de esta información, según criterios distintos.

c) El mismo análisis de partida nos sugiere que el uso variado de la información disponible en las fuentes léxicas para los sentidos es beneficiosa para el proceso de DSA. A partir de *EuroWordNet*, hemos derivado un nuevo recurso léxico, denominado *Discriminadores de Sentido*, en que los sentidos de una palabra se caracterizan a través de conjuntos disjuntos de palabras. Este recurso nos permite desarrollar el proceso de DSA a través de la explotación de información adquirida de corpus no etiquetados con sentidos, mediante nuestro algoritmo de la Prueba de Conmutabilidad: la sustitución de una palabra ambigua con un discriminador asociado a un sentido significa que la palabra puede tener el respectivo sentido en el marco del patrón léxico-sintáctico.

d) Partiendo de la observación que el conocimiento lingüístico puede intervenir en la aplicación de los algoritmos de DSA, hemos procedido en este sentido para utilizar el algoritmo de la Marca de Especificidad de Montoyo y Palomar (2000). El algoritmo fue diseñado inicialmente para que opere sobre las nombres de la oración en que aparece la palabra por desambiguar, sin tener en cuenta la estructura sintáctica de la oración. Nosotros hemos implementado el algoritmo sobre los conjuntos de tipo paradigmático y sintagmático que se obtienen del corpus a partir de los patrones léxico-sintácticos de la palabra ambigua.

e) Las consideraciones previas nos llevan a diseñar un sistema de DSA en el que se juntan varias heurísticas. Las heurísticas se agrupan en dos clases: heurísticas relacionadas con los patrones léxico-sintácticos y heurísticas relacionadas con la oración. Su combinación se hace respectando esta agrupación: dentro de cada clase y finalmente entre las dos clases. La modalidad efectiva para desarrollar las diferentes fases de la combinación se ha establecido por vía empírica, a través de la evaluación de las variantes. En la evaluación, hemos obtenido una precisión muy buena, hasta el 96%. En cambio, la cobertura está limitada por la casuística parcial tratada actualmente: se tratan exclusivamente los nombres y, de éstos, sólo las ocurrencias implicadas en una relación sintáctica con otro nombre, con un adjetivo o con un verbo en participio; no se han tratado las relaciones con verbos en modos predicativos.

Nuestra propuesta es independiente de un corpus etiquetado a nivel de sentidos, sólo necesita un preprocesamiento previo del texto por desambiguar y del corpus de búsqueda a nivel de categorías morfosintácticas. Por estas razones, el método es fácilmente implementable y transferible a cualquier lengua que dispone de un etiquetador a nivel de categorías morfosintácticas y de una taxonomía como *WordNet*. Además, siempre trabajamos con textos reales, nunca creamos ejemplos a través de la sustitución y por lo tanto no corremos el riesgo de introducir “ruido” en los datos que estamos usando.

f) Otra ventaja de nuestra propuesta es que permite la reutilización de la desambiguación (total o parcial) adquirida para una palabra integrada dentro de un patrón léxico-sintáctico: cada vez que la palabra aparecerá en este patrón léxico-sintáctico, tendrá el mismo sentido o los mismos sentidos. Proponemos la creación de una base de datos de patrones léxico-sintácticos etiquetados con sentidos como nueva fuente léxica aprovechable en la DSA y en otras tareas del PLN.

g) Hemos demostrado que la incorporación de conocimiento lingüístico en la tarea de DSA es útil. El método tiene implicaciones teóricas, ya que confirma la incidencia de la sintaxis en el significado de las palabras. Se confirman las suposiciones sobre la interrelación entre el significado de las palabras y el de las unidades sintácticas superiores (sintagma y oración) y se confirman resultados obtenidos previamente en la DSA explotando la información sintáctica. A la vez, el método permite el estudio de criterios objetivos para la identificación de los sentidos.

Creemos que la investigación desarrollada aporta cierta evidencia sobre algunos factores más que intervienen en la asignación del sentido a una palabra en el texto. Por una parte, hemos revelado la interacción compleja entre la información paradigmática y sintagmática asociada a la ocurrencia ambigua, si bien esta interacción necesita un estudio mucho más profundo. Hemos esbozado posibles métodos de estudio. Por otra parte, parece que algún cálculo de frecuencia interviene en la asignación de sentido: antes de la aplicación del algoritmo, como filtro sobre la información vinculada a la ocurrencia ambigua, o en la aplicación misma del algoritmo Prueba de Conmutabilidad, considerando

la frecuencia de los Discriminadores de Sentido en estos conjuntos de información. En consecuencia, los resultados parecen indicar que es necesario combinar el conocimiento lingüístico con la información estadística para obtener asignación de sentido con un alto grado de fiabilidad.

h) Como hemos mostrado, nuestra propuesta tiene varias posibles aplicaciones en el campo de la DSA y más general, del PLN. Una aplicación inmediata es la ampliación de la información disponible en las fuentes léxicas estructuradas en relación con los sentidos, a través de incorporación de ejemplos etiquetados con sentidos o de información sintagmática asociada a los sentidos, ampliación de colocaciones, etc. o bien a través del establecimiento de correspondencias, a nivel de sentido, entre diferentes fuentes léxicas. Así, la estrategia de DSA que proponemos determina una relación dialéctica con las fuentes léxicas estructuradas: la desambiguación de palabras en el interior de patrones léxico-sintácticos lleva al enriquecimiento de las fuentes léxicas con información sumamente útil para el proceso de desambiguación, lo que implica la mejora cualitativa de la tarea de DSA. El proceso es cíclico, de modo que en este enfoque la desambiguación se autoalimenta y mejora de manera reflexiva. Dentro de la misma área de la DSA, nuestra propuesta puede contribuir a mejoras en el enfoque basado en corpus, que necesita corpus anotados a nivel de sentido, debido a que se puede ver igualmente como un método de etiquetado semántico completamente automático. Otras consecuencias de nuestra propuesta se dan en las diferentes tareas de PLN cuando la DSA es un módulo incorporado: Traducción Automática, Preguntas-Respuestas (*Question–Answering*), Recuperación y Extracción de Información, Resumen Automático, etc.

* * *

Nuestra posición en la presente tesis es que entre la semántica léxica y la lingüística del corpus, por una parte, y el área de la Desambiguación Semántica Automática, por otra, hay una relación dialéctica. El sistema de DSA que proponemos se nutre de esta convención y a su vez la fortalece. Concluimos nuestro estudio en la línea de Edmonds y Kilgarriff (2002): apenas empezamos a entender la complejidad de la ambigüedad léxica y de la semántica léxica; la misión básica de la DSA es desarrollar nuestro entendimiento del léxico y en general del lenguaje.

Bibliografía

Agirre, E. (1998), *Formalization of Concept-Relatedness Using Ontologies: Applications in the construction of lexical knowledge bases, word sense disambiguation and automatic spelling correction*, tesis doctoral, Universidad del País Basco, Donostia

Agirre, E., O. Ansa, D. Martínez, E. Hovy (2001), "Enriching *WordNet* concepts with topic signatures", en *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh

Agirre, E. y O. López de Lacalle (2004a), "Clustering *WordNet* Word Senses", en *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP'03)*, Bulgaria

Agirre, E. y O. López de Lacalle (2004b), "Publicily Available Topic Signatures for all *WordNet* Nominal Senses" en *Proceedings of LREC 2004*, ELRA, Lisabona

Agirre, E. y D. Martínez (2000), "Exploring automatic word sense disambiguation with decision lists and the web", *Proceedings of the COLING Workshop on Semantic Annotation and Intelligent Content*, Saarbrüken

Agirre, E. y D. Martínez (2001a), "Knowledge Sources for Word Sense Disambiguation", *Proceedings of the Fourth International Conference TSD 2001*, Plzen (Pilsen), Czech Republic

Agirre, E. y D. Martínez (2001b), "Learning class-to-class selectional preferences", en *Proceedings of the ACL CONLL Workshop*, Tolouse

Agirre E. y D. Martínez (2002), "Integrating Selectional Preferences in *WordNet*", en *Proceedings of First International WordNet Conference*, Mysore (India)

Agirre E. y D. Martínez (2004), "The effect of bias on an automatically-built word sense corpus", en *Proceedings of LREC 2004*, ELRA, Lisabona

Agirre, E. y G. Rigau (1995), "A proposal for Word Sense Disambiguation using Conceptual Distance", en *Proceedings of Recent Advances in NLP (RANLP95)*, 258-264, Tzigov Chark (Bulgaria)

Agirre, E. y G. Rigau (1996), "Word Sense Disambiguation using Conceptual Density", en P*roceedings of COLING'96*, Copenhagen, 16-22

Agirre, E., G. Rigau, L. Padró, J. Atserias (2000), "Combining Supervised and Unsupervised Lexical Knowledge Methods for WSD", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs*, **34 (1-2)**, 103-108

Ahlswede, T.E. y M. Evans (1988), "Generating a Relational Lexicon from a Machine-Readable Dictionary", en *International Journal of Lexicography*, 1 (3), 214-237

Allen, J. (1995), *Natural Language Understanding*, 2nd edition, The Benjamin/Cummings Publishing Company, Redwood City

Alonge, A. , N. Calzolari, P. Vossen, L. Bloksma, I. Castellón, M.A. Martí y W. Peters (1998), "The Linguistic Design of the *EuroWordNet* Database", en P. Vossen (ed. invitado), *Computers and the Humanities, Double Special Issue on EuroWordNet*, **32 (2-3)**, 91-115

Anderson, J.R. (1976), *Language, Memory, and Thought*, Lawrence Erlbaum and Associates, Hillsdale, NJ

Anderson, J.R. (1983), "A spreading activation theory of memory", en *Journal of Verbal Learning and Verbal Behavior*, **22(3)**:261-95

Apresjian, J.D. (1974), "Regular Polysemy", *Linguistics*, **142**, 5-32

Asher, R.E. (ed.) (1994), *Encyclopedia of Language and Linguistics*, 10 vols., Oxford, Pergamon Press

Atkins, B. (1987), "Semantic ID tags: Corpus evidence for dictionary senses", en *Proceedings of the Third Annual Conference of the UW Center for the New OED*, 17--36, Waterloo, Canada.

Atkins, B. y B. Levin (1988), "Admitting Impediments", en *Proceedings of the 4th Annual Conference of the UW Center for the New OED*, Oxford, 1988

Basili, R., R. Catizone, M.T. Pazienza, M. Stevensom, P. Velardi, M. Vindigni, Y. Wilks (1998), "An Empirical Approach to Lexical Tuning", en *LREC'1998 (Proceedings of First International Conference on Language Resources and Evaluation)*, Granada

Basili, R., A. Cucchiarelli, C. Consoli, M.T. Pazienza y P. Velardi (1998), "Automatic adaptation of *WordNet* to sublanguages and computational tasks", en *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal

Bertagna, F., C. Soria y N. Calzolari (2001), "The Italian Lexical Sample Task", en *Proceedings of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, Toulouse, 29-32

Black, E. (1988), "An experiment in computational discrimination of English word senses", en *IBM Journal of Research and Development*, **32 (2)**, 185-194

Boguraev, B., y T. Briscoe (1989) (eds.), *Computational Lexicography for Natural Language processing*, London, Harlow

Boguraev, B. y J. Pustejovsky (1996), "Issues in Text-based Lexicon Acquisition", en *idem* (eds.), *Corpus Processing for Lexical Acquisition*, The MIT Press, Cambridge-London

Brill, E. y J. Wu (1998), "Classifier Combination for Improved Lexical Disambiguation", en Proceedings of the 17th COLING, 191-195

Brown, P., S. Della Pietra, V.J. Della Pietra, y R. Mercer (1991), "Word sense disambiguation using statistical methods", en *Proceedings of the 12th International Computational Linguistics*, Berkeley

Brown, P., V.J. Della Pietra, P. DeSouza, J.C. Lai, y R. Mercer (1992), "Class-based *n*-grams models of natural language", en Ide y Véronis (eds.), *Computational Lingustics*, **18(4)**, 467-479

Bruce, R., Y. Wilks, L. Guthrie, B. Slator, T. Dunning, (1992), *NounsSense - A Disambiguated Noun Taxonomy with a Sense of Humour. Research Report MCCS-92-246*, Computing Research Laboratory, New Mexico State University

Bruce, R. y J. Wiebe (1994), "Word Sense Disambiguation using decomposable models", en *Proceedings of the 32th Annual Meeting*, Association for Computational Linguistics, Las Cruces, NM, 139-145

Budanitsky, A. y G. Hirst (2001), "Semantic distance in *WordNet*: an experimental, application-oriented evaluation of five measures", *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh

Buitelaar, P. (1997), "A lexicon for underspecified semantic tagging", en *Proceedings of the ACL-SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, DC

Buitelaar, P. (1998), *CoreLex: Systematic Polysemy and Underspecification*, PhD Thesis, Computer Science, Brandeis University,
http://www.cs.brandeis.edu/~paulb/CoreLex/corelex.html

Buitelaar, P. y B. Sacaleanu, "Ranking and selecting synsets by domain relevance", en *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh

Calzolari, N., C. Soria, F. Bertagna, F. Barsotti (2002), "Evaluating lexical resources using SENSEVAL", en *Natural Language Engineering*, 8 (4), 279-291

Charniak, E. (1993), *Statistical language Learning*, The MIT Press, Cambridge - London

Chen, J.N. y J.S. Chang (1998), "Topical Clustering of MRD Senses Based on Information Retrieval Techniques", en Ide y Véronis (eds.) *Computational Lingustics. Special Issue on Word Sense Disambiguation*, **24(1)**, 63-95

Chodorow, M., C. Leacock, G.A. Miller (2000), "A Topical/Local Classifier for Word Sense Identification", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs*, **34 (1-2)**

Chomsky, N. (1965), *Aspects of the Theory of Syntax***,** Cambridge, M.I.T. Press

Chugur, I., J. Gonzalo, F. Verdejo (2002), "Polysemy and Sense Proximity in the Senseval-2 Test Suite", en *Proceedings of the ACL-2002 Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions"*, Pennsylvania

Collins, A. y E. Loftus (1975), "A spreading activation theory of semantic processing", en *Psychological Review*, **82(6)**, 407-428

Copestake y T. Briscoe (1995), "Semi-productive Polysemy and Sense Extension", *Journal of Semantics*, 12, 15-67

Corazzari, O., N.Calzolari, A.Zampolli (2000), "An Experiment of Lexical-Semantic Tagging of an Italian Corpus", en *LREC'2000 (Proceedings of Second International Conference on Language Resources and Evaluation)*, Athens

Cowie, J., J. Guthrie y L. Guthrie (1992), "Lexical disambiguation using simulated annealing", en *Proceedings of the 14th International Conference on Computational Linguistics, COLING'92*, volume 1, 359-365, Nantes, France

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

Cruse, A. (1995), "Polysemy and related phenomena from a cognitive linguistic viewpoint", en P. Saint-Dizier y E. Viegas (eds.), *Computational Lexical Semantics*, Cambridge University Press

Cruse, A. (2000a), *Meaning in language: An introduction to semantics and pragmatics*, New York: Oxford University Press

Cruse, A. (2000a), "Aspects of the Micro-Structure of Word Meanings", en Y. Ravin y C. Leacock (2000) (eds.), *Polysemy: Theoretical and Computational Approaches*, Oxford University Press

Dahlgren, K. (1988), *Naive Semantics for Natural Language Understanding*, Kluwer Academic Publishers, Norwell, MA

Dagan, I (2000), "Contextual Word Similarity", en R. Dale, H. Moisl y H. Somers (eds.), *Handbook of Natural Language Processing*, Marcel Dekker Inc., cap. 19, 459-476

Dagan, I. y A. Itai (1994), "Word sense disambiguation using a second language monolingual corpus", en *Computational Linguistics*, **20 (4)**, 563-596

Dagan, I., A. Itai y U. Schwall (1991), "Two languages are more informative than one", en *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 130-137

Dagan, I., F. Pereira y L. Lee (1994), "Similarity-based estimation of word cooccurrence probabilities", en *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 272-278

Dale, R., H. Moisl, y H. Somers (eds.) (2000), *Handbook of Natural Language Processing*, Marcel Dekker, New York-Basel

Deane, P. (1988), "Which NPs are there unusual possibilities for extraction from?", en MacLeod, L., G. Larson, y D. Brentari (eds.), *CLS 24: Papers from the 24th Annual Regional Meeting of the Chicago Linguistics Society-Part One: The General Session*, Chicago, Chicago Linguistics Society

Diab, M. y P. Resnik (2002), "An Unsupervised Method for Word Sense Tagging using Parallel Corpora", en *Proceedings of ACL*, 2002

Dolan, W. (1994), "Word Sense Ambiguation: clustering related senses", en *Proceedings of COLING94*, 712-716

Dolan, W., L. Vanderwende, y S. Richardson (2000), "Polysemy in a Broad-Coverage Natural Language", en Ravin, Y. y C. Leacock, C. (eds.), *Polysemy: Theoretical and Computational Approaches*, Oxford University Press

EAGLES (1998) (The EAGLES Lexicon Interest Group), *Preliminary Recommendations on Semantic Encoding*, 1998, versión en línea

Edmonds, P. y S. Cotton (2001), "SENSEVAL-2: Overview", en *Proceedings of 2nd International Workshop "Evaluating Word Sense Disambiguation Systems"*, Toulouse

Edmonds, P. y A. Kilgarriff (2002), "Introduction to the special issue on evaluating word sense disambiguation systems", en *Natural Language Engineering*, 8 (4), 279-291

Ellman, J., I. Kincke y J.Tait (2000), "Word Sense Disambiguation by Filtering and Extraction", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs*, **34 (1-2)**, 127-134

Escandell Vidal, M.V. (2004), *Fundamentos de semántica composicional*, Barcelona, Ariel

Escudero, G., L. Màrquez y G. Rigau (2000), "A Comparison between Supervised Learning Algorithms for Word Sense Disambiguation", en *Proceedings of Fourth Computational Natural Language Learning Workshop (CoNLL-2000)*, Lisbon

Escudero, G., L. Màrquez y G. Rigau (2000), "An Empirical Study of the Domain Dependence of the Supervised Word Sense Disambiguation Systems", *Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'00)*, Hong Kong

Espinal, M.Y. y J. Mateu (2002), "Lexicologia I. La informació semàntica de les unitats lèxiques", en M.T. Espinal (coord.), *Semàntica. Del significat del mot al significat de l´oració*, Barcelona,

Ariel
Federici, F., S. Montemagni, V. Pirrelli (2000), “ROMANSEVAL: Results for Italian by SENSE”, en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs,* **34 (1-2)**
Fellbaum, C. (1998) (ed.), *WordNet: An Electronical Lexical Database*, The MIT Press
Fernández-Amorós, D. (2004), *Anotación Semántica no Supervisada*, tesis doctoral, Universidad Nacional de Educación a Distancia, Madrid
Fernández-Amorós, D., Gonzalo, J. and Verdejo, F. (2001), “The role of conceptual relations in Word Sense Disambiguation“, en *Proceedings of the 6th International Workshop on Applications of Natural Language for Information Systems (NLDB-01)*
Fillmore, Ch.J. (1968), “The case for case” en E. Bach y R.T. Harms (eds.), *Universals in Linguistic Theory*, New York, Holt, Rinehart, and Winston, 1-88
Fillmore, Ch.J. (1982), “Towards a descriptive framework for spatial deixis”, en R.J. Jarvella y W. Klein (eds.), *Speech, Place and Actions; Studies in Deixis and Related Topics*, John Wiley and Sons, New York, 31-59
Fillmore, C.F. y B.T.S. Atkins (2000), “Describing Polysemy: The Case of ‘Crawl’”, en Ravin, Y. y C. Leacock, C. (eds.), *Polysemy: Theoretical and Computational Approaches*, Oxford University Press
Fillmore, Ch.J. y C.F. Baker (2001), “Frame Semantics for Text Understanding”, en *Proceedings of WordNet and Other Lexical Resources Workshop*, NAACL, Pittsburgh
Firth, J.R. (1951), “Modes of meaning”, en *Papers in Linguistics 1934-51*, 190-215, Oxford University Press, Oxford
Florian, R., S. Cucerzan, C. Schafer, D. Yarowsky (2002), “Combining Classifiers for word sense disambiguation”, en *Natural Language Engineering*, 8 (4), 327-341
Gale, W. y K. Church (1991), “Identifying word correspondance in parallel text”, en *Proceedings of the DARPA NLP Workshop*
Gale, W.A., K. W. Church, y D. Yarowsky (1992), “One sense per discourse”, *en Proceedings of DARPA speech and Natural Language Workshop*, Harriman, NY
Gale, W.A., K. W. Church, y D. Yarowsky (1993), “A method for disambiguating word senses in a large corpus”, en *Computer and the Humanities*, **26**, 415-439
Gayral, F. y P. Saint-Dizier (1999), “Peut-on 'couper' `a la polysémie verbale”, en *Proceedings of TALN’99*, Cargese
Gildea, D. (2001), “Corpus Variation and Parser Performance”, en *2001 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Pittsburgh
Gildea, D. y D. Jurafsky (2000), “Automatic Labeling of Semantic Roles”, en *Proceedings of the 38th Annual Conference of the Association for Computational Linguistics (ACL-00)*, Hong Kong; *Computational Linguistics*, **28(3)**, 245-288
Grice, H. P. 1975. "Logic and conversation", en P. Cole y J. L. Morgan (eds.), *Syntax and semantics: Speech acts*, vol. 3, New York, Academia, 41–58
Guthrie, J.A., L. Guthrie, Y. Wilks, H. Aidinejad (1991), “Subject-dependent co-occurrence and word sense disambiguation”, en *Proceedings of the 29th conference on Association for Computational Linguistics*, Berkeley, California, 146-152
Guthrie, L., J. Guthrie, Y. Wilks J. Cowie, D. Farwell, B. Slator y R. Bruce (1992), “A research program on machine-tractable dictionaries and their application to text analysis”, en *Literary and Linguistic Computing*, **8 (4)**
Habert, B. A.Nazarenko y A.Salem (1997), *Les linguistiques de corpus*, Armand Colin, Paris
Hanks, P. (2000), "Do Word Meaning Exist?", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs,* **34 (1-2)**, 205-215
Harabagiu, S., G. Miller, D. Moldovan (1999), "*WordNet*2 - a morphologically and semantically enhanced resource", en *Proceedings of SIGLEX-99*, University of Mariland
Harris, Z. (1954), “Distributional Structure”, en *Word*, **10**, 146-162
Hayes, P.J. (1976), “A process to implement some word-sense disambiguation”, Working paper 23, Institut pour les Etudes Sémantiques et Cognitives, Université de Genève
Hayes, P.J. (1977), *Some Association-based Techniques for Lexical Disambiguation by Machine*, Doctoral dissertation, Département de Mathématiques, Ecole Polytechnique Fédérale de Lausanne

Hearst, M. (1991), “Noun Homograph Disambiguation using Local Context in Large Text Corpora”, en *Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*, Oxford

Hirst, G. (1987), *Semantics Interpretation and the Resolution of the Ambiguity*. *Studies in Natural Language Processing*, Cambridge University Press, Cambridge

Hirst, G. (2000), “Context as a Spurious Concept”, en *Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, 273-287

Hoste, V., I. Hendrickx, W. Daelemans, A. Van den Bosch (2002), “Parameter optimization for machine-learning of word sense disambiguation”, en *Natural Language Engineering,* **8 (4)**, 311-325

Ide, N. (1999), “Parallel translations as sense discriminators”, en *SIGLEX99: Standardizing Lexical Resources*, ACL99 Workshop, College Park, Maryland

Ide, N. (2000), "Cross-Lingual Sense Determination: Can It Work?", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs,* **34 (1-2)**, 223-234

Ide, N. y J. Véronis (eds.) (1998), *Computational Linguistics. Special Issue on Word Sense Disambiguation*, **24 (1)**, 1998

Ide, N. y J. Véronis (1998), "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art", en *Computational Linguistics. Special Issue on Word Sense Disambiguation* , **24(1)**, 1-40

Jorgensen, J. (1990), “The psychological reality of word senses”, en *Journal of Psycholinguistic Research*, **19**: 167-190

Justeson, L.S. y S. Katz (1995), "*Technical terminology: some linguistic properties and an algorithm for identification in text*", en *Natural Language Engineering*, 1, 9-27

Karov, Y. y S.Edelman (1998), "Similarity-based Word Sense Disambiguation", en Ide y Véronis (eds.), *Computational Lingustics. Special Issue on Word Sense Disambiguation* , **24(1)**, 41-59

Katz, J. (1972), *Semantic Theory*, New York, Harper y Row

Katz, J.J., y J.A. Fodor (1963) "The structure of a semantic theory", en *Language*, **39**, 170-210

Kawamoto, A.H. (1988), “Distributed representations of ambiguous words and their resolution in a connectionist network”, en S. Small, G.W. Cottrell y M.K. Tanenhaus (eds.), *Lexical Ambiguity Resolution: Perspectives from Psycholinguistics, Neuropsychology, and Artificial Intelligence,* Morgan Kaufman, San Mateo, CA, 195-228

Kempson, Ruth M. (1977), *Semantic theory*, Cambridge University Press, Cambridge

Kilgarriff, A. (1993), “Dictionary Sense Distinctions: An Enquiry into Their Nature”, *Computers and the Humanities*, **26**, 365-387

Kilgarriff, A. (1997), “I Don´t Believe in Word Senses”, *Computers and the Humanities*, **31(2)**

Kilgarriff, A. (1998), "Bridging the gap between lexicon and corpus: convergence of formalisms", en *LREC'1998 (Proceedings of First International Conference on Language Resources and Evaluation)*, Granada

Kilgarriff, A. (1999), "95% Replicability for Manual Word Sense Tagging", en *Proceedings of EACL'99*, Morgan Kaufman Publishers, San Francisco

Kilgarriff, A y M. Palmer (eds. invitados) (2000), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs*, **34 (1-2)**

Kilgarriff, A y M. Palmer (2000), "Introduction to the Special Issue on SENSEVAL", en Kilgarriff y Palmer (eds.), *Computers and the Humanities*, **34 (1-2)**

Kilgarriff, A. y J. Rozenweig (2000), "Framework and Results for English SENSEVAL", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs,* **34 (1-2)**

Klavans, J. y P. Resnik (1997), “Introduction”, en J .Klavans y P. Resnik 1997 (eds.), *The Balancing Act*, MIT Press, Cambridge

Krowetz, B. (1997), “Homonymy and Polysemy in Information Retrieval,” en *35th Annual Meeting of the Association for Computational Linguistics*, 72-79

Krowetz, B. (1998), “More than One Sense per Discourse”, en *NEC Princeton NJ Labs., Research Memorandum*

Lakoff, G. (1987), *Women, fire and dangerous things ..., what the categories reveal about*

mind,University of Chicago, 1987
Larson, R y G. Segal (1995), Knowledge of Meaning, Cambridge, Massachussets, MIT Press
Leacock, C.y M. Chodorow (1998a), "Combining local context and *WordNet* similarity for word sense identification", en C. Fellbaum (ed.), *WordNet: An Electronic Lexical Database*, MIT Press
Leacock, C., M. Chodorow, G.A. Miler (1998), "Using Corpus Statistics and *WordNet* Relations for Sense Identification", en Ide y Véronis (eds.), *Computational Linguistics. Special Issue on Word Sense Disambiguation*, **24(1)**, 147-166
Leacock, C., G. Towell, y E.M. Voorhees (1993), "Corpus-based statistical sense resolution", en *Proceedings of the ARPA Human Languages Technology Workshop*
Leacock, C., G. Towell, y E.M. Voorhees (1995), "Towards building contextual representations of word senses using statistical models', en B. Boguraev y J. Pustejovsky (eds.), *Corpus Processing for Lexical Acquisition*, Cambridge, MIT Press
Lesk, M. (1986), "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone", en *Proceedings of SIGDOC*, 24-26
Levinson, S.C. (1983), *Pragmatics*, Cambridge, Cambridge University
Li, X., S. Szpakowicz y S. Matwin (1995), "A *WordNet*-based algorithm for word sense disambiguation", en *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, 1368 - 1374
Lin, D. (1997), "Using Syntactic Dependency as Local Context to Resolve Word Sense Ambiguity", en *Proceedings of ACL and EACL'97*, Morgan Kaufman Publishers, San Francisco
Löbner, S. (2002), *Understanding Semantics*, Arnold, London
Lyons, J. (1977), *Semantics*, Cambridge, Cambridge University Press
Magnini, B. y G. Cavaglià (2000), "Integrating Subject Field Codes into *WordNet*", en *Proceedings of LREC-2000*, Athens, 1413-1418
Magnini, B. y C. Strapparava (2001), "Using *WordNet* to improve user modelling in a web document recommender system", en *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh
Magnini, B., C. Strapparava, G. Pezzulo, A. Gliozzo (2002), "The role of domain information in Word Sense Disambiguation", en *Natural Language Engineering*, **8 (4)**, 359-373
Mahesh, K., S. Nirenburg y S. Beale (1997), "If You Have It, Flaunt It: Using Full Ontological Knowledge for Word Sense Disambiguation", en *Ranlp'1997 (Proceedings of International Conference on Recent Advances in Natural Language Processing), Tzigov Chark*, Bulgaria
Manning, C.D. y H. Schütze (1999), *Foundations of Statistical Natural Language Processing*, 3rd printing, MIT Press, Cambridge-London, cap. 7: *Word Sense Disambiguation*, 229-263
Màrquez, L. (2002), "Aprendizaje automático y procesamiento del lenguaje natural", en M.A. Martí y J. Llisteri (eds.), *Tratamiento del lenguaje natural*, Edicions de la Universitat de Barcelona
Màrquez, L., M. Taulé, A. Martí, M. García, F. Real y D. Ferrés (2004a), "Senseval-3: The Catalan lexical sample task", en *Proceedings of SENSEVAL-3*, Barcelona, ACL
Màrquez, L., M. Taulé, A. Martí, M. García, F. Real y D. Ferrés (2004b), "Senseval-3: The Spanish lexical sample task", en *Proceedings of SENSEVAL-3*, Barcelona, ACL
L. Màrquez, M. Taulé, L. Padró, L. Villarejo y M.A. Martí (2004), "On the Quality of Lexical Resources for Word Sense Disambiguation", en vía de publicación en *Proceedings of the EsTAL Conference*, Alicante
Martí, M.A., "Consideraciones sobre la polisemia", en Martí, M.A., G. Vázquez, A. Fernández (eds.) (2002), *Lexicografía computacional y semántica*, Edicions de la Universitat de Barcelona
Martínez, D. y E. Agirre (2000), "One Sense per Collocation and Genre/Topic Variations", en *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large* Corpora, Hong Kong
Martínez D., E. Agirre y L. Màrquez (2002), "Syntactic Features for High Precision Word Sense Disambiguation", en *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taiwan
McClelland, J.L. y D.E. Rumelhart, D. E. (1981), "An interactive activation model of context effects in letter perception. Part 1. An account of basic findings", en *Psychological Review*, *88*(5), 375-407
McEnery, T. y A. Wilson (1996), *Corpus Linguistics*, Edinburgh, Edinburgh University Press
McRoy, S. (1992), "Using Multiple Knowledge Sources for Word Sense Discrimination", en

Computational Linguistics, **18 (1)**, 1-30
Melamed, D. (1997), "Measuring Semantic Entropy", en *Proceedings of the ACL-SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington
Melamed, D. y P. Resnik (2000), " Tagger evaluation given hierarchical tag sets", en *Computers and the Humanities 34(1–2)*
Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller (1953), "Equation of State Calculations by Fast Computing Machines", en *Journal of Chemical Physics*, **21 (6),** 1087-1092
Mihalcea, R. y D. Moldovan (1998), "Word Sense Disambiguation Based on Semantic Density", en *Proceedings of COLING-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, 16-22
Mihalcea, R. y D. Moldovan (1999), "Automatic Acquisition of Sense Tagged Corpora", en *Proceedings of Flairs '99*, Orlando, FL
Mihalcea, R. y D. Moldovan (1999), "An Automatic Method for Generating Sense Tagged Corpora" In: *Proceedings of AAAI '99*, Orlando, FL, 461-466
Mihalcea, R. y D. Moldovan (2000), "An Iterative Approach to Word Sense Disambiguation" In: *Proceedings of Flairs 2000*, Orlando, FL, 219-223
Mihalcea, R. y D. Moldovan (2001a), "Automatic Generation of a Coarse Grained *WordNet*", en *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, Pittsburgh
Mihalcea, R. y D. Moldovan (2001b), "eXtended *WordNet*: progress report", en *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh
Mihalcea, R. y D. Moldovan (2001c), "Highly Accurate Bootstrapping Algorithm for Word Sense Disambiguation", en *International Journal on Artificial Intelligence Tools*, **10 (1-2)**, 5-21
Mihalcea, R. (2002a), "Bootstrapping Large Sense Tagged Corpora", en *Proceedings of the 3rd International Conference on Languages Resources and Evaluations (LREC 2002)*, Las Palmas
Mihalcea, R. (2002b), "Word Sense Disambiguation Using Pattern Learning and Automatic Feature Selection", en *Natural Language and Engineering (JNLE)*
Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross y K.J. Miller (1990), "Introduction to *WordNet*: an on-line lexical database", en *International Journal of Lexicography*, **3 (4)**, 235 – 244
Miller, G.A. y W.G. Charles (1991), "Contextual correlates of semantic similarity", en *Language and Cognitive Processes*, **6(1)**, 1-28
Miller, G.A., M. Chodorow, S. Landes, C. Leacock y R.G. Thomas (1994), "Using a semantic concordance for sense identification", en *Proceedings of the ARPA Workshop on Human Language Technology*, 240-243, Plainsboro, NJ
Miller, G.A. y C. Leacock (2000), "Lexical representations for sentence processing", en Y. Ravin y C. Leacock (2000) (eds.), *Polysemy: Theoretical and Computational Approaches*, Oxford University Press
Miller, G.A., C. Leacock, R. Tengi, y R. Bunker (1993), "A semantic concordance", en *Proceedings of the ARPA Human Language Technology Workshop*, Princeton, NJ, 303-308
Moisl, H. (2000), "NLP Based on Artificial Neural Metworks: Introduction", en Dale *et al.* (eds.) (2000)
Moldovan, D.I. y R. Girju (2001), "An Interactive Tool for the Rapid Development of Knowledge Bases", en *International Journal on Artificial Intelligence Tools*, **10(1-2)**, March 2001
Montemagni S., S. Federici, V. Pirrelli (2000), "Example-based Word Sense Disambiguation: a Paradigm-Driven Approach", *Proceedings of COLING-96*, Copenhagen
Montoyo, A. (2000), "Método basado en Marcas de Especificidad para WSD**",** *Procesamiento del Lenguaje Natural*, **26**, 53-60
Montoyo, A. (2002), *Desambiguación léxica mediante Marcas de Especificidad*, tesis doctoral, Universidad de Alicante
Montoyo, A. y M. Palomar (2000a), "Word Sense Disambiguation with Specification Marks in Unrestricted Texts", en *Proceedings of 2nd International Workshop on Natural Language and Information Systems (NLIS'2000) in conjunction with the 11th International Conference on Database and Expert Systems Applications (DEXA'2000)*, Greenwich
Montoyo, A. y M. Palomar (2000b), "WSD Algorithm applied to a NLP System", 5th International Conference on Application of Natural Language to Information Systems (NLDB´2000), en *Lecture Notes in Computer Science*, Springer-Verlag, volume 1959, 54-65

Montoyo A., M. Palomar (2001), "Specification Marks for Word Sense Disambiguation: New Development", 2nd International conference on Intelligent Text Processing and Computational Linguistics CICLing-2001, en *Lecture Notes in Computer Science*, Springer-Verlag, volume 2004, 182-191

Montoyo A., M. Palomar, G. Rigau (2001), "Method for *WordNet* Enrichment Using WSD", 4th International Conference on Text Speech and Dialogue TSD'2001, *Lecture Notes in Artificial Intelligent*, volume 2166, 180-186

Montoyo, A y A. Suárez (2001), "The University of Alicante word sense disambiguation system", en *Proceedings of Second International Workshop on Evaluating Word Sense Disambiguation Systems, SENSEVAL-2*, ACL, Toulouse

Moon, R. (2000), "Lexicography and Disambigution. The Size of the Problem", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs,* **34 (1-2)**, 99-102

Mooney, R. (1996), "Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning", en *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ACL, New Jersey

Mooney, R. (2003), "Machine Learning", en R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, cap. 20, Oxford University Press

Ng, H.T. (1997), "Getting Serious about Word Sense Disambiguation", en *Proceedings of the ACL-SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington, DC

Ng, H.T. y H.B. Lee (1996), "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-Based Approach", en *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*

Nica, I. (2002a), *Desambiguación semántica automática*, trabajo para la obtención del Diploma de Estudios Avanzados, Universidad de Barcelona (XTRACT-WP-02/04)

Nica, I. (2002b), *Estrategia para la desambiguación semántica de las unidades léxicas de los diccionarios a partir de las relaciones meronímicas*, trabajo para la obtención del Diploma de Estudios Avanzados, Universidad de Barcelona (XTRACT-WP-02/05)

Nica, I., M. A. Martí y A. Montoyo (2003a), "Automatic sense (pre-)tagging by syntagmatic patterns", en *Proceedings of RANLP'03*, Bulgaria, 334-338

Nica, I., M. A. Martí y A. Montoyo (2003b), "Colaboración entre información paradigmática y sintagmática en la Desambiguación Semántica Automática", XIX Congreso de la SEPLN 2003, Alcalá de Henares, publicado en *Procesamiento del Lenguaje Natural*, **31**, 133-140

Nica, I., M. A. Martí, A. Montoyo y S. Vázquez (2004a), "An Unsupervised WSD Algorithm for a NLP System", Congreso de NLDB'04, Reino Unido, publicado en *Lecture Notes in Computer Science*, Springer-Verlag, volume 3136, 288-298

Nica, I., M. A. Martí, A. Montoyo y S. Vázquez (2004b), "Combining *EWN* and corpora for WSD", Congreso de CICLing'04, Korea, publicado en *Lecture Notes in Computer Science*, Springer-Verlag, volume 2945, 188-200

Nica, I., M. A. Martí, A. Montoyo y S. Vázquez (2004c), "Enriching *EWN* with syntagmatic information", en *Proceedings of LREC 2004*, ELRA, Lisabona

Nica, I., M. A. Martí, A. Montoyo y S. Vázquez (2004d), "Intensive Use of Lexicon and Corpus for WSD", Congreso de la SEPLN 2004, Barcelona, publicado en *Procesamiento del Lenguaje Natural*, **33**, 147-154

Nica, I., M. A. Martí, A. Montoyo y S. Vázquez (2004e), "Towards filling the gap between lexicon and corpus in Word Sense Disambiguation", Workshop "Beyond Named Entity Recognition. Semantic labelling for NLP tasks" asociado al Congreso de LREC'04, Portugal

Ooi, V.B.Y. (1998), *Computer Corpus Lexicography*, Edinburgh University Press

Palmer, M. (1998), "Are *WordNet* sense distinctions appropriate for computational lexicons?", en *Proceedings of Senseval, Siglex98*, Brighton

Palmer, M. (2000), "Consistent Criteria for Sense Distinctions", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs,* **34 (1-2)**, 217-222

Patrick, A. (1985), *An Exploration of Abstract Thesaurus Instantiation*, M. Sc. thesis, University of

Kansas, Lawrence, KS
Pedersen, T. (2000), “A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambiguation”, en *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-00)*, Seattle, WA
Pedersen, T. (2002), *A Baseline Methodology for Word Sense Disambiguation*: hhtp://www.d.umn.edu/~tpedersen
Pedersen, T., R. Bruce y J. Wiebe (1997), "Sequential Model Selection for Word Sense Disambiguation", en *Proceedings of Fifth Conference on Applied Natural Language Processing*, Morgan Kaufman Publishers, San Francisco
Pedersen, T., S. Banerjee y S. Patwardhan (2003), “Maximizing Semantic Relatedness to Perform Word Sense Disambiguation”, submitted
Patwardhan, S., S. Banerjee y T. Pedersen (2003), “Using measures of semantic relatedness for word sense disambiguation”, en Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City
Peh, L.S. y H.T. Ng (1997), “Domain-specific semantic class disambiguation using *WordNet*”, en *Proceedings of the 5th Workshop on Very Large Corpora*, Beijing
Pereira, F., N. Tishby, L. Lee (1993), “Distributional Clustering of English Words”, en *Proceedings of the 31st Annual Meeting*, Ohio State University, Association for Computational Linguistics, 183-190
Pereira, F. y N. Tishby (1992), “Distributional similarity, phase transitions and hierarchical clustering”, en *Working Notes of the AAI Symposium on Probabilistic Approaches to Natural Language*, Cambridge, MA, 108-112
Proceedings of SENSEVAL-2. Second International Workshop on Evaluating Word Sense Disambiguation Systems, ACL SIGLEX, Toulouse, 2001
Proceedings of SENSEVAL-3, Barcelona, ACL, 2004
Pustejovsky, J. (1991), “The generative lexicon”, en *Computational Linguistics*, **17 (4),** 409-441
Pustejovsky, J. (1995), *The Generative Lexicon*, MIT Press, Cambridge
Pustejovsky, J. y B. Boguraev (1996) (eds.), *Lexical Semantics: The Problem of Polysemy*, Oxford University Press, Oxford
Pustejovsky, J. y B. Boguraev (1996), “Introduction: Lexical Semantics in Context”, en Pustejovsky y Boguraev (1996) (eds.), *Lexical Semantics: The Problem of Polysemy*, Oxford University Press, Oxford
Pustejovsky, J., B. Boguraev y M. Johnston (1995), “A core lexical engine: The contextual determination of word sense”, Technical Report, Department of Computer Science, Brandeis University
Quillian, M. Ross (1961), “A design for an understanding machine”, en The Semantic Problems in Natural Language colloquium, King's College, Cambridge University, Cambridge, UK
Rada, R., H. Mili, E. Bicknell y M. Blettner (1989), “Development and application of a metric on semantic nets”, en *IEEE Transactions on Systems, Man, and Cibernetics*, **19(1)**, 17-30
Ravin, Y. y C. Leacock (2000) (eds.), *Polysemy: Theoretical and Computational Approaches*, Oxford University Press
Ravin, Y. y C. Leacock (2000), "Polysemy: An overview", en Ravin y Leacock (2000) (eds.)
Resnik, P. (1993), *Selection and Information: A Class-Based Approach to Lexical Relationships*, Doctoral Dissertation, Department of Computer and Information Science, University of Pennsylvania
Resnik, P. (1995), “Using Information Content to Evaluate Semantic Similarity in a Taxonomy”, en *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*
Resnik, P. y D.Yarowsky (1997), "A Perspective on Word Sense Disambiguation Methods and Their Evaluation", en *Proceedings of ACL SiglexWorkshop on Tagging Text with Lexical Semantics: Why, What, and How?*, versión en línea
Resnik, P. y D.Yarowsky, (2000), ”Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation”, en *Natural Language Engineering* **5(2),** 113-133
Rigau, G. (1998), “*Automatic Acquisition of Lexical Knowledge from MRDs*”, tesis doctoral, Universitat Politècnica de Catalunya, Barcelona

Rigau, G., M. Taulé, J. Gonzalo y A. Fernández (2001), "Framework and results for the Spanish SENSEVAL", en *Proceedings of the SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems*, ACL SIGLEX, Toulouse

Rigau, G. (2002), *Resolución automática de la ambigüedad semántica de las palabras*, Fundación Duques de Soria, curso de Tecnologías de la lengua, 2002

Rigau G., B. Magnini, E. Agirre, P. Vossen y J. Carroll (2002), "MEANING: A Roadmap to Knowledge Technologies", en *Proceedings of COLING Workshop "A Roadmap for Computational Linguistics"*, Taipei, Taiwan

Rodríguez, H. (2001), *Lingüística y estadística ¿incompatibles?*, tutorial presentado en el curso de Industrias de la Lengua, Soria, 2001

Rodríguez, H. (2002), *Similitud Semántica*, tutorial presentado en el curso de Industrias de la Lengua, Soria, 2002

Rodríguez, H., S. Climent, P. Vossen, L. Bloksma, W. Peters, A. Alonge, F. Bertagna y A. Roventini (1998), "The Top-Down Strategy for Building *EuroWordNet*: Vocabulary Coverage, Base Concepts and Top Ontology", en N. Ide y D. Greenstein (eds.), *Computers and the Humanities, Double Special Issue on EuroWordNet*, **32 (2-3)**

Rodríguez, H. y M.A. Martí (1998), *Lexicografía Computacional*, Fundación Duques de Soria, curso de Tecnologías de la lengua, 1998

Rosch, E. (1977), "Human Categorization", en Warren, N. (ed.), *Advances in Cross-Cultural Psychology*, vol.7, London, Academic Press

Saussure, F. de (1916), *Cours de linguistique générale*, Paris

Schütze,H. (1992), "Dimensions of Meaning", en *Proceedings of Supercomputing '92*, Los Alamitos, California: IEEE Computer Society Press, 787-96

Schütze,H. (1995), "Word Space", en Hanson, S.J., J.D. Cowan, y C.L. Giles (eds.), *Advances in Neural Information Processing Systems 5*, San Mateo, California, Morgan Kaufmann, 895-902

Schütze, H. (1995), *Ambiguity Resolution and Language Learning: Computational and Cognitive Models*, Ph.D. thesis, Stanford University

Schütze, H. (1997), *Ambiguity Resolution in Language Learning*, Stanford, CSLI

Schütze, H. (1998), "Automatic Word Sense Disambiguation", en N. Ide y J. Véronis (eds.), *Computational Lingustics. Special Issue on Word Sense Disambiguation*, **24(1)**, 97-124

Schütze, H. (2000), "Disambiguation and Connectionism", en Ravin, y C. Leacock (eds.), *Polysemy. Theoretical and Computational Approach*, Oxford University Press, Oxford

Sebastián, N., M.A. Martí, M. F. Carreiras, F. Cuetos Gómez (2000), *Lexesp, léxico informatizado del español*, Edicions de la Universitat de Barcelona

Sedelow, S.Y. y D.W. Mooney (1988), "Knowledge retrieval from domain transcendent expert systems: II. research results", en *Proceedings of the 51st Annual Meeting of the American Society of Information Science*, 209-212

Segond, F. A. Shiller, G. Grefenstette, J.-P. Chanod (1997), "An Experiment in Semantic Tagging using Hidden Markov Model Tagging", versión en línea

Segond, F., E. Aimelet, V. Lux, C. Jean (2000), "Dictionary-Driven Semantic Look-up", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs*, **34 (1-2)**, 193-197

Singleton, D. (2000), *Language and the lexicon: An introduction*, New York, Oxford University Press

Slator, B.M. y Y.A. Wilks (1987), "Towards semantic structures from dictionary entries", en *Proceedings of the 2nd Annual Rocky Mountain Conference on Artificial Intelligence*, Boulder, CO, 85-96

Small, S.L. y C. Rieger (1982), "Parsing and comprehending with word experts (a theory and its realization)", en W. Lenhert y M. Ringle (eds.), *Strategies for Natural Language Processing. Lawrence Erlbaum and Associates*, Hillsdale, NJ, 89-147

Stevenson, M. (1999), "A Corpus-Based Approach to Deriving Lexical Mappings", en *Proceedings of EACL'99*, Morgan Kaufman Publishers, San Francisco

Stevenson, M. (2003), *Word Sense Disambiguation: The Case for Combinations of Knowledge Sources*, CSLI Publications, Stanford

Stevenson, M. y Y. Wilks (1997), "Combining Independent Knowledge Sources for Word Sense Disambiguation", en *Proceddings of the Conference Recent Advances in Natural Language*

Processing, Tzigov Chark, Bulgaria
Stevenson, M. y Y. Wilks (1997), "Sense Tagging: Semantic Tagging with a Lexicon", en *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics"*, Washington
Stevenson, M. y Y. Wilks (1999), "Combining Weak Knowledge Sources for Sense Disambiguation", en *Proceedings of the International Joint Conference for Artificial Intelligence (IJCAI-99)*, Stockholm
Stevenson, M. y Y. Wilks (2000a), "Combining Independent Knowledge Sources for Word Sense Disambiguation", en R. Mitkov (ed.), *Recent Advances in Natural Language Processing*, John Benjamins Publishers
Stevenson, M. y Y. Wilks (2000b), "Large Vocabulary Word Sense Disambiguation", en Y. Ravin y C. Leacock (eds.), *Polysemy. Theoretical and Computational Approach*, Oxford, Oxford University Press
Stevenson, M. y Y. Wilks (2001), "The Interaction of Knowledge Sources in Word Sense Disambiguation", *Computational Linguistics*, 27(3)
Stevenson M. y Y. Wilks (2003), "Word Sense Disambiguation", en R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford University Press
Stubbs, M. (2001), *Words and Phrases: Corpus Studies in Lexical Semantics*, Oxford, Blackwell
Suárez, A. y A. Montoyo (2001), "Estudio de cooperación de métodos de desambiguación léxica: marcas de especificidad *vs.* máxima entropía", en *Procesamiento del Lenguaje Natural*, 27
Suárez, A. y M. Palomar (2002), "A maximum entropy-based word sense disambiguation system", en *Proceedings of the 19th International Conference on Computational Linguistics COLING 2002*, vol. 2, 960-966
Suderman, K. (2000),"Simple Word Sense Disambiguation", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs*, **34 (1-2)**, 165-170
Sussna, M. (1993), "Word sense disambiguation for free-text indexing using a massive semantic network", en *Proceedings of the second international conference on Information and knowledge management*, 67-74, Washington, D.C.
Taulé, M. (2002). *Especificación de los criterios y la metodología seguida en la organización del SENSEVAL-II español*. X-Tract WP-08/02, Universitat de Barcelona
Taulé, M., M. García, N. Artigas y M.A. Martí (2004), "Evaluating lexical resources for WSD", en *Euralex Proceedings*, Le Paquebot, Lorient, France
Tavares da Silva, J.L. y V.L. Strube de Lima (1997), "An Alternative Approach to Lexical Categorial Disambiguation Using a Multi-Agent System Architecture", en *Ranlp'1997 (Proceedings of InternationalConference on Recent Advances in Natural Language Processing), Tzigov Chark,* Bulgaria
Taylor, J.R. (1995), *Linguistic Categorization: Prototypes in Linguistic Theory*, Oxford, Oxford University Press
Teo, E., C. Ting, H.-B. Lee y L.-S. Peh (1997), "Probabilistic Word Sense Disambiguation: A Portable Approach Using Minimum Knowledge", en *Ranlp'1997 (Proceedings of International Conference on Recent Advances in Natural Language Processing),* Tzigov Chark, Bulgaria
Tufis, D. (2002), "Dezambiguizarea automata a cuvintelor din corpusuri paralele folosind echivalentii de traducere", en D. Tufis y F. Gh. Filip (eds.), *Limba Româna în Societatea Informationala - Societatea Cunoasterii*, Editura Expert, Academia Româna, 235-267
Tufis, D. (2004), "Term Translations in Parallel Corpora: Discovery and Consistency Check", en *Proceedings of LREC 2004*, ELRA, Lisabona
Vázquez, S., A. Montoyo, G. Rigau (2003), "Método de desambiguación léxica basada en el recurso léxico Dominios Relevantes", XIX Congreso de la SEPLN 2003, Alcalá de Henares, publicado en *Procesamiento del Lenguaje Natural*, 31
Vázquez, S., R. Romero, A. Suárez, A. Montoyo, I. Nica, A. Martí (2004), "The University of Alicante systems at SENSEVAL-3", en *Proceedings of Senseval-3*, ACL, 243-247
Veenstra, J., A. van den Bosch, S. Bucholz, W. Daelemans y J. Zavrel (2000), "Memory-Based Word Sense Disambiguation", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs*, **34 (1-2)**, 171-177
Véronis, J. (1998), "A study of polysemy judgements and inter-annotator agreement", en *Programme*

and advanced papers of the Senseval workshop, Herstmonceux Castle, Inglaterra
Véronis, J. (2001), "Sense tagging: does it make sense?", trabajo presentado en *Corpus Linguistics'2001 Conference*, Lancaster
Véronis, J. (2000), "Sense tagging: Don't look for the meaning but for the use", *Computational Lexicography and Multimedia Dictionaries (COMLEX'2000)*, Kato Achia, Greece, 1-9
Voorhees, E.M. (1993), "Using *WordNet* to disambiguate word senses for text retrieval", en Korfhage, R., E. y P. Willett (eds.), *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, 171 – 180
Voorhees, E.M., C. Leacock, y G. Towell (1995), "Learning context to disambiguate word senses", en Petsche, T., S.J. Hanson, y J. Shavlik (eds.), *Computational Learning Theory and Natural Learning Systems*, Cambridge, MIT Press
Vossen, P. (1998) (ed.), *EUROWORDNET. A Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Dordrecht
Vossen, P., L. Bloksna, A. Alonge, E. Marinai, C. Peters, I. Castellón, M.A. Martí y G. Rigau (1998), "Compatibility in Interpretation of Relations in *EuroWordNet*", en N. Ide y D. Greenstein (eds.), *Computers and the Humanities*, *Double Special Issue on EuroWordNet*, **32 (2-3)**, 153-184
Walker, D.E. (1987), "Knowledge resource tools for accessing large text files", en S. Nirenburg (ed.), *Machine Translation: Theoretical and Methodological Issues*, Cambridge, Cambridge University Press, 247-261
Walker, D.E. y R.A. Amsler (1986), "The use of machine-readable dictionaries in sublanguage analysis", en R. Grishman y R. Kittedge (eds.), *Analysing Language in restricted domains*, Lawrence Erlbaum, Hillsdale, NJ
Wanner, L., M. Alonso y A. Martí (2004), "Enriching the Spanish *EuroWordNet* by Collocations", en *Proceedings of LREC 2004*, ELRA, Lisabona
Weaver, W. (1955), "Translation", en W.N. Locke y A.D. Booth (eds.), *Machine Translation of Languages*, John Wiley & Sons, New York, 15-23
Weinreich, U. (1964), "Webster's Third: A Critique of its Semantics", en *International Journal of American Linguistic*, 30, 405-409
Wierzbicka, A. (1989), "*Semantic Primitives and Lexical Universals*", en *Quaderni di Semantica*, **10, 1**, Bologna, Il Mulino
Wilks, Y. (1973), "An artificial intelligence approach to machine translation", en R. Schank y K. Colby (eds.), *Computer Models of Thought and Language*, W. H. Freeman, San Francisco, 114-151
Wilks, Y. (1975), "Preference semantics", en E. L. Keenan III (ed.), *Formal Semantics of Natural Language*, Cambridge University Press, 329-348
Wilks, Y. (1997), "Senses and Texts", en *Computers and the Humanities*, **31**, 77-90
Wilks, Y. (2000), "Is Word Sense Disambiguation Just One More NLP Task?", en Kilgarriff y Palmer (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs*, **34 (1-2)**, 235-243
Wilks, Y., D. Fass, C.M. Guo, J. McDonald, T. Plate, y B. Slator (1990), *A tractable machine dictionary as a basis for computational semantics, Journal of Machine Translation*, **5**, 99-154
Wilks, Y., D. Fass, C.M. Guo, J. McDonald, T. Plate, y B. Slator (1993), *Machine tractable dictionary tools*, en J. Pustejovsky (ed.), *Semantics and the Lexicon*, Dordrecht, Kluwer
Wilks, Y.A., B.M. Slator, y L.M. Guthrie (1996), *Electric Words. Dictionaries, Computers and Meaning*, The MIT Press, Cambridge-London
Wilks, Y. y M. Stevenson (1997), "Sense Tagging: Semantic Tagging with a Lexicon", en *Proceedings of ACL SiglexWorkshop on Tagging Text with Lexical Semantics: Why, What, and How?*, versión en línea
Wilks, Y. y M. Stevenson (1997), "Combining Independent Knowledge Sources for Word Sense Disambiguation", en *Ranlp'1997 (Proceedings of International Conference on Recent Advances in Natural Language Processing)*, Tzigov Chark, Bulgaria
Wilks, Y. y M. Stevenson (2000), "Word Sense Disambiguation", en R. Mitkov (ed.), *Recent Advances in Natural Language Processing*, John Benjamins Publishers
Wilson, A. y J.Thomas (1997), "Semantic Annotation", en R. Garside *et al.* (eds.), *Corpus Annotation. Linguistic Information from Computer Text Corpora*, Longman, London-New York
Wittgenstein, L. (1953), *Philosophical Investigations*, New York, MacMillan

Yarowsky, D. (1992), “Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora”, en *Proceedings of COLING-92*

Yarowsky, D. (1993), “One Sense per Collocation”, en *DARPA Workshop on Human Language Technology*, Princeton, NJ, 266-271

Yarowsky, D. (1994), “Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French”, en *Proceedings of the 32nd Meeting of the Association for Computational Linguistics*, Las Cruces, NM, 88-95

Yarowsky, D. (1995), “Unsupervised word sense disambiguation rivaling supervised methods”, en *Proceedings of the 33rd Conference of the Association for Computational Linguistics*, Cambridge, Massachusetts, 189-196

Yarowsky, D. (2000a), "Hierarchical Decision Lists for Word Sense Disambiguation", en N. Ide y E. Mylonas (eds.), *Computers and the Humanities. Special Issue: Evaluating Word Sense Disambiguation Programs,* **34 (1-2)**, 179-186

Yarowsky, D. (2000b), "Word Sense Disambiguation", en Dale, R., H. Moisl, y H. Somers (eds.), *Handbook of Natural Language Processing*, Marcel Dekker, New York-Basel, 2000, 629-654

Yarowsky, D. y R. Florian (2002), “Evaluating sense disambiguation across diverse parameter spaces”, en *Natural Language Engineering*, 8 (4), 293-310

Yngve, V.H. (1955), “Syntax and the problem of multiple meaning”, en W.N. Locke y A.D. Booth (eds.), *Machine Translation of Languages*, John Wiley & Sons, New York, 208-226

Yokoi, T (1995), “The EDR Electronic Dictionary”, en *Communications of the ACM*, **38(11)**

Zadeh, L.A. (1982), “Test-score semantics for natural languages and meaning-representation via PRUF”, en *Proceedings of COLING 82*, Prague, 425-430

Zavrel, J. y W. Daelemans (1997), "Memory-Based Learning: Using Similarity for Smoothing", en *Proceedings of the 35th Annual Meeting of the ACL*, Madrid (ACl/EACL'97)

Zernik, U. (1991), “Train1 *vs.* Train2: Tagging word senses in a corpus”, en *Proceedings of Intelligent Text and Image Handling, RIAO'91*, 567-585, Barcelona

ANEXOS

Anexo 1. Esquemas de búsqueda

I. Generales:

Tipos básicos de patrones morfosintácticos en que puede intervenir un nombre:

P1

N1 ADV ADJ**,** DET DET **N2**
N1 ADV VPART**,** DET DET **N2**
N1 ADJ **,** DET DET **N2**
N1 VPART **,** DET DET **N2**
N1 , DET DET **N2**

N1 , DET DET ADV ADJ **N2**
N1 , DET DET ADV VPART **N2**
N1 , DET DET ADJ **N2**
N1 , DET DET PART **N2**
N1 , DET DET **N2**

N1 ADV ADV ADJ**,** DET DET **N2**
N1 ADV ADV VPART**,** DET DET **N2**

N1 , DET DET ADV ADV ADJ **N2**
N1 , DET DET ADV ADV VPART **N2**

- Para el patrón P2 = [N CONJ* N], donde: CONJ* = {*y, o, e, u, ni*} (conjunciones coordinativas)

N1 ADV ADJ **CONJ*** DET **N2**
N1 ADV VPART **CONJ*** DET **N2**
N1 ADJ **CONJ*** DET **N2**
N1 VPART **CONJ*** DET **N2**
N1 CONJ* DET **N2**
N1 ADV ADJ **CONJ* N2**
N1 ADV VPART **CONJ* N2**
N1 ADJ **CONJ* N2**
N1 VPART **CONJ* N2**
N1 CONJ* N2

N1 CONJ* DET ADV ADJ **N2**
N1 CONJ* DET ADV VPART **N2**
N1 CONJ* ADV ADJ **N2**
N1 CONJ* ADV VPART **N2**
N1 CONJ* DET ADJ **N2**
N1 CONJ* DET VPART **N2**
N1 CONJ* ADJ **N2**
N1 CONJ* VPART **N2**
N1 CONJ* DET **N2**

N1 ADV ADJ **CONJ*** DET DET **N2**
N1 ADV VPART **CONJ*** DET DET **N2**
N1 ADJ **CONJ*** DET DET **N2**
N1 VPART **CONJ*** DET DET **N2**
N1 CONJ* DET DET **N2**

N1 CONJ* DET DET ADV ADJ **N2**
N1 CONJ* DET DET ADV VPART **N2**
N1 CONJ* DET DET ADJ **N2**
N1 CONJ* DET DET VPART **N2**
N1 CONJ* DET DET **N2**

N1 ADV ADV ADJ **CONJ*** DET **N2**
N1 ADV ADV VPART **CONJ*** DET **N2**
N1 ADV ADV ADJ **CONJ* N2**
N1 ADV ADV VPART **CONJ* N2**

N1 CONJ* DET ADV ADV ADJ **N2**

N1 CONJ* DET ADV ADV VPART **N2**
N1 CONJ* ADV ADV ADJ **N2**
N1 CONJ* ADV ADV VPART **N2**

N1 ADV ADV ADJ **CONJ*** DET DET **N2**
N1 ADV ADV VPART **CONJ*** DET DET **N2**

N1 CONJ* DET DET ADV ADV ADJ **N2**
N1 CONJ* DET DET ADV ADV VPART **N2**

- Para el patrón P3 = [N PREP N]:

N1 PREP DET ADV ADJ **N2**
N1 PREP DET ADV VPART **N2**
N1 PREP DET ADJ **N2**
N1 PREP DET VPART **N2**
N1 PREP ADJ **N2**
N1 PREP VPART **N2**
N1 PREP DET **N2**
N1 PREP N2

N1 PREP DET DET ADV ADJ **N2**
N1 PREP DET DET ADV VPART **N2**
N1 PREP DET DET ADJ **N2**
N1 PREP DET DET VPART **N2**
N1 PREP DET DET **N2**

N1 PREP DET ADV ADV ADJ **N2**
N1 PREP DET ADV ADV VPART **N2**

N1 PREP DET DET ADV ADV ADJ **N2**
N1 PREP DET DET ADV ADV VPART **N2**

- Para el patrón P4 = [N ADJ] & [ADJ N]:

N ADJ
N ADV **ADJ**

N ADV ADV **ADJ**

N ADV ADJ1 CONJ* ADV **ADJ2**
N ADV VPART CONJ* ADV **ADJ**
N ADJ1 CONJ* ADV **ADJ2**
N VPART CONJ* ADV **ADJ**
N ADV ADJ1 CONJ* **ADJ2**
N ADV VPART CONJ* **ADJ**
N ADJ1 CONJ* **ADJ2**
N VPART CONJ* **ADJ**

N ADV ADV ADJ1 CONJ* ADV **ADJ2**
N ADV ADV VPART CONJ* ADV **ADJ**
N ADV ADV ADJ1 CONJ* **ADJ2**
N ADV ADV VPART CONJ* **ADJ**

N ADV ADJ1 CONJ* ADV ADV **ADJ2**
N ADV VPART CONJ* ADV ADV **ADJ**
N ADJ1 CONJ* ADV ADV **ADJ2**
N VPART CONJ* ADV ADV **ADJ**

N ADV ADV ADJ1 CONJ* ADV ADV **ADJ2**
N ADV ADV VPART CONJ* ADV ADV **ADJ**

N ADJ1 **ADJ2**
N ADV ADJ1 **ADJ2**
N ADJ1 ADV **ADJ2**

ADJ2 CONJ* ADV ADJ1 **N**
ADJ CONJ* ADV VPART **N**
ADJ2 CONJ* ADJ1 **N**
ADJ CONJ* VPART **N**
ADJ N

ADJ2 CONJ* ADV ADV ADJ1 **N**
ADJ CONJ* ADV ADV VPART **N**

- Para el patrón P5 = [N VPART] & [VPART N]:

N VPART
N ADV **VPART**

N ADV ADV **VPART**

N ADV ADJ CONJ* ADV **VPART**
N ADV VPART1 CONJ* ADV **VPART2**
N ADJ CONJ* ADV **VPART**
N VPART1 CONJ* ADV **VPART2**
N ADV ADJ CONJ* **VPART**
N ADV VPART1 CONJ* **VPART2**
N ADJ CONJ* **VPART**
N VPART1 CONJ* **VPART2**

N ADV ADV ADJ CONJ* ADV **VPART**
N ADV ADV VPART1 CONJ* ADV **VPART2**
N ADV ADV ADJ CONJ* **VPART**
N ADV ADV VPART1 CONJ* **VPART2**

N ADV ADJ CONJ* ADV ADV **VPART**
N ADV1 VPART1 CONJ* ADV ADV **VPART2**
N ADJ CONJ* ADV ADV **VPART**
N VPART1 CONJ* ADV ADV **VPART2**

N ADV ADV ADJ CONJ* ADV ADV **VPART**
N ADV ADV VPART1 CONJ* ADV ADV **VPART2**

N ADJ **VPART**
N ADV ADJ **VPART**
N ADJ ADV **VPART**
VPART CONJ* ADV ADJ **N**
VPART2 CONJ* ADV VPART1 **N**
VPART CONJ* ADJ **N**
VPART2 CONJ* VPART1 **N**
VPART N

VPART CONJ* ADV ADV ADJ **N**
VPART2 CONJ* ADV ADV VPART1 **N**

II. Para un nombre fijado (*lema0*)

Tipos básicos de patrones morfosintácticos en que puede intervenir el nombre *lema0*

A la derecha:
- P1 = [*lema0*-N, N];
- P2 = [*lema0*-N CONJ* N], donde: CONJ* = {*y, o, e, u, ni*} (conjunciones coordinativas);
- P3 = [*lema0*-N PREP N];
- P4 = [*lema0*-N ADJ];
- P5 = [*lema0*-N VPART];

A la izquierda:
- P6 = [N, *lema0*-N];
- P7 = [N CONJ* *lema0*-N], donde: CONJ* = {*y, o, e, u, ni*} (conjunciones coordinativas);
- P8 = [N PREP *lema0*-N];
- P9 = [ADJ *lema0*-N];
- P10 = [VPART *lema0*-N],

Esquemas de búsqueda para la identificación de los patrones léxico-sintácticos del nombre *lema0*

A la derecha:

- Para el patrón P1 = [*lema0*-N, N]:

***lema0*-N** ADV ADJ**,** DET **N**
***lema0*-N** ADV VPART**,** DET **N**
***lema0*-N** ADJ **,** DET **N**
***lema0*-N** VPART **,** DET **N**
***lema0*-N , DET N**
***lema0*-N** ADV ADJ **, N**
***lema0*-N** ADV VPART **, N**
***lema0*-N** ADJ **, N**
***lema0*-N** VPART **, N**
***lema0*-N , N**

***lema0*-N ,** DET ADV ADJ **N**
***lema0*-N ,** DET ADV VPART **N**
***lema0*-N ,** ADV ADJ **N**
***lema0*-N ,** ADV VPART **N**
***lema0*-N ,** DET ADJ **N**
***lema0*-N ,** DET PART **N**
***lema0*-N ,** ADJ **N**
***lema0*-N ,** PART **N**
***lema0*-N ,** DET **N**

***lema0*-N** ADV ADV ADJ**,** DET **N**
***lema0*-N** ADV ADV VPART**,** DET **N**
***lema0*-N** ADV ADV ADJ **, N**
***lema0*-N** ADV ADV VPART **, N**

***lema0*-N ,** DET ADV ADV ADJ **N**
***lema0*-N,** DET ADV ADV VPART **N**

***lema0*-N , ADV ADV ADJ N**
***lema0*-N , ADV ADV VPART N**

***lema0*-N ADV ADJ, DET DET N**
***lema0*-N ADV VPART, DET DET N**
***lema0*-N ADJ , DET DET N**
***lema0*-N VPART , DET DET N**
***lema0*-N , DET DET N**

***lema0*-N , DET DET ADV ADJ N**
***lema0*-N , DET DET ADV VPART N**
***lema0*-N , DET DET ADJ N**
***lema0*-N , DET DET PART N**
***lema0*-N , DET DET N**

***lema0*-N ADV ADV ADJ, DET DET N**
***lema0*-N ADV ADV VPART, DET DET N**

***lema0*-N , DET DET ADV ADV ADJ N**
***lema0*-N , DET DET ADV ADV VPART N**

- Para el patrón P2 = [*lema0*-N CONJ* N], donde: CONJ* = {*y, o, e, u, ni*}:

***lema0*-N ADV ADJ CONJ* DET N**
***lema0*-N ADV VPART CONJ* DET N**
***lema0*-N ADJ CONJ* DET N**
***lema0*-N VPART CONJ* DET N**
***lema0*-N CONJ* DET N**
***lema0*-N ADV ADJ CONJ* N**
***lema0*-N ADV VPART CONJ* N**
***lema0*-N ADJ CONJ* N**
***lema0*-N VPART CONJ* N**
***lema0*-N CONJ* N**

***lema0*-N CONJ* DET ADV ADJ N**
***lema0*-N CONJ* DET ADV VPART N**
***lema0*-N CONJ* ADV ADJ N**
***lema0*-N CONJ* ADV VPART N**
***lema0*-N CONJ* DET ADJ N**
***lema0*-N CONJ* DET VPART N**
***lema0*-N CONJ* ADJ N**
***lema0*-N CONJ* VPART N**
***lema0*-N CONJ* DET N**

***lema0*-N ADV ADJ CONJ* DET DET N**
***lema0*-N ADV VPART CONJ* DET DET N**
***lema0*-N ADJ CONJ* DET DET N**
***lema0*-N VPART CONJ* DET DET N**
***lema0*-N CONJ* DET DET N**

***lema0*-N CONJ* DET DET ADV ADJ N**
***lema0*-N CONJ* DET DET ADV VPART N**
***lema0*-N CONJ* DET DET ADJ N**
***lema0*-N CONJ* DET DET VPART N**
***lema0*-N CONJ* DET DET N**

***lema0*-N ADV ADV ADJ CONJ* DET N**
***lema0*-N ADV ADV VPART CONJ* DET N**
***lema0*-N ADV ADV ADJ CONJ* N**

***lema0*-N** ADV ADV VPART **CONJ* N**

lema0*-N CONJ DET ADV ADV ADJ **N**
lema0*-N CONJ DET ADV ADV VPART **N**
lema0*-N CONJ ADV ADV ADJ **N**
lema0*-N CONJ ADV ADV VPART **N**

lema0*-N** ADV ADV ADJ **CONJ DET DET **N**
lema0*-N** ADV ADV VPART **CONJ DET DET **N**

lema0*-N CONJ DET DET ADV ADV ADJ **N**
lema0*-N CONJ DET DET ADV ADV VPART **N**

- Para el patrón P3 = [*lema0*-N PREP N]:

***lema0*-N PREP** DET ADV ADJ **N**
***lema0*-N PREP** DET ADV VPART **N**
***lema0*-N PREP** DET ADJ **N**
***lema0*-N PREP** DET VPART **N**
***lema0*-N PREP** ADJ **N**
***lema0*-N PREP** VPART **N**
***lema0*-N PREP** DET **N**
***lema0*-N PREP N**

***lema0*-N PREP** DET DET ADV ADJ **N**
***lema0*-N PREP** DET DET ADV VPART **N**
***lema0*-N PREP** DET DET ADJ **N**
***lema0*-N PREP** DET DET VPART **N**
***lema0*-N PREP** DET DET **N**

***lema0*-N PREP** DET ADV ADV ADJ **N**
***lema0*-N PREP** DET ADV ADV VPART **N**

***lema0*-N PREP** DET DET ADV ADV ADJ **N**
***lema0*-N PREP** DET DET ADV ADV VPART **N**

- Para el patrón P4 = [*lema0*-N ADJ]:

***lema0*-N ADJ**
***lema0*-N** ADV **ADJ**

***lema0*-N** ADV ADV **ADJ**

***lema0*-N** ADV ADJ1 CONJ* ADV **ADJ2**
***lema0*-N** ADV VPART CONJ* ADV **ADJ**
***lema0*-N** ADJ1 CONJ* ADV **ADJ2**
***lema0*-N** VPART CONJ* ADV1 **ADJ**
***lema0*-N** ADV ADJ1 CONJ* **ADJ2**
***lema0*-N** ADV VPART CONJ* **ADJ**
***lema0*-N** ADJ1 CONJ* **ADJ2**
***lema0*-N** VPART CONJ* **ADJ**

***lema0*-N** ADV ADV ADJ1 CONJ* ADV **ADJ2**
***lema0*-N** ADV ADV VPART CONJ* ADV **ADJ**
***lema0*-N** ADV ADV ADJ1 CONJ* **ADJ2**
***lema0*-N** ADV ADV VPART CONJ* **ADJ**

***lema0*-N** ADV ADJ1 CONJ* ADV ADV **ADJ2**

***lema0*-N** ADV VPART CONJ* ADV ADV **ADJ**
***lema0*-N** ADJ1 CONJ* ADV ADV **ADJ2**
***lema0*-N** VPART CONJ* ADV ADV **ADJ**

***lema0*-N** ADV ADV ADJ1 CONJ* ADV ADV **ADJ2**
***lema0*-N** ADV ADV VPART CONJ* ADV ADV **ADJ**

***lema0*-N** ADJ1 **ADJ2**
***lema0*-N** ADV ADJ1 **ADJ2**
***lema0*-N** ADJ1 ADV **ADJ2**

- Para el patrón P5 = [*lema0*-N VPART]:

***lema0*-N VPART**
***lema0*-N** ADV **VPART**

***lema0*-N** ADV ADV **VPART**

***lema0*-N** ADV ADJ1 CONJ* ADV2 **VPART**
***lema0*-N** ADV VPART1 CONJ* ADV **VPART2**
***lema0*-N** ADJ CONJ* ADV **VPART**
***lema0*-N** VPART1 CONJ* ADV **VPART2**
***lema0*-N** ADV ADJ CONJ* **VPART**
***lema0*-N** ADV VPART1 CONJ* **VPART2**
***lema0*-N** ADJ CONJ* **VPART**
***lema0*-N** VPART1 CONJ* **VPART2**

***lema0*-N** ADV ADV ADJ CONJ* ADV **VPART**
***lema0*-N** ADV ADV VPART1 CONJ* ADV **VPART2**
***lema0*-N** ADV ADV ADJ CONJ* **VPART**
***lema0*-N** ADV ADV VPART1 CONJ* **VPART2**

***lema0*-N** ADV ADJ CONJ* ADV ADV **VPART**
***lema0*-N** ADV VPART1 CONJ* ADV ADV **VPART2**
***lema0*-N** ADJ CONJ* ADV ADV **VPART**
***lema0*-N** VPART1 CONJ* ADV ADV **VPART2**

***lema0*-N** ADV ADV ADJ CONJ* ADV ADV **VPART**
***lema0*-N** ADV ADV VPART1 CONJ* ADV ADV **VPART2**

***lema0*-N** ADJ **VPART**
***lema0*-N** ADV ADJ **VPART**
***lema0*-N** ADJ ADV **VPART**

A la izquierda:

- Para el patrón P6 = [N, *lema0*-N]:

N ADV ADJ**,** DET ***lema0*-N**
N ADV VPART**,** DET ***lema0*-N**
N ADJ **,** DET ***lema0*-N**
N VPART **,** DET ***lema0*-N**
N , DET ***lema0*-N**
N ADV ADJ **,** ***lema0*-N**
N ADV VPART **,** ***lema0*-N**
N ADJ **,** ***lema0*-N**
N VPART **,** ***lema0*-N**

N , *lema0*-N

N , DET ADV ADJ ***lema0*-N**
N , DET ADV VPART ***lema0*-N**
N , ADV ADJ ***lema0*-N**
N , ADV VPART ***lema0*-N**
N , DET ADJ ***lema0*-N**
N , DET PART ***lema0*-N**
N , ADJ ***lema0*-N**
N , PART ***lema0*-N**
N , DET ***lema0*-N**

N ADV ADV ADJ**,** DET ***lema0*-N**
N ADV ADV VPART**,** DET ***lema0*-N**
N ADV ADV ADJ **,** ***lema0*-N**
N ADV ADV VPART **,** ***lema0*-N**

N , DET ADV ADV ADJ ***lema0*-N**
N , DET ADV ADV VPART ***lema0*-N**
N , ADV ADV ADJ ***lema0*-N**
N , ADV ADV VPART ***lema0*-N**

N ADV ADJ**,** DET DET ***lema0*-N**
N ADV VPART**,** DET DET ***lema0*-N**
N ADJ **,** DET DET ***lema0*-N**
N VPART **,** DET DET ***lema0*-N**
N , DET DET ***lema0*-N**

N , DET DET ADV ADJ ***lema0*-N**
N , DET DET ADV VPART ***lema0*-N**
N , DET DET ADJ ***lema0*-N**
N , DET DET PART ***lema0*-N**
N , DET DET ***lema0*-N**

N ADV ADV ADJ**,** DET DET ***lema0*-N**
N ADV ADV VPART**,** DET DET ***lema0*-N**

N , DET DET ADV ADV ADJ ***lema0*-N**
N , DET DET ADV ADV VPART ***lema0*-N**

- Para el patrón P7 = [N CONJ* *lema0*-N], donde: CONJ* = {*y, o, e, u, ni*}:

N ADV ADJ **CONJ*** DET ***lema0*-N**
N ADV VPART **CONJ*** DET ***lema0*-N**
N ADJ **CONJ*** DET ***lema0*-N**
N VPART **CONJ*** DET ***lema0*-N**
N CONJ* DET ***lema0*-N**
N ADV ADJ **CONJ*** ***lema0*-N**
N ADV VPART **CONJ*** ***lema0*-N**
N ADJ **CONJ*** ***lema0*-N**
N VPART **CONJ*** ***lema0*-N**
N CONJ* ***lema0*-N**

N CONJ* DET ADV ADJ ***lema0*-N**
N CONJ* DET ADV VPART ***lema0*-N**
N CONJ* ADV ADJ ***lema0*-N**
N CONJ* ADV VPART ***lema0*-N**
N CONJ* DET ADJ ***lema0*-N**
N CONJ* DET VPART ***lema0*-N**

N CONJ* ADJ ***lema0*-N**
N CONJ* VPART ***lema0*-N**
N CONJ* DET ***lema0*-N**

N ADV ADJ **CONJ*** DET DET ***lema0*-N**
N ADV VPART **CONJ*** DET DET ***lema0*-N**
N ADJ **CONJ*** DET DET ***lema0*-N**
N VPART **CONJ*** DET DET ***lema0*-N**
N CONJ* DET DET ***lema0*-N**

N CONJ* DET DET ADV ADJ ***lema0*-N**
N CONJ* DET DET ADV VPART ***lema0*-N**
N CONJ* DET DET ADJ ***lema0*-N**
N CONJ* DET DET VPART ***lema0*-N**
N CONJ* DET DET ***lema0*-N**

N ADV ADV ADJ **CONJ*** DET ***lema0*-N**
N ADV ADV VPART **CONJ*** DET ***lema0*-N**
N ADV ADV ADJ **CONJ*** ***lema0*-N**
N ADV ADV VPART **CONJ*** ***lema0*-N**

N CONJ* DET ADV ADV ADJ ***lema0*-N**
N CONJ* DET ADV ADV VPART ***lema0*-N**
N CONJ* ADV ADV ADJ ***lema0*-N**
N CONJ* ADV ADV VPART ***lema0*-N**

N ADV ADV ADJ **CONJ*** DET DET ***lema0*-N**
N ADV ADV VPART **CONJ*** DET DET ***lema0*-N**

N CONJ* DET DET ADV ADV ADJ ***lema0*-N**
N CONJ* DET DET ADV ADV VPART ***lema0*-N**

- Para el patrón P8 = [N PREP *lema0*-N]:

N PREP DET ADV ADJ ***lema0*-N**
N PREP DET ADV VPART ***lema0*-N**
N PREP DET ADJ ***lema0*-N**
N PREP DET VPART ***lema0*-N**
N PREP ADJ ***lema0*-N**
N PREP VPART ***lema0*-N**
N PREP DET ***lema0*-N**
N PREP *lema0*-N

N PREP DET DET ADV ADJ ***lema0*-N**
N PREP DET DET ADV VPART ***lema0*-N**
N PREP DET DET ADJ ***lema0*-N**
N PREP DET DET VPART ***lema0*-N**
N PREP DET DET ***lema0*-N**

N PREP DET ADV ADV ADJ ***lema0*-N**
N PREP DET ADV ADV VPART ***lema0*-N**

N PREP DET DET ADV ADV ADJ ***lema0*-N**
N PREP DET DET ADV ADV VPART ***lema0*-N**

- Para el patrón P9 = [ADJ *lema0*-N]:

ADJ2 CONJ* ADV ADJ1 ***lema0*-N**

ADJ CONJ* ADV VPART ***lema0*-N**
ADJ2 CONJ* ADJ1 ***lema0*-N**
ADJ CONJ* VPART ***lema0*-N**
ADJ *lema0*-N

ADJ2 CONJ* ADV ADV ADJ1 ***lema0*-N**
ADJ CONJ* ADV ADV VPART ***lema0*-N**

- Para el patrón P10 = [VPART *lema0*-N]:

VPART CONJ* ADV ADJ ***lema0*-N**
VPART2 CONJ* ADV VPART1 ***lema0*-N**
VPART CONJ* ADJ ***lema0*-N**
VPART2 CONJ* VPART1 ***lema0*-N**
VPART *lema0*-N

VPART CONJ* ADV ADV ADJ ***lema0*-N**
VPART2 CONJ* ADV ADV VPART1 ***lema0*-N**

Anexo 2. Reglas de descomposición

I. Generales

N1 ADV ADJ**,** DET **N2**	→ **N1** ADJ	+	**N1 , N2**
N1 ADV VPART**,** DET **N2**	→ **N1** VPART	+	**N1 , N2**
N1 ADJ **,** DET **N2**	→ **N1** ADJ	+	**N1 , N2**
N1 VPART **,** DET **N2**	→ **N1** VPART	+	**N1 , N2**
N1 , DET **N2**	→ **N1 , N2**		
N1 ADV ADJ **, N2**	→ **N1** ADJ	+	**N1 , N2**
N1 ADV VPART **, N2**	→ **N1** VPART	+	**N1 , N2**
N1 ADJ **, N2**	→ **N1** ADJ	+	**N1 , N2**
N1 VPART **, N2**	→ **N1** VPART	+	**N1 , N2**
N1 , N2	→ **N1 , N2**		
N1 , DET ADV ADJ **N2**	→ **N1 , N2**	+	ADJ **N2**
N1 , DET ADV VPART **N2**	→ **N1 , N2**	+	VPART **N2**
N1 , ADV ADJ **N2**	→ **N1 , N2**	+	ADJ **N2**
N1 , ADV VPART **N2**	→ **N1 , N2**	+	VPART **N2**
N1 , DET ADJ **N2**	→ **N1 , N2**	+	ADJ **N2**
N1 , DET PART **N2**	→ **N1 , N2**	+	VPART **N2**
N1 , ADJ **N2**	→ **N1 , N2**	+	ADJ **N2**
N1 , PART **N2**	→ **N1 , N2**	+	VPART **N2**
N1 , DET **N2**	→ **N1 , N2**		
N1 ADV ADV ADJ**,** DET **N2**	→ **N1 , N2**	+	**N1** ADJ
N1 ADV ADV VPART**,** DET **N2**	→ **N1 , N2**	+	**N1** VPART
N1 ADV ADV ADJ **, N2**	→ **N1 , N2**	+	**N1** ADJ
N1 ADV ADV VPART **, N2**	→ **N1 , N2**	+	**N1** VPART
N1 , DET ADV ADV ADJ **N2**	→ **N1 , N2**	+	ADJ **N2**
N1 , DET ADV ADV VPART **N2**	→ **N1 , N2**	+	VPART **N2**
N1 , ADV ADV ADJ **N2**	→ **N1 , N2**	+	ADJ **N2**
N1 , ADV ADV VPART **N2**	→ **N1 , N2**	+	VPART **N2**
N1 ADV ADJ**,** DET DET **N2**	→ **N1 , N2**	+	**N1** ADJ
N1 ADV VPART**,** DET DET **N2**	→ **N1 , N2**	+	**N1** VPART
N1 ADJ **,** DET DET **N2**	→ **N1 , N2**	+	**N1** ADJ
N1 VPART **,** DET DET **N2**	→ **N1 , N2**	+	**N1** VPART
N1 , DET DET **N2**	→ **N1 , N2**		
N1 , DET DET ADV ADJ **N2**	→ **N1 , N2**	+	ADJ **N2**
N1 , DET DET ADV VPART **N2**	→ **N1 , N2**	+	VPART **N2**
N1 , DET DET ADJ **N2**	→ **N1 , N2**	+	ADJ **N2**
N1 , DET DET PART **N2**	→ **N1 , N2**	+	VPART **N2**
N1 , DET DET **N2**	→ **N1 , N2**		
N1 ADV ADV ADJ**,** DET DET **N2**	→ **N1 , N2**	+	**N1** ADJ
N1 ADV ADV VPART**,** DET DET **N2**	→ **N1 , N2**	+	**N1** VPART
N1 , DET DET ADV ADV ADJ **N2**	→ **N1 , N2**	+	ADJ **N2**
N1 , DET DET ADV ADV VPART **N2**	→ **N1 , N2**	+	VPART **N2**

N1 ADV ADJ **CONJ*** DET **N2**	→ **N1 CONJ* N2**	+	**N1** ADJ
N1 ADV VPART **CONJ*** DET **N2**	→ **N1 CONJ* N2**	+	**N1** VPART
N1 ADJ **CONJ*** DET **N2**	→ **N1 CONJ* N2**	+	**N1** ADJ
N1 VPART **CONJ*** DET **N2**	→ **N1 CONJ* N2**	+	**N1** VPART
N1 CONJ* DET **N2**	→ **N1 CONJ* N2**		
N1 ADV ADJ **CONJ* N2**	→ **N1 CONJ* N2**	+	**N1** ADJ
N1 ADV VPART **CONJ* N2**	→ **N1 CONJ* N2**	+	**N1** VPART
N1 ADJ **CONJ* N2**	→ **N1 CONJ* N2**	+	**N1** ADJ
N1 VPART **CONJ* N2**	→ **N1 CONJ* N2**	+	**N1** VPART
N1 CONJ* N2	→ **N1 CONJ* N2**		
N1 CONJ* DET ADV ADJ **N2**	→ **N1 CONJ* N2**	+	ADJ **N2**
N1 CONJ* DET ADV VPART **N2**	→ **N1 CONJ* N2**	+	VPART **N2**
N1 CONJ* ADV ADJ **N2**	→ **N1 CONJ* N2**	+	ADJ **N2**
N1 CONJ* ADV VPART **N2**	→ **N1 CONJ* N2**	+	VPART **N2**
N1 CONJ* DET ADJ **N2**	→ **N1 CONJ* N2**	+	ADJ **N2**
N1 CONJ* DET VPART **N2**	→ **N1 CONJ* N2**	+	VPART **N2**
N1 CONJ* ADJ **N2**	→ **N1 CONJ* N2**	+	ADJ **N2**
N1 CONJ* VPART **N2**	→ **N1 CONJ* N2**	+	VPART **N2**
N1 CONJ* DET **N2**	→ **N1 CONJ* N2**		
N1 ADV ADJ **CONJ*** DET DET **N2**	→ **N1 CONJ* N2**	+	**N1** ADJ
N1 ADV VPART **CONJ*** DET DET **N2**	→ **N1 CONJ* N2**	+	**N1** VPART
N1 ADJ **CONJ*** DET DET **N2**	→ **N1 CONJ* N2**	+	**N1** ADJ
N1 VPART **CONJ*** DET DET **N2**	→ **N1 CONJ* N2**	+	**N1** VPART
N1 CONJ* DET DET **N2**	→ **N1 CONJ* N2**		
N1 CONJ* DET DET ADV ADJ **N2**	→ **N1 CONJ* N2**	+	ADJ **N2**
N1 CONJ* DET DET ADV VPART **N2**	→ **N1 CONJ* N2**	+	VPART **N2**
N1 CONJ* DET DET ADJ **N2**	→ **N1 CONJ* N2**	+	ADJ **N2**
N1 CONJ* DET DET VPART **N2**	→ **N1 CONJ* N2**	+	VPART **N2**
N1 CONJ* DET DET **N2**	→ **N1 CONJ* N2**		
N1 ADV ADV ADJ **CONJ*** DET **N2**	→ **N1 CONJ* N2**	+	**N1** ADJ
N1 ADV ADV VPART **CONJ*** DET **N2**	→ **N1 CONJ* N2**	+	**N1** VPART
N1 ADV ADV ADJ **CONJ* N2**	→ **N1 CONJ* N2**	+	**N1** ADJ
N1 ADV ADV VPART **CONJ* N2**	→ **N1 CONJ* N2**	+	**N1** VPART
N1 CONJ* DET ADV ADV ADJ **N2**	→ **N1 CONJ* N2**	+	ADJ **N2**
N1 CONJ* DET ADV ADV VPART **N2**	→ **N1 CONJ* N2**	+	VPART **N2**
N1 CONJ* ADV ADV ADJ **N2**	→ **N1 CONJ* N2**	+	ADJ **N2**
N1 CONJ* ADV ADV VPART **N2**	→ **N1 CONJ* N2**	+	VPART **N2**
N1 ADV ADV ADJ **CONJ*** DET DET **N2**	→ **N1 CONJ* N2**	+	**N1** ADJ
N1 ADV ADV VPART **CONJ*** DET DET **N2**	→ **N1 CONJ* N2**	+	**N1** VPART
N1 CONJ* DET DET ADV ADV ADJ **N2**	→ **N1 CONJ* N2**	+	ADJ **N2**
N1 CONJ* DET DET ADV ADV VPART **N2**	→ **N1 CONJ* N2**	+	VPART **N2**
N1 PREP DET ADV ADJ **N2**	→ **N1 PREP N2**	+	ADJ **N2**
N1 PREP DET ADV VPART **N2**	→ **N1 PREP N2**	+	VPART **N2**
N1 PREP DET ADJ **N 2**	→ **N1 PREP N2**	+	ADJ **N2**
N1 PREP DET VPART **N2**	→ **N1 PREP N2**	+	VPART **N2**
N1 PREP ADJ **N2**	→ **N1 PREP N2**	+	ADJ **N2**
N1 PREP VPART **N2**	→ **N1 PREP N2**	+	VPART **N2**
N1 PREP DET **N2**	→ **N1 PREP N2**		
N1 PREP N2	→ **N1 PREP N2**		
N1 PREP DET1 DET2 ADV2 ADJ2 **N2**	→ **N1 PREP N2**	+	ADJ **N2**

N1 PREP DET1 DET ADV2 VPART2 **N2**	→ **N1 PREP N2**	+	VPART **N2**
N1 PREP DET DET ADJ **N2**	→ **N1 PREP N2**	+	ADJ **N2**
N1 PREP DET DET VPART **N2**	→ **N1 PREP N2**	+	VPART **N2**
N1 PREP DET DET **N2**	→ **N1 PREP N2**		
N1 PREP DET1 ADV ADV2 ADJ2 **N2**	→ **N1 PREP N2**	+	ADJ **N2**
N1 PREP DET1 ADV ADV2 VPART2 **N2**	→ **N1 PREP N2**	+	VPART **N2**
N1 PREP DET DET1 ADV ADV2 ADJ2 **N2**	→ **N1 PREP N2**	+	ADJ **N2**
N1 PREP DET DET1 ADV ADV2 VPART2 **N2**	→ **N1 PREP N2**	+	VPART **N2**
N ADV **ADJ**	→ **N ADJ**		
N ADV **VPART**	→ **N VPART**		
N ADJ	→ **N ADJ**		
N VPART	→ **N VPART**		
N ADV ADV **ADJ**	→ **N ADJ**		
N ADV ADV **VPART**	→ **N VPART**		
N ADV ADJ1 CONJ* ADV **ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N ADV VPART CONJ* ADV **ADJ**	→ **N ADJ**	+	**N** VPART
N ADJ1 CONJ* ADV **ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N VPART CONJ* ADV **ADJ**	→ **N ADJ**	+	**N** VPART
N ADV ADJ1 CONJ* **ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N ADV VPART CONJ* **ADJ**	→ **N ADJ**	+	**N** VPART
N ADJ1 CONJ* **ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N VPART CONJ* **ADJ**	→ **N ADJ**	+	**N** VPART
N ADV ADJ CONJ* ADV **VPART**	→ **N ADJ**	+	**N** VPART
N ADV VPART1 CONJ* ADV **VPART2**	→ **N VPART1**	+	**N** VPART2
N ADJ CONJ* ADV **VPART**	→ **N ADJ**	+	**N** VPART
N VPART1 CONJ* ADV **VPART2**	→ **N VPART1**	+	**N** VPART2
N ADV ADJ CONJ* **VPART**	→ **N ADJ**	+	**N** VPART
N ADV VPART1 CONJ* **VPART2**	→ **N VPART1**	+	**N** VPART2
N ADJ CONJ* **VPART**	→ **N ADJ**	+	**N** VPART
N VPART1 CONJ* **VPART2**	→ **N VPART1**	+	**N** VPART2
N ADV ADV ADJ1 CONJ* ADV **ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N ADV ADV VPART CONJ* ADV **ADJ**	→ **N ADJ**	+	**N** VPART
N ADV ADV ADJ1 CONJ* **ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N ADV ADV VPART CONJ* **ADJ**	→ **N ADJ**	+	**N** VPART
N ADV ADV ADJ CONJ* ADV **VPART**	→ **N ADJ**	+	**N** VPART
N ADV ADV VPART1 CONJ* ADV **VPART2**	→ **N VPART1**	+	**N** VPART2
N ADV ADV ADJ CONJ* **VPART**	→ **N ADJ**	+	**N** VPART
N ADV ADV VPART1 CONJ* **VPART2**	→ **N VPART1**	+	**N** VPART2
N ADV ADJ1 CONJ* ADV ADV **ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N ADV VPART CONJ* ADV ADV **ADJ**	→ **N ADJ**	+	**N** VPART
N ADJ1 CONJ* ADV ADV **ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N VPART CONJ* ADV ADV **ADJ**	→ **N ADJ**	+	**N** VPART
N ADV ADJ CONJ* ADV ADV **VPART**	→ **N ADJ**	+	**N** VPART
N ADV VPART1 CONJ* ADV ADV **VPART2**	→ **N VPART1**	+	**N** VPART2
N ADJ CONJ* ADV ADV **VPART**	→ **N ADJ**	+	**N** VPART
N VPART1 CONJ* ADV ADV **VPART2**	→ **N VPART1**	+	**N** VPART2
N ADV ADV ADJ1 CONJ* ADV ADV **ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N ADV ADV VPART CONJ* ADV ADV **ADJ**	→ **N ADJ**	+	**N** VPART

N ADV ADV ADJ CONJ* ADV ADV **VPART**	→ **N ADJ**	+	**N** VPART
N ADV ADV VPART1 CONJ* ADV ADV **VPART2**	→ **N VPART1**	+	**N** VPART2
ADJ2 CONJ* ADV ADJ1 **N**	→ **ADJ1 N**	+	ADJ2 **N**
ADJ CONJ* ADV VPART **N**	→ **ADJ N**	+	VPART **N**
ADJ2 CONJ* ADJ1 **N**	→ **ADJ1 N**	+	ADJ2 **N**
ADJ CONJ* VPART **N**	→ **ADJ N**	+	VPART **N**
ADJ N	→ **ADJ N**		
VPART CONJ* ADV ADJ **N**	→ **VPART N**	+	ADJ **N**
VPART2 CONJ* ADV VPART1 **N**	→ **VPART1 N**	+	VPART2 **N**
VPART CONJ* ADJ **N**	→ **VPART N**	+	ADJ **N**
VPART2 CONJ* VPART1 **N**	→ **VPART1 N**	+	VPART2 **N**
VPART N	→ **VPART N**		
ADJ2 CONJ* ADV ADV ADJ1 **N**	→ **ADJ1 N**	+	ADJ2 **N**
ADJ CONJ* ADV ADV VPART **N**	→ **ADJ N**	+	VPART **N**
VPART CONJ* ADV ADV ADJ **N**	→ **ADJ N**	+	VPART **N**
VPART2 CONJ* ADV ADV VPART1 **N**	→ **VPART1 N**	+	VPART2 **N**
N ADJ1 ADJ2	→ **N ADJ1**	+	**N** ADJ2
N ADV **ADJ1 ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N ADJ1 ADV **ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N ADV **ADJ1** ADV **ADJ2**	→ **N ADJ1**	+	**N** ADJ2
N ADJ VPART	→ **N ADJ**	+	**N** VPART
N ADV **ADJ VPART**	→ **N ADJ**	+	**N** VPART
N ADJ ADV **VPART**	→ **N ADJ**	+	**N** VPART
N ADV **ADJ** ADV **VPART**	→ **N ADJ**	+	**N** VPART

II. Para un nombre fijado (*lema0*)

[-A la derecha:]

***lema0*-N** ADV ADJ**,** DET **N** → ***lema0*-N** ADJ + ***lema0*-N , N**
***lema0*-N** ADV VPART**,** DET **N** → ***lema0*-N** VPART + ***lema0*-N , N**
***lema0*-N** ADJ **,** DET **N** → ***lema0*-N** ADJ + ***lema0*-N , N**
***lema0*-N** VPART **,** DET **N** → ***lema0*-N** VPART + ***lema0*-N , N**
***lema0*-N , N** DET **N** → ***lema0*-N , N**
***lema0*-N** ADV ADJ **, N** → ***lema0*-N** ADJ + ***lema0*-N , N**
***lema0*-N** ADV VPART **, N** → ***lema0*-N** VPART + ***lema0*-N , N**
***lema0*-N** ADJ **, N** → ***lema0*-N** ADJ + ***lema0*-N , N**
***lema0*-N** VPART **, N** → ***lema0*-N** VPART + ***lema0*-N , N**
***lema0*-N , N** → ***lema0*-N , N**

***lema0*-N ,** DET ADV ADJ **N** → ***lema0*-N , N**
***lema0*-N ,** DET ADV VPART **N** → ***lema0*-N , N**
***lema0*-N ,** ADV ADJ **N** → ***lema0*-N , N**
***lema0*-N ,** ADV VPART **N** → ***lema0*-N , N**
***lema0*-N ,** DET ADJ **N** → ***lema0*-N , N**
***lema0*-N ,** DET PART **N** → ***lema0*-N , N**
***lema0*-N ,** ADJ **N** → ***lema0*-N , N**
***lema0*-N ,** PART **N** → ***lema0*-N , N**
***lema0*-N ,** DET **N** → ***lema0*-N , N**

***lema0*-N** ADV ADV ADJ**,** DET **N** → ***lema0*-N** ADJ + ***lema0*-N , N**
***lema0*-N** ADV ADV VPART**,** DET **N** → ***lema0*-N** VPART + ***lema0*-N , N**
***lema0*-N** ADV ADV ADJ **, N** → ***lema0*-N** ADJ + ***lema0*-N , N**
***lema0*-N** ADV ADV VPART **, N** → ***lema0*-N** VPART + ***lema0*-N , N**

***lema0*-N ,** DET ADV ADV ADJ **N** → ***lema0*-N , N**
***lema0*-N,** DET ADV ADV VPART **N** → ***lema0*-N , N**
***lema0*-N ,** ADV ADV ADJ **N** → ***lema0*-N , N**
***lema0*-N ,** ADV ADV VPART **N** → ***lema0*-N , N**

***lema0*-N** ADV ADJ**,** DET DET **N** → ***lema0*-N** ADJ + ***lema0*-N , N**
***lema0*-N** ADV VPART**,** DET DET **N** → ***lema0*-N** VPART + ***lema0*-N , N**
***lema0*-N** ADJ **,** DET DET **N** → ***lema0*-N** ADJ + ***lema0*-N , N**
***lema0*-N** VPART **,** DET DET **N** → ***lema0*-N** VPART + ***lema0*-N , N**
***lema0*-N ,** DET DET **N** → ***lema0*-N , N**

***lema0*-N ,** DET DET ADV ADJ **N** → ***lema0*-N , N**
***lema0*-N ,** DET DET ADV VPART **N** → ***lema0*-N , N**
***lema0*-N ,** DET DET ADJ **N** → ***lema0*-N , N**
***lema0*-N ,** DET DET PART **N** → ***lema0*-N , N**
***lema0*-N ,** DET DET **N** → ***lema0*-N , N**

***lema0*-N** ADV ADV ADJ**,** DET DET **N** → ***lema0*-N** ADJ + ***lema0*-N , N**
***lema0*-N** ADV ADV VPART**,** DET DET **N** → ***lema0*-N** VPART + ***lema0*-N , N**

***lema0*-N ,** DET DET ADV ADV ADJ **N** → ***lema0*-N , N**
***lema0*-N ,** DET DET ADV ADV VPART **N** → ***lema0*-N , N**

lema0*-N** ADV ADJ **CONJ DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** ADJ
lema0*-N** ADV VPART **CONJ DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** VPART
lema0*-N** ADJ **CONJ DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** ADJ
lema0*-N** VPART **CONJ DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** VPART
lema0*-N CONJ DET **N** → ***lema0*-N CONJ* N**
***lema0*-N** ADV ADJ **CONJ* N** → ***lema0*-N CONJ* N** + ***lema0*-N** ADJ

***lema0*-N** ADV VPART **CONJ* N** → ***lema0*-N CONJ* N** + ***lema0*-N** VPART
***lema0*-N** ADJ **CONJ* N** → ***lema0*-N CONJ* N** + ***lema0*-N** ADJ
***lema0*-N** VPART **CONJ* N** → ***lema0*-N CONJ* N** + ***lema0*-N** VPART
***lema0*-N CONJ* N** → ***lema0*-N CONJ* N**

lema0*-N CONJ DET ADV ADJ **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ DET ADV VPART **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ ADV ADJ **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ ADV VPART **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ DET ADJ **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ DET VPART **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ ADJ **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ VPART **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ DET **N** → ***lema0*-N CONJ* N**

lema0*-N** ADV ADJ **CONJ DET DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** ADJ
lema0*-N** ADV VPART **CONJ DET DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** VPART
lema0*-N** ADJ **CONJ DET DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** ADJ
lema0*-N** VPART **CONJ DET DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** VPART
lema0*-N CONJ DET DET **N** → ***lema0*-N CONJ* N**

lema0*-N CONJ DET DET ADV ADJ **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ DET DET ADV VPART **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ DET DET ADJ **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ DET DET VPART **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ DET DET **N** → ***lema0*-N CONJ* N**

lema0*-N** ADV ADV ADJ **CONJ DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** ADJ
lema0*-N** ADV ADV VPART **CONJ DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** VPART
***lema0*-N** ADV ADV ADJ **CONJ* N** → ***lema0*-N CONJ* N** + ***lema0*-N** ADJ
***lema0*-N** ADV ADV VPART **CONJ* N** → ***lema0*-N CONJ* N** + ***lema0*-N** VPART

lema0*-N CONJ DET ADV ADV ADJ **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ DET ADV ADV VPART **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ ADV ADV ADJ **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ ADV ADV VPART **N** → ***lema0*-N CONJ* N**

lema0*-N** ADV ADV ADJ **CONJ DET DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** ADJ
lema0*-N** ADV ADV VPART **CONJ DET DET **N** → ***lema0*-N CONJ* N** + ***lema0*-N** VPART

lema0*-N CONJ DET DET ADV ADV ADJ **N** → ***lema0*-N CONJ* N**
lema0*-N CONJ DET DET ADV ADV VPART **N** → ***lema0*-N CONJ* N**

***lema0*-N PREP** DET1 ADV ADJ **N** → ***lema0*-N PREP N**
***lema0*-N PREP** DET1 ADV VPART **N** → ***lema0*-N PREP N**
***lema0*-N PREP** DET ADJ **N** → ***lema0*-N PREP N**
***lema0*-N PREP** DET VPART **N** → ***lema0*-N PREP N**
***lema0*-N PREP** ADJ **N** → ***lema0*-N PREP N**
***lema0*-N PREP** VPART **N** → ***lema0*-N PREP N**
***lema0*-N PREP** DET **N** → ***lema0*-N PREP N**
***lema0*-N PREP N** → ***lema0*-N PREP N**

***lema0*-N PREP** DET DET ADV ADJ **N** → ***lema0*-N PREP N**
***lema0*-N PREP** DET DET ADV VPART **N** → ***lema0*-N PREP N**
***lema0*-N PREP** DET DET ADJ **N** → ***lema0*-N PREP N**
***lema0*-N PREP** DET DET VPART **N** → ***lema0*-N PREP N**
***lema0*-N PREP** DET DET **N** → ***lema0*-N PREP N**

***lema0*-N PREP** DET ADV ADV ADJ **N** → ***lema0*-N PREP N**
***lema0*-N PREP** DET ADV ADV VPART **N** → ***lema0*-N PREP N**

***lema0*-N PREP** DET DET ADV ADV ADJ **N** → ***lema0*-N PREP N**
***lema0*-N PREP** DET DET ADV ADV VPART **N** → ***lema0*-N PREP N**

***lema0*-N** ADV **ADJ** → ***lema0*-N ADJ**
***lema0*-N** ADV **VPART** → ***lema0*-N VPART**
***lema0*-N ADJ** → ***lema0*-N ADJ**
***lema0*-N VPART** → ***lema0*-N VPART**

***lema0*-N** ADV ADV **ADJ** → ***lema0*-N ADJ**
***lema0*-N** ADV ADV **VPART** → ***lema0*-N VPART**

***lema0*-N** ADV ADJ1 CONJ* ADV **ADJ2** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2
***lema0*-N** ADV VPART CONJ* ADV **ADJ** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** ADJ1 CONJ* ADV **ADJ2** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2
***lema0*-N** VPART CONJ* ADV **ADJ** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** ADV ADJ1 CONJ* **ADJ2** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2
***lema0*-N** ADV VPART CONJ* **ADJ** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** ADJ1 CONJ* **ADJ2** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2
***lema0*-N** VPART CONJ* **ADJ** → ***lema0*-N ADJ** + ***lema0*-N** VPART

***lema0*-N** ADV ADJ CONJ* ADV **VPART** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** ADV VPART1 CONJ* ADV **VPART2** → ***lema0*-N VPART1** + ***lema0*-N** VPART2
***lema0*-N** ADJ CONJ* ADV **VPART** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** VPART1 CONJ* ADV **VPART2** → ***lema0*-N VPART1** + ***lema0*-N** VPART2
***lema0*-N** ADV ADJ CONJ* **VPART** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** ADV VPART1 CONJ* **VPART2** → ***lema0*-N VPART1** + ***lema0*-N** VPART2
***lema0*-N** ADJ CONJ* **VPART** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** VPART1 CONJ* **VPART2** → ***lema0*-N VPART1** + ***lema0*-N** VPART2

***lema0*-N** ADV ADV ADJ1 CONJ* ADV **ADJ2** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2
***lema0*-N** ADV ADV VPART CONJ* ADV **ADJ** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** ADV ADV ADJ1 CONJ* **ADJ2** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2
***lema0*-N** ADV ADV VPART CONJ* **ADJ** → ***lema0*-N ADJ** + ***lema0*-N** PART

***lema0*-N** ADV ADV ADJ CONJ* ADV **VPART** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** ADV ADV VPART1 CONJ* ADV **VPART2** → ***lema0*-N VPART1** + ***lema0*-N** VPART2
***lema0*-N** ADV ADV ADJ CONJ* **VPART** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** ADV ADV VPART1 CONJ* **VPART2** → ***lema0*-N VPART1** + ***lema0*-N** VPART2

***lema0*-N** ADV ADJ1 CONJ* ADV ADV **ADJ2** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2
***lema0*-N** ADV VPART CONJ* ADV ADV **ADJ** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** ADJ1 CONJ* ADV ADV **ADJ2** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2
***lema0*-N** VPART CONJ* ADV ADV1 **ADJ** → ***lema0*-N ADJ** + ***lema0*-N** VPART

***lema0*-N** ADV ADJ CONJ* ADV ADV **VPART** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** ADV VPART1 CONJ* ADV ADV **VPART2** → ***lema0*-N VPART1** + ***lema0*-N** VPART2
***lema0*-N** ADJ CONJ* ADV ADV **VPART** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** VPART1 CONJ* ADV ADV **VPART2** → ***lema0*-N VPART1** + ***lema0*-N** VPART2

***lema0*-N** ADV ADV ADJ1 CONJ* ADV ADV **ADJ2** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2
***lema0*-N** ADV ADV VPART CONJ* ADV ADV **ADJ** → ***lema0*-N ADJ** + ***lema0*-N** VPART

***lema0*-N** ADV ADV ADJ CONJ* ADV ADV **VPART** → ***lema0*-N ADJ** + ***lema0*-N** VPART
***lema0*-N** ADV ADV VPART1 CONJ* ADV ADV **VPART2** → ***lema0*-N VPART1** + ***lema0*-N** VPART2

***lema0*-N ADJ ADJ** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2
***lema0*-N** ADV **ADJ ADJ** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2
***lema0*-N ADJ** ADV **ADJ** → ***lema0*-N ADJ1** + ***lema0*-N** ADJ2

***lema0*-N ADJ VPART**	→ ***lema0*-N ADJ**	+	***lema0*-N** VPART
***lema0*-N** ADV **ADJ VPART**	→ ***lema0*-N ADJ**	+	***lema0*-N** VPART
***lema0*-N ADJ** ADV **VPART**	→ ***lema0*-N ADJ**	+	***lema0*-N** VPART

[-A la izquierda:]

N ADV ADJ**,** DET ***lema0*-N**	→ **N , *lema0*-N**		
N ADV VPART**,** DET ***lema0*-N**	→ **N , *lema0*-N**		
N ADJ **,** DET ***lema0*-N**	→ **N , *lema0*-N**		
N VPART **,** DET ***lema0*-N**	→ **N , *lema0*-N**		
N , DET ***lema0*-N**	→ **N , *lema0*-N**		
N ADV ADJ **,** ***lema0*-N**	→ **N , *lema0*-N**		
N ADV VPART **,** ***lema0*-N**	→ **N , *lema0*-N**		
N ADJ **,** ***lema0*-N**	→ **N , *lema0*-N**		
N VPART **,** ***lema0*-N**	→ **N , *lema0*-N**		
N , ***lema0*-N**	→ **N , *lema0*-N**		
N , DET ADV ADJ ***lema0*-N**	→ **N , *lema0*-N**	+	ADJ ***lema0*-N**
N , DET ADV VPART ***lema0*-N**	→ **N , *lema0*-N**	+	VPART ***lema0*-N**
N , ADV ADJ ***lema0*-N**	→ **N , *lema0*-N**	+	ADJ ***lema0*-N**
N , ADV VPART ***lema0*-N**	→ **N , *lema0*-N**	+	VPART ***lema0*-N**
N , DET ADJ ***lema0*-N**	→ **N , *lema0*-N**	+	ADJ ***lema0*-N**
N , DET VPART ***lema0*-N**	→ **N , *lema0*-N**	+	VPART ***lema0*-N**
N , ADJ ***lema0*-N**	→ **N , *lema0*-N**	+	ADJ ***lema0*-N**
N , VPART ***lema0*-N**	→ **N , *lema0*-N**	+	VPART ***lema0*-N**
N , DET ***lema0*-N**	→ **N , *lema0*-N**		
N ADV ADV ADJ**,** DET ***lema0*-N**	→ **N , *lema0*-N**		
N ADV ADV VPART**,** DET ***lema0*-N**	→ **N , *lema0*-N**		
N ADV ADV ADJ **,** ***lema0*-N**	→ **N , *lema0*-N**		
N ADV ADV VPART **,** ***lema0*-N**	→ **N , *lema0*-N**		
N , DET ADV ADV ADJ ***lema0*-N**	→ **N , *lema0*-N**	+	ADJ ***lema0*-N**
N , DET ADV ADV VPART ***lema0*-N**	→ **N , *lema0*-N**	+	VPART ***lema0*-N**
N , ADV ADV ADJ ***lema0*-N**	→ **N , *lema0*-N**	+	ADJ ***lema0*-N**
N , ADV ADV VPART ***lema0*-N**	→ **N , *lema0*-N**	+	VPART ***lema0*-N**
N ADV ADJ**,** DET DET ***lema0*-N**	→ **N , *lema0*-N**		
N ADV VPART**,** DET DET ***lema0*-N**	→ **N , *lema0*-N**		
N ADJ **,** DET DET ***lema0*-N**	→ **N , *lema0*-N**		
N VPART **,** DET DET ***lema0*-N**	→ **N , *lema0*-N**		
N , DET DET ***lema0*-N**	→ **N , *lema0*-N**		
N , DET DET ADV ADJ ***lema0*-N**	→ **N , *lema0*-N**	+	ADJ ***lema0*-N**
N , DET DET ADV VPART ***lema0*-N**	→ **N , *lema0*-N**	+	VPART ***lema0*-N**
N , DET DET ADJ ***lema0*-N**	→ **N , *lema0*-N**	+	ADJ ***lema0*-N**
N , DET DET PART ***lema0*-N**	→ **N , *lema0*-N**	+	VPART ***lema0*-N**
N , DET DET ***lema0*-N**	→ **N , *lema0*-N**		
N ADV ADV ADJ**,** DET DET ***lema0*-N**	→ **N , *lema0*-N**		
N ADV ADV VPART**,** DET DET ***lema0*-N**	→ **N , *lema0*-N**		
N , DET DET ADV ADV ADJ ***lema0*-N**	→ **N , *lema0*-N**	+	ADJ ***lema0*-N**
N , DET DET ADV ADV VPART ***lema0*-N**	→ **N , *lema0*-N**	+	VPART ***lema0*-N**
N ADV ADJ **CONJ*** DET ***lema0*-N**	→ **N CONJ* *lema0*-N**		
N ADV VPART **CONJ*** DET ***lema0*-N**	→ **N CONJ* *lema0*-N**		
N ADJ **CONJ*** DET ***lema0*-N**	→ **N CONJ* *lema0*-N**		
N VPART **CONJ*** DET ***lema0*-N**	→ **N CONJ* *lema0*-N**		

N CONJ* DET ***lema0*-N** → **N CONJ*** ***lema0*-N**
N ADV ADJ **CONJ*** ***lema0*-N** → **N CONJ*** ***lema0*-N**
N ADV VPART **CONJ*** ***lema0*-N** → **N CONJ*** ***lema0*-N**
N ADJ **CONJ*** ***lema0*-N** → **N CONJ*** ***lema0*-N**
N VPART **CONJ*** ***lema0*-N** → **N CONJ*** ***lema0*-N**
N CONJ* ***lema0*-N** → **N CONJ*** ***lema0*-N**

N CONJ* DET ADV ADJ ***lema0*-N** → **N CONJ*** ***lema0*-N** + ADJ ***lema0*-N**
N CONJ* DET ADV VPART ***lema0*-N** → **N CONJ*** ***lema0*-N** + VPART ***lema0*-N**
N CONJ* ADV ADJ ***lema0*-N** → **N CONJ*** ***lema0*-N** + ADJ ***lema0*-N**
N CONJ* ADV VPART ***lema0*-N** → **N CONJ*** ***lema0*-N** + VPART ***lema0*-N**
N CONJ* DET ADJ ***lema0*-N** → **N CONJ*** ***lema0*-N** + ADJ ***lema0*-N**
N CONJ* DET VPART ***lema0*-N** → **N CONJ*** ***lema0*-N** + VPART ***lema0*-N**
N CONJ* ADJ ***lema0*-N** → **N CONJ*** ***lema0*-N** + ADJ ***lema0*-N**
N CONJ* VPART ***lema0*-N** → **N CONJ*** ***lema0*-N** + VPART ***lema0*-N**
N CONJ* DET ***lema0*-N** → **N CONJ*** ***lema0*-N**

N ADV ADJ **CONJ*** DET DET ***lema0*-N** → **N CONJ*** ***lema0*-N**
N ADV VPART **CONJ*** DET DET ***lema0*-N** → **N CONJ*** ***lema0*-N**
N ADJ **CONJ*** DET DET ***lema0*-N** → **N CONJ*** ***lema0*-N**
N VPART **CONJ*** DET DET ***lema0*-N** → **N CONJ*** ***lema0*-N**
N CONJ* DET DET ***lema0*-N** → **N CONJ*** ***lema0*-N**

N CONJ* DET DET ADV ADJ ***lema0*-N** → **N CONJ*** ***lema0*-N** + ADJ ***lema0*-N**
N CONJ* DET DET ADV VPART ***lema0*-N** → **N CONJ*** ***lema0*-N** + VPART ***lema0*-N**
N CONJ* DET DET ADJ ***lema0*-N** → **N CONJ*** ***lema0*-N** + ADJ ***lema0*-N**
N CONJ* DET DET VPART ***lema0*-N** → **N CONJ*** ***lema0*-N** + VPART ***lema0*-N**
N CONJ* DET DET ***lema0*-N**

N ADV ADV ADJ **CONJ*** DET ***lema0*-N** → **N CONJ*** ***lema0*-N**
N ADV ADV VPART **CONJ*** DET ***lema0*-N** → **N CONJ*** ***lema0*-N**
N ADV ADV ADJ **CONJ*** ***lema0*-N** → **N CONJ*** ***lema0*-N**
N ADV ADV VPART **CONJ*** ***lema0*-N** → **N CONJ*** ***lema0*-N**

N CONJ* DET ADV ADV ADJ ***lema0*-N** → **N CONJ*** ***lema0*-N** + ADJ ***lema0*-N**
N CONJ* DET ADV ADV VPART ***lema0*-N** → **N CONJ*** ***lema0*-N** + VPART ***lema0*-N**
N CONJ* ADV ADV ADJ ***lema0*-N** → **N CONJ*** ***lema0*-N** + ADJ ***lema0*-N**
N CONJ* ADV ADV VPART ***lema0*-N** → **N CONJ*** ***lema0*-N** + VPART ***lema0*-N**

N ADV ADV ADJ **CONJ*** DET DET ***lema0*-N** → **N CONJ*** ***lema0*-N**
N ADV ADV VPART **CONJ*** DET DET ***lema0*-N** → **N CONJ*** ***lema0*-N**

N CONJ* DET DET ADV ADV ADJ ***lema0*-N** → **N CONJ*** ***lema0*-N** + ADJ ***lema0*-N**
N CONJ* DET DET ADV ADV VPART ***lema0*-N** → **N CONJ*** ***lema0*-N** + VPART ***lema0*-N**

N PREP DET ADV ADJ ***lema0*-N** → **N PREP** ***lema0*-N** + ADJ ***lema0*-N**
N PREP DET ADV VPART ***lema0*-N** → **N PREP** ***lema0*-N** + VPART ***lema0*-N**
N PREP DET ADJ ***lema0*-N** → **N PREP** ***lema0*-N** + ADJ ***lema0*-N**
N PREP DET VPART ***lema0*-N** → **N PREP** ***lema0*-N** + VPART ***lema0*-N**
N PREP ADJ ***lema0*-N** → **N PREP** ***lema0*-N** + ADJ ***lema0*-N**
N PREP VPART ***lema0*-N** → **N PREP** ***lema0*-N** + VPART ***lema0*-N**
N PREP DET ***lema0*-N** → **N PREP** ***lema0*-N**
N PREP ***lema0*-N** → **N PREP** ***lema0*-N**

N1 PREP DET DET ADV ADJ ***lema0*-N** → **N PREP** ***lema0*-N** + ADJ ***lema0*-N**
N1 PREP DET DET ADV VPART ***lema0*-N** → **N PREP** ***lema0*-N** + VPART ***lema0*-N**
N1 PREP DET DET ADJ ***lema0*-N** → **N PREP** ***lema0*-N** + ADJ ***lema0*-N**
N1 PREP DET DET VPART ***lema0*-N** → **N PREP** ***lema0*-N** + VPART ***lema0*-N**
N1 PREP DET DET ***lema0*-N** → **N PREP** ***lema0*-N**

N1 PREP DET ADV ADV ADJ ***lema0*-N** → **N PREP *lema0*-N** + ADJ ***lema0*-N**
N1 PREP DET ADV ADV VPART ***lema0*-N** → **N PREP *lema0*-N** + VPART ***lema0*-N**

N1 PREP DET DET ADV ADV ADJ ***lema0*-N** → **N PREP *lema0*-N** + ADJ ***lema0*-N**
N1 PREP DET DET ADV ADV VPART ***lema0*-N** → **N PREP *lema0*-N** + VPART ***lema0*-N**

ADJ2 CONJ* ADV ADJ1 ***lema0*-N** → **ADJ1 *lema0*-N** + ADJ2 ***lema0*-N**
ADJ CONJ* ADV VPART ***lema0*-N** → **ADJ *lema0*-N** + VPART ***lema0*-N**
ADJ2 CONJ* ADJ1 ***lema0*-N** → **ADJ1 *lema0*-N** + ADJ2 ***lema0*-N**
ADJ CONJ* VPART ***lema0*-N** → **ADJ *lema0*-N** + VPART ***lema0*-N**
ADJ *lema0*-N → **ADJ *lema0*-N**

VPART CONJ* ADV ADJ ***lema0*-N** → **VPART *lema0*-N** + ADJ ***lema0*-N**
VPART2 CONJ* ADV VPART1 ***lema0*-N** → **VPART1 *lema0*-N** + VPART2 ***lema0*-N**
VPART CONJ* ADJ ***lema0*-N** → **VPART *lema0*-N** + ADJ ***lema0*-N**
VPART2 CONJ* VPART1 ***lema0*-N** → **VPART1 *lema0*-N** + VPART2 ***lema0*-N**
VPART *lema0*-N → **VPART *lema0*-N**

ADJ1 CONJ* ADV ADV ADJ2 ***lema0*-N** → **ADJ1 *lema0*-N** + ADJ2 ***lema0*-N**
ADJ CONJ* ADV ADV VPART ***lema0*-N** → **ADJ *lema0*-N** + VPART ***lema0*-N**

VPART CONJ* ADV ADV ADJ ***lema0*-N** → **ADJ *lema0*-N** + VPART ***lema0*-N**
VPART2 CONJ* ADV ADV VPART1 ***lema0*-N** → **VPART1 *lema0*-N** + VPART2 ***lema0*-N**

Anexo 3. Discriminadores de Sentido

ARTE

Sentido 1

aguatinta
cerámica
creación_artística
modelado
moldura
producción_artística
recreación
serigrafía

Sentido 2

Estatua_de_la_Libertad
abstracción
acompañamiento
acuarela
adagio
adaptación
aire
alto_relieve
altorrelieve
anaglifo
antífona
aria
arietta
armonización
armonía
arte_abstracto
arte_comercial
arte_popular
arte_étnico
artes_plásticas
baile
bajo
bajo-relieve
bajo_relieve
bajorrelieve
balada
ballet
barcarola
bellas_artes
blues
boceto
bodegón
bolero
bordado
bosquejo
bronce
bugui
bugui-bugui
bulería
busto
cadencia
camafeo
cancioncilla
canción
canción_de_amor
canción_de_borracho
canción_religiosa
canción_tradicional
cantata
cantinela
canto
canto_fúnebre
canto_gregoriano
canto_litúrgico
carboncillo
carbón
cencerrada
chotis
clásico
comedia_musical
composición
concierto
conga
contrapunto
coral
coreografía
cosido
costura
croquis
cuadrilla
cuadro
cuarteto
cántico
danza
delineación
desnudo
diagrama
dibujo_esquemático
dies_irae
diseño
divertimento
doxología
drama_musical
díptico
dúo
elevación
endecha
esbocito
esbozo
escena
esfinge
espiritual_negro
estampa
estatua
estribillo
estudio
expresionismo
expresionismo_abstracto
fantasía
fauvismo
figura
flamenco
flor_artificial
folk
fotograbado
fotolitografía
frase
frase_musical
fuga
gavota
giga
glifo
glisando
glíptica
gore
gorigori
gouache
grabado_en_acero
grabado_en_madera
grafismo
grafismo_por_ordenador
gráficos
guache
género
himno
himno_nacional
homofonía
horóscopo
icono
impresionismo
impromptu
innovación
instrumentación
intermezzo
invención
invento
jota
joya
kitsch
la_Marsellesa
la_internacional
labor
largo
leitmotiv
lied
lienzo
madrigal
magnificat
makemono
matriz
mazurca
medio_relieve
melodía
mezcla
mezzotinto
minué
misa
modernismo
modulación
molde
monodia

mosaico
motivo
movimiento
mural
musical
muñeira
mármol
móvil
música
música_clásica
música_country
música_de_cámara
música_flamenca
música_folk
música_pop
música_popular
música_religiosa
nana
naturaleza_muerta
naturalismo
nocturno
nueva_versión
obertura
obra
obra_de_arte
obra_de_talla
octava
opereta
oratorio
original
orquestación
paisaje
paisaje_marítimo
paisajismo
pasaje
pasodoble
pastiche
pastoral
pavana
perspectiva
pieza
pieza_musical
pintarrajo
pintura_abstracta
pintura_monocroma
pintura_mural
pintura_paisajística
plano
planta
poema_sinfónico
polca
polifonía
politonalismo
pop
popularismo
popurrí
posmodernismo
preludio
primitivismo
producción
producto
quinteto
realismo
realismo_mágico
recitado
recitativo
reggae
reharmonización
relieve
remake
representación
retrato
rondó
rumba
réquiem
saloma
samba
sardana
seguidilla
septeto
serenata
sevillana
sexteto
silueta
sinfonía
sintetismo
sintonía
sobreimpresión
soleá
solo
sonata
sonata_para_piano
suite
superrealismo
surrealismo
talla_en_madera
tallado_en_madera
tango
tarantela
te_deum
tema
tema_melódico
tema_musical
tesoro
tocata
tonada
tonadilla
trío
tríptico
vaciado
vals
vanguardismo
variación
villancico
visión
vista
voz
xilografía
zarabanda
ópera
ópera_bufa
ópera_cómica

Sentido 3

alfabetismo
arte_teatral
artesanía
equitación
escenotecnia
esgrima
marinería
maña
náutica
puntería
teatralidad

Sentido 4

cetrería
disección
juglaría
relojería
taxidermia
ventriloquia

Sentido 5

ardid
argucia
artería
astucia
camelo
dilación
disimulación
doble_juego
doblez
duplicidad
embaucamiento
engañifa
engaño
estafa
estratagema
fachada
farfolla
farol
fingimiento
fraude
fullería
ilusión
imitación
impostura
mistificación
obscurantismo
pastrana
patraña
picaresca
simulación
sofisma
timo
trampa
treta

AUTORIDAD

Sentido 1

-

Sentido 2

E.E.U.U
Estados_Unidos
Gobierno_de_los_Estados_Unidos
Imperio_Romano
Washington
administración
ayuntamiento
capitolio
consistorio
cuerpo_legislativo
departamento
departamento_gubernamental
división
estado
estado_del_bienestar
gobierno
gobierno_en_el_exilio
gobierno_estatal
gobierno_federal

gobierno_local
imperio
judicatura
municipalidad
palacio
papado
pleno_del_ayuntamiento
pontificado
régimen
sección

Sentido 3

agente_agrónomo
agrónomo
analista
anatomista
anticuario
apreciador
arquero
as
asesor
asesor_agrónomo
asesor_de_inversiones
asesor_de_moda
catador
cazatalentos
cinturón_negro
comentarista
conocedor
consejero
consultor
consultor_en_dirección_de_empresas
cosmetólogo
crack
crítico
crítico_musical
crítico_teatral
curador
ducho
entendido
especialista
estafador
esteta
esteticista
estrella
evaluador
examinador
flechero
genealogista
genio
geógrafo
hacha
horticultor
informático
jurado
jurisconsulto
jurisperito
jurista
lapidario
liturgista
lógico
maestra
maestro
manitas
miembro_del_jurado
mitologista
mitólogo
musicógrafo
musicólogo
ojeador
padre_de_la_Iglesia
parlamentario
perito
pistolero
portavoz_del_jurado
presidenta_del_jurado
presidente_del_jurado
probador
profesional
pronosticador
reclutador
sanador
sargento_de_reclutamiento
talento
tasador
terapeuta
técnico
valuador
veterano
virutoso
árbitro

Sentido 4

alcalde
alcaldesa
autoridad_civil
bajá
burgomaestre
concejal
edil
ex-alcalde
gobernador_general
munícipe
pachá
procónsul
sátrapa
teniente_de_alcalde
vicealcalde
vicegobernador
virreina
virrey

Sentido 5

carta_blanca
control
dominación
dominio
rienda
señoría

Sentido 6

Abraham_Lincoln
Adolfo_Suárez
Ardanza
Aznar
Bacon
Bella_Durmiente
Berkeley
Bismarck
Calvo_Sotelo
Charles_De_Gaulle
Chirac
Churchill
Cleopatra
Companys
Cromwell
César
Dalai_Lama
De_Gaulle
Demóstenes
Donne
Felipe_González
Francesc_Macià
Francis_Bacon
Franklin_Delano_Roosevelt
François_Mitterrand
Gandhi
Garaikoetxea
George_Berkeley
George_Washington
Georges_Pompidou
Giulio_Andreotti
González
Gran_Lama
Gregorio_I
Gregorio_XIII
Isabel_I
Isabel_II
Jacques_Chirac
Jawaharlal_Nehru
John_Donne
John_Major
John_Wesley
Jordi_Pujol
Josep_Tarradellas
José_Antonio_Ardanza
José_María_Aznar
Julio_César
Karlos_Garaikoetxea
King
Lenin
Leopoldo_Calvo_Sotelo
Lluís_Companys
Lucius_Annaeus_Seneca
Macià
Mahatma_Gandhi
Mao_Tse-tung
Mao_Zedong
Maquiavelo
Margaret_Thatcher
Martin_Luther_King
Massimo_D'Alema
Mitterrand
Míster_PESC
Nehru
Oliver_Cromwell
Pompidou
Príncipe_de_Gales
Pujol
Richelieu
Romano_Prodi
Stalin
Suárez
Séneca
Tarradellas
Theodore_Roosevelt
Tony_Blair
Ulysses_Grant
Victoria
Wesley
Winston_Churchill
abanderado
adherente
administrador
adulador

aficionada
agresor
amaestrador
amiga
animador
apasionada
apasionado
aprista
arcediano
archidiácono
archiduque
archiduquesa
arcipreste
aristócrata
arzobispo
azafato
bacante
baronesa
baronet
barón
bolchevique
brujo
caballero
caballero_andante
caballero_templario
cabeza
cacique
camarera
camarlengo
canciller
candidato
canónigo
capataz
capataza
capellán
capitoste
caporal
cardenal
caudillo
censor
centroizquierdista
chamán
cheer-leader
cheer_leader
cheerleader
cicerone
clérigo
cobista
comunista
conde
condesa
confesor
congresista
convergente
coordinador
cura
decano
dechado
delfín
demócrata
devota
devoto
deán
diocesano
dios
diputado
director
director_de_empresa
director_de_escena
director_de_escuela
director_de_investigación
director_ejecutivo
directora
dirigente_mayoritario
dirigente_minoritario
discípulo
diácono
domador
druida
duce
duque
duquesa
eclesiástico
ejemplo
empresaria
empresario
encargado
entrenador
entrenador_de_baloncesto
entrenador_de_béisbol
entrenador_de_fútbol
entrenador_de_hockey
entrenador_de_tenis
entusiasta
estadista
estafetera
euskadicoesquerro
evangelista
evangelizador
ex_presidente
fabianista
fanática
fanático
favorito
federalista
gerente
gran_duque
gran_duquesa
grande
hechicero
hombre_de_estado
hostelero
hotelero
imán
infanta
inferior
iniciador
instigador
internacionalista
izquierdista
jefa
jefe
jefe_de_correos
jefe_de_departamento
jefe_de_estación
jefe_de_estado
jefe_de_policía
jefe_de_seguridad
jerarca
laboralista
lama
legislador
lehendakari
lendakari
lord
líder_espiritual
líder_mayoritario
líder_minoritario
mago
maharajá
manager
mandado
mandatario
marajá
mariscal
marquesa
marqués
marxista
mayoral
miembro_del_partido
miladi
milor
ministro
ministro_de_Agricultura_y_Pesca
ministro_de_Asuntos_Exteriores
ministro_de_Defensa
ministro_de_Economía_y_Hacienda
ministro_de_Educación_y_Cultura
ministro_de_Fomento
ministro_de_Industria_y_Energía
ministro_de_Medio_Ambiente
ministro_de_Sanidad
ministro_de_agricultura
ministro_de_asuntos_exteriores
ministro_de_comercio
ministro_de_cultura
ministro_de_defensa
ministro_de_economía
ministro_de_educación
ministro_de_energía
ministro_de_hacienda
ministro_de_industria
ministro_de_interior
ministro_de_justicia
ministro_de_medio_ambiente
ministro_de_obras_públicas
ministro_de_sanidad
ministro_de_trabajo
ministro_de_transporte
misionero
mitrado
modelo
moderador
monarca
monseñor
mosén
míster
negrero
neocomunista
newtoniano
noble
noble_vitalicio
obispo
obispo_sufragáneo
ordenando
organizador
padre
papa
parásito
pastor
patriarca

patricio
patrona
patrono
patrón
pelota
pelotillero
penitenciario
peronista
político
pontífice
popular
portahachón
prebendado
preboste
predicador
prefecto
prelado
premier
preparador
presbítero
president
presidente
presidente_de_Francia
presidente_de_la_Generalitat_de_Catal unya
presidente_de_la_República_de_Italia
presidente_de_la_República_francesa
presidente_de_la_república
presidente_de_los_Estados_Unidos
presidente_del_Gobierno_español
presidente_del_gobierno
primado
primer_ministro
princesa
privado
priísta
promulgador
príncipe
príncipe_consorte
príncipe_heredero
párroco
rabino
rabí
racionero
rajá
rector
regente
regidor
reina
reina_consorte
reina_madre
reina_regente
republicano
responsable
reverendo
revisionista
rojo
sacerdote
sacerdotisa
sanguijuela
satélite
secretaria_general
secretario
secretario_de_estado
secretario_general
semidiós
senador
senador_del_estado
sir
sire
socialdemócrata
socialista
subdirector
subsecretario
superhombre
superintendente
superior
supervisor
súbdito
tecnócrata
templario
triunviro
vasallo
vestal
vicario
vicario_apostólico
vicecanciller
vicejefe
viceministro
vicepresidente
vicesecretario
vizconde
vizcondesa
zalamero
zarina

BANDA

Sentido 1

atadura
cadena
cadenas
cierre
cinturón_salvavidas
coyunda
cuerdas
freno
freno_aerodinámico
ligadura
limitación
mimbre
mordaza
pastilla_de_frenos
restricción
salvavidas
seguro
sujeción
tope
trabas
traílla
zapata

Sentido 2

abertura
acolchamiento
adorno
alzacuello
anacronismo
anilla
antigüedad
arcén
aro_de_nariz
artefacto
artículo
atadero
bloque
borde
burlete
camino
celo
cercado
chismes
cincha
cinta
cinta_adhesiva
cinta_aislante
cinta_para_máquina_de_escribir
claro
collar_de_perro
cono
construcción
cosa
cosas
creación
cubierta
cuello_alto
cuello_de_cisne
cuneta
decoración
droga
encarte
esfera
esparadrapo
estructura
excavación
flotador
fármaco
grapa
indicador
instalación
instalación_fija
instrumental
interlínea
juguete
llanta
llavero
lámina
línea
marca
margen
material_de_construcción
moldura
mueble_fijo
neumático
neumático_antideslizante
neumático_radial
neumático_sin_cámara
orillo
orla
ornamento
parteluz
pavimento
paño
peso
recinto
regleta
ribete
servicio
servilletero
señal
superficie
tejido

tela
tira
tirita
trapo
trastos
trayectoria
utillaje
vía

Sentido 3

-

Sentido 4

-

Sentido 5

banda_de_frecuencia
canal
canal_de_transmisión
conexión
contacto
coordinación
enlace

Sentido 6

apacheta
asociación
camorra
club
clube
consorcio
cooperativa
cámara_de_comercio
fraternidad
gang
gremio
gángster
gánster
institución
instituto
ladronera
legión
mafia_siciliana
mafioso
mano_negra
pistolero
pool
sindicato
sociedad

Sentido 7

-

Sentido 8

aglomeración
asonada
canalla
chusma
ejército
garulla
garullada
gentualla
gentuza
gentío
horda
masa
montón
muchedumbre
multitud
nube
patulea
populacho
tropa
turba
turbamulta

Sentido 9

línea_de_banda

Sentido 10

-

BOMBA

Sentido 1

artefacto_explosivo
bomba_de_demolición
bomba_de_dispersión
bomba_de_goma
bomba_de_humo
bomba_de_racimo
bomba_de_relojería
bomba_fétida
bomba_incendiaria
carga_de_profundidad
coche-bomba
cóctel_molotov
detonador
detonante
explosivo
fuegos_artificiales
fuegos_de_artificio
granada
granada_de_mano
mecanismo_explosivo
mina
obús
petardo
pirotecnia

Sentido 2

acoplamiento
aspersor
bigudí
bomba_al_vacío
bomba_aspirante
bomba_auxiliar
bomba_centrífuga
bomba_de_agua
bomba_de_aspiración
bomba_de_bicicleta
bomba_de_mano
bomba_estomacal
bomba_hidráulica
bomba_manual
calzador
cambio_de_marchas
carburador
cartuchera
chicho
columpio
compresor
consolador
devanadera
diafragma
dispositivo_mecánico
engranaje
engrasadora_a_presión
ensanchador
escape
escobilla
esparcidor
estator
fuelle
férula
gancho
garfio
horma
hélice
lengüeta
limpiaparabrisas
máquina
máquina_de_circulación_extracorpórea
máquina_de_montaje
máquina_de_votar
obturador
palanca_de_cambio
pantógrafo
parachoques
percusor
percutor
pistola_engrasadora
pistón
propulsor
rulo
sembradora
surtidor_de_gasolina
tablilla
trinquete
veleta
vibrador
émbolo

CANAL

Sentido 1

-

Sentido 2

-

Sentido 3

compuerta

Sentido 4

-

Sentido 5

abertura
agujero
alcachofa_de_la_ducha
barbacana
campana
canilla
cavidad
cañonera
chorrera
claro
escote
gárgola
hendidura
intersticio
ostíolo
pico
pitorro
pitón
porta
recámara
rejilla
resquicio
salida
surtidor
toma
tronera

ventana
ventanilla

Sentido 6

acueducto_cerebral
acueducto_de_Silvio
canal_hepático
canal_lacrimal
canal_pancreático
canal_pancreático_accesorio
canal_torácico
canal_vertebral
canalículo
conducto_biliar
conducto_biliar_común
conducto_eyaculador
conducto_seminal
cordón_umbilical
epidídimo
esófago
fauces
fístula
meato
oviducto
poro
seno_cavernoso
seno_coronario
seno_sigmoideo
seno_tentorial
seno_transverso
seno_venoso
shunt
tubo_digestivo
tubo_uterino
uretra
uréter
vagina
vaso_deferente
vaso_linfático
vía

Sentido 7

banda
banda_de_frecuencia
canal_de_transmisión
comunicación_electrónica
conexión
contacto
coordinación
difusión
divulgación
enlace
fibroóptica
impartición
media
medios_de_comunicación
modulación
multimedia
propagació
telecomunicación
transmisión
trasmisión

Sentido 8

arroyada
arroyo
bajo
barranca
barranco
cráter
cuenca
depresión
fondo
hoya
hoyo
lecho
riachuelo
socavón
surco
tierra_baja
torrentera
valle
zanja

Sentido 9

Dardanelos
Helesponto
abismo
abismo_oceánico
abra
agua
aguazal
arteria
bahía
bajío
balsa
cala
canal_Tajo-Segura
canal_de_María_Cristina
canal_del_Trasvase
cascada
catarata
charca
charco
corriente
curso
curso_del_agua
encharcada
ensenada
estanque
estrecho
estrecho_de_Corea
estrecho_de_Gibraltar
estuario
fosa
fosa_oceánica
golfo
lago
lagunajo
mar
océano
poza
pozanco
rada
rebalsa
seno
vado

Sentido 10

alcantarilla
alcantarillado
algalia
boca
boca_de_incendio
boca_de_riego
bureta
capilar
catéter
cañería
cañón
cañón_de_agua
cerbatana
chimenea
conducto_de_admisión
corredor
cuentagotas
cámara
cánula
difusor
embornal
entrada
fuste_de_chimenea
galería
gasoducto
humero
imbornal
lanzatorpedos
manguera
manguera_de_bomberos
narguile
oleoducto
pasaje
pasillo
paso
pipa
pipa_de_brezo
pipa_de_cerámica
pipa_de_la_paz
pipeta
prioridad
prioridad_de_paso
probeta
rampa_de_caída_de_carbón
sifón
tobogán
tobogán_de_agua
tubería
tubería_general_del_agua
tubería_general_del_gas
tuberías
tubo_acústico
tubo_de_admisión
tubo_de_estufa
tubo_de_lámpara
tubo_en_espiral

CIRCUITO

---

cassette
chip
circuitería
circuito_abierto
circuito_cerrado
circuito_de_ordenador
circuito_eléctrico
circuito_impreso
circuito_integrado
codificador
condensador
condensador_eléctrico
corto_circuito
cuadro_de_mandos
cátodo
célula
delco
derivación_eléctrica
detector
dispositivo_periférico
distribuidor
distribuidor_eléctrico
ecualizador
electrodo
electroimán
enchufe
equipo_auxiliar
escobilla
filtro
fototelégrafo
fusible
generador_electrostático
inductor
instalación_eléctrica
intercomunicador
interfaz_de_usuario
interficie
jack
lector
lector_de_CD
lente_de_electrones
mezclador
microchip
microcircuito
microplaqueta
modem
monitor
módem
módulo
obturador
oscilador
osciloscopio
panel_solar
pila
platina
plomo
polo
procesador
puente
radio-gramófono
radiocasete
radiocassete
reactor
rectificador
relé
replay
reproductor_de_CD
reproductor_de_casete
reproductor_de_vídeo
resistencia
resonador
sistema_de_audio
sistema_de_seguridad
supresor
tablero
tarjeta
tarjeta_de_cpu
tarjeta_de_pc
telefotógrafo
teléfono
terminal
transductor
transformador
transpondedor
trasformador
unidad_central_de_proceso
videoconsola
videojuego
videoplayer
videotelevisor
visionador
válvula_termoiónica
ánodo

Sentido 2

autódromo
campo
campo_de_golf
pista
pista_de_ceniza
terreno_de_juego
tramo

Sentido 3

-

Sentido 4

-

Sentido 5

-

Sentido 6

andadura
andar
andares
arrastramiento
caminar
carrera
corrida
círculo
gateado
locomoción
paso
paso_de_baile
vuelta

COLUMNA

Sentido 1

atlante
basa
basamento
capitel
cariátide
elemento_vertical
friso
fuste
jamba
pilote
plinto

Sentido 2

abrigo
aduja
albañilería
altar
andamio
arcada
arco
atracadero
balcón
brollador
carena
carrocería
casa_prefabricada
columnata
complejo
construcción
contrapeso
cornisamento
cornisamiento
crucero
defensa
divisor
edificio
equilibrio
espiral
establecimiento
estadio
estructura
estructura_defensiva
fuente
grada
hélice
impedimento
infraestructura
inmueble
instalaciones
letrero
manantial
monumento
monumento_conmemorativo
morón
obelisco
obras_públicas
observatorio
obstrucción
piso
planta
porche
protuberancia
puente
pórtico
quilla
refugio
retallo
rosca
saliente
separador
soportal
superestructura
surtidor
torre
transepto
viviendas
voluta

zona
área

Sentido 3

abrazadera
aguilón
anaquel
apoyo
arreos
asiento
balancín
balaustre
balaústre
base
cabestrillo
cojinete
colgadero
cortafuego
culata
dique
escalón
espigón
estante
estantería
estribera
estribo
faldón
fundamento
gablete
gancho
hastial
malecón
medianería
morillo
muralla
muro
pared_medianera
pata
peana
peldaño
percha
pie
pila
puntal
radio
respaldo
rompeolas
soporte
sujetalibros
tapia
varilla

Sentido 4

fila_india

Sentido 5

batería
calendario
espectro
matriz
orden
panoplia
renglón
tabla
tabla_estadística
tabla_periódica

Sentido 6

aljófar
conexión
distorsión
enlace
figura
forma
forma_angular
línea
plano
sólido
toroide
vuelo

Sentido 7

artículo
artículo_de_fondo
artículo_de_revista
consultorio
gaceta
separata
vespertino

Sentido 8

alineación
cola
flanco
formación_militar

CORAZÓN

Sentido 1

as
bastos
copas
diamante
espadas
figura
naipe
oros
picas
triunfo
trébol

Sentido 2

arteria_coronaria
buche
corazón_biauricular
estómago
intestino
músculo_cardíaco
panza
tripa
vesícula_biliar
válvula
válvula_del_corazón
órgano_excretor
órgano_excretorio
órgano_interno
órgano_respiratorio

Sentido 3

asunto
contenido_mental
convicción
creencia
cuestión
cultura
descreimiento
educación
enjundia
esencia
experiencia
falta_de_fe
fin
finalidad
formación
hipóstasis
idea
ignorancia
imagen_mental
instrucción
materia
meollo
meta
metaconocimiento
objetivo
pensamiento
propósito
quid
quintaesencia
saber_popular
sabiduría
sustancia
tema
teoría
tradición
universo
universo_del_discurso
área_de_conocimiento

Sentido 4

asaduras
lechecilla
lechecillas
lengua
menudillos
menudos
mollejas
sesos

Sentido 5

ancladero
casco_antiguo
centro_financiero
centrocampo
cuadrante
escena
espacio
esquina
fondeadero
mediocampo
ojo
refugio
región
retiro
rincón
santa_sede
sección
sede
tierra_de_nadie
zona_franca
área
área_geográfica
área_recreativa

Sentido 6

baricentro
centro_de_gravedad
centro_de_la_curvatura
diana
gaza
laza
ojo_de_la_tormenta
ombligo

Sentido 7

admirador
alma_gemela
amado
amante
amor
besucador
besucón
cariño
chica
cielo
compañera
compañera_sentimental
compañero
compañero_del_alma
favorito
novia
novio
pareja
pareja_sentimental
predilecto
preferido
prometida
prometido
querida
querido
sol

Sentido 8

diagrama_arbóreo
elipsoide
estrella
figura_bidimensional
figura_oblonga
figura_plana
paraboloide
polígono
sección_cónica
sector
semicírculo
árbol

Sentido 9

alma
intuición
presentimiento

Sentido 10

interior

CORONA

Sentido 1

diadema
joyas_de_la_corona
tiara

Sentido 2

arreglo_floral
cadena_de_flores
corona_de_laurel
guirnalda
lauréola
ramillete
ramita
ramo

Sentido 3

$
amuleto
cifra
divisa
emblema
estigma
gráfico
mancha
marca
número
sello
señal
simbolismo
símbolo
símbolo_gráfico
tipo
vara
variable

Sentido 4

aureola
auréola
resplandor

Sentido 5

cambio
centavo
chelín
cinco_peniques
cuarto_de_dólar
cuarto_de_penique
cuatro_peniques
céntimo
diez_peniques
doblón
dos_peniques
ducado
duro
euro
guinea
maría
media_corona
medio_dólar
medio_penique
moneda
nueve_peniques
níquel
ocho_peniques
penique
peseta
pieza
real
real_de_a_ocho
seis_peniques
tres_peniques
águila

GEMELO

Sentido 1

cuatrillizo
gemelo
hermanastro
hermano
mellizo
quintillizo
trillizo

Sentido 2

-

Sentido 3

-

GRACIA

Sentido 1

elegancia
estilo
gala
galanura
modernidad

Sentido 2

beneficencia

Sentido 3

-

Sentido 4

arranque
broma_de_mal_gusto
cachondeo
caricatura
chanza
chiste_negro
chiste_verde
chiste_visual
chocarrería
coña
equívoco
gag
golpe
humor_gráfico
humor_obsceno
imitación
ingenio
ingeniosidad
ironía
ocurrencia
rapidez
salida
sarcasmo
sátira

Sentido 5

acción
actividad
antagonismo
autonomia
cargo
circunstancias
condición
delegación
dependencia
desempleo
desocupación
desorden
destreza
destrucción
disponibilidad
empleo
enemistad
espíritu
estado
estado_de_gracia
estado_de_la_materia
estado_de_ánimo
estado_emocional
estado_físico
estado_natural
estado_salvaje
estado_temporal
estatismo
estatus
etapa

existencia
fase
final
grado
habilidad
hostilidad
humor
iluminación
imperfección
inacción
inactividad
inminencia
inmovilidad
integridad
isomería
libertad
liquidación
lugar
madurez
movimiento
muerte
naturaleza
nivel
ocupación
omnipotencia
omnisciencia
orden
paro
perfección
poder
posición
punto
rango
relación
representación
situación
status
tierra_virgen
totalidad
trabajo
tramo
trato
unidad
unión
utopía
ánimo

Sentido 6

cumplido
tomadura_de_pelo

Sentido 7

andares
desmaño
inflexibilidad
porte
rigidez
torpeza
torpor
tosquedad

GRANO

Sentido 1

cloroplasto
cromoplasto
gránulo
insignificancia
menudencia
miaja
microsoma
partícula
pellizco
pizca
plasto

Sentido 2

café
castaña_de_Indias
hueso
pepita
semilla
semilla_comestible
semillas_oleaginosas

Sentido 3

-

Sentido 4

cuarto_de_libra
kilotón
libra
megatón

Sentido 5

dracma
escrúpulo

Sentido 6

decagramo
decigramo
dg
g
gr
gramo
hectogramo
hg
kg
kilo
kilogramo
mg
microgramo
miligramo
quilate
quilo
quilogramo
tonelada_métrica
unidad_de_peso

Sentido 7

acné_adenoida
acné_adolescentium
acné_agminata
acné_alba
acné_artificial
acné_atrófica
acné_brómica
acné_caquéctica
acné_ciliar
acné_clórica
acné_coagminata
acné_congestiva
acné_córnea
acné_decalvans
acné_diseminada
acné_efébica
acné_elefantiásica
acné_eritematosa
acné_florida
acné_frontalis
acné_general
acné_granulosa
acné_halógena
acné_hipertrófica
acné_hordeolaris
acné_luposa
acné_miliaris
acné_necrótica
acné_pancreática
acné_picealis
acné_queloide
acné_queratosa
acné_rosácea
acné_sebácea
acné_sifilítica
acné_solaris
acné_telangiectodes
acné_varioliforme
acné_vulgar
barrillo
enfermedad_de_Quinquaud
granito
milio
picadura
pupa
seborrea

Sentido 8

absceso
acedía
acidez_de_estómago
ahíto
alforza
amenorrea
anaplasia
anemia
ansia
apnea
apostema
arcada
ardentía
ardor_de_estómago
ardores
arritmia
asco
atrofia
aura
basca
bulto
calambrazo
cardiomegalia
chiribita
cicatriz
congestión
descalabradura
descomposición
desvanecimiento
diarrea
dipsnea
dispepsia
dolor
dureza
efecto_secundario
empacho
entumecimiento
erupción
espasmo
estornudo
estreñimiento
exoftalmia

exoftalmos
fiebre
gangrena
habón
hematuria
hemorragia
hinchazón
hiperglucemia
hipertrofia
hipo
hipoglucemia
ictericia
indigestión
inflamación
lacra
mareo
mono
mononucleosis
náusea
palpitación
parálisis
perlesía
piel_de_gallina
pirexia
postema
pródromo
punzada
purulencia
roncha
sarpullido
soplo
síndrome
síndrome_de_abstinencia
síntoma
taquiarritmia
taquicardia
tiritera
tiritona
tos
urticaria
vahído
vértigo

HERMANO

Sentido 1

hermana_mayor
hermana_pequeña
hermanastra
hermanita

Sentido 2

comisionado
conciliar
ejecutivo
francmasón
masón
miembro
miembro_del_clan
miembro_del_club
miembro_del_comité
miembro_fundador

Sentido 3

antecesor
antepasado
ascendiente
cercano
consanguíneo
consorte
cuatrillizo
cónyuge
cónyugues
descendiente
familiar
gemelo
línea_materna
línea_paterna
matrimonio
mellizo
pareja
parienta
pariente_político
primo
primo_carnal
primo_hermano
primo_segundo
quintillizo
trillizo
vástago

Sentido 4

sobrino
tío

Sentido 5

acompañante
amiga
amigo
amigota
amigote
amigueta
amiguete
amiguito
camarada
colega
compa
compadre
compañera
compañero
compañero_de_colegio
compañero_de_habitación
compañero_de_juego
compañero_de_piso
compinche
confidente
cuate
manito
mano
ninchi
queli
socia
socio
tronco
álter_ego
íntimo

Sentido 6

-

LETRA

Sentido 1

caligrafía
cursiva
código
escarabajo
escritura
estenografía
garabato
garambaina
garrapato
impresión
letra_inglesa
mecanografía
notación
ortografía
redondilla
taquigrafía

Sentido 2

a
alfa
alfabeto_Braille
alfabeto_Morse
alfabeto_arábico
alfabeto_cirílico
alfabeto_fonético
alfabeto_griego
alfabeto_hebreo
alfabeto_romano
alfabeto_ruso
alfabeto_telegráfico
alfabeto_árabe
alógrafo
asterisco
b
be
beta
braille
c
caja_alta
caja_baja
carácter
carácter_ascii
ce
ce_hache
ceda
ceta
ch
che
chi
consonante
cu
d
de
delta
doble_erre
doble_uve
dígrafo
e
efe
ele
elle
eme
ene
equis
erre
erre_doble
ese
espacio
eta
eñe
f
fi
g
gamma

ge
grafía
h
hache
i
i_griega
i_latina
ideograma
inicial
iniciales
iota
j
jota
k
ka
kappa
l
lambda
letra_doble
letra_inicial
letra_mayúscula
letra_minúscula
ll
m
mayúscula
mi
minúscula
monograma
n
ni
o
obelisco
omega
p
pe
pi
pictograma
polifonía
psi
q
r
radical
ro
rr
runa
s
sigma
subíndice
superíndice
símbolo_fonético
símbolo_matemático
t
tanto_por_ciento
tau
te
tetragrama
theta
tipo
u
uve
uve_doble
v
ve
ve_doble
vocal
w
x
xi
y
ye
z
zeda
zeta
épsilon
ípsilon
ñ
ómicron

Sentido 3

aria
balada
barcarola
borrador
cancioncilla
canción_de_amor
canción_de_borracho
canción_tradicional
cantinela
canto_fúnebre
carta
endecha
entrega
estancia
estrofa
gorigori
himno_nacional
lied
nana
réquiem
saloma
serenata
texto
tonadilla

Sentido 4

abonamiento
abono
adelanto
anticipo
arras
bacalada
cohecho
crédito
derechos_de_autor
derechos_de_patente
desembolso
devolución
entrada
envío
finiquito
giro
gratificación
incentivo
liquidación
miseria
paga_y_señal
pago
penalización
pensión
plazo
prima
pronto_pago
recompensa
reembolso
regalías
remisión
retribución
royalty
saldo
señal
soborno
subscripción
subsidio

Sentido 5

autorización_bancaria
cheque
cheque_certificado
cheque_en_blanco
cheque_personal
constancia
cédula_de_dividendo
descubierto
documento
letra_de_cambio
libramiento
libranza
orden_de_pago
sobregiro
talón

MASA

Sentido 1

absorción
elasticidad
fugacidad
imperceptibilidad
inaudibilidad
inducción
inelasticidad
inercia
invisibilidad
luminosidad
masa_atómica
masa_atómica_relativa
masa_en_reposo
masa_gravitatoria
masa_inercial
masa_molecular
perceptibilidad
peso
peso_atómico
peso_equivalente
peso_molecular
propiedad_física
reflectividad
reflexión
sensibilidad
solubilidad
temperatura
volumen

Sentido 2

amplitud
cuantía
dimensión
extensión
magnitud
medida
mole
multiplicidad
proporción
tamaño

Sentido 3

lengua_chádica
lengua_chádica_occidental
lengua_chádica_oriental

Sentido 4

-

Sentido 5

amasijo
harina
mezcla
preparado
rebozado
relleno

Sentido 6

acumulación
agrupación
ciudadanía
claca
colección
colectivo
conjunto
edición
especie_humana
etnia
fiel
gente
grupo
grupo_biológico
grupo_social
género_humano
hatajo
hombre
humanidad
laicado
lectorado
lectores
mundo
personas
población
pueblo
raza
raza_humana
reino
santoral
seguidores
sistema
subgrupo
telespectador
televidente

Sentido 7

cromosoma
partícula

Sentido 8

acompañamiento
afluencia
aglomeración
asamblea
asistencia
asociación
asonada
atasco
banda
basca
campamento
canalla
caravana
casa
chusma
clase
colectividad
comitiva
comuna
comunidad
concurrencia
contingente
convocatoria
corporación
cortejo
cuadrilla
cuarteto
cuatro_gatos
curso
embotellamiento
encuentro_multitudinario
familia
feria
garulla
garullada
gentualla
gentuza
grupito
gusanera
junta
municipio
nube
octeto
panda
pandilla
pareja
patulea
populacho
promoción
público
quinteto
quórum
reparto
retención
reunión
reunión_social
revista
septeto
sexteto
séquito
tapón
trinca
trinidad
tropa
tríada
trío
turbamulta

Sentido 9

avalancha
barbaridad
batallón
bestialidad
burrada
cargamento
enormidad
fajo
infinidad
lote
límite
mar
mogollón
montones
mucho
máximo
océano
pequeña_fortuna
pila
resma
tanda

Sentido 10

legión
manada

MINA

Sentido 1

grafito
plombagina

Sentido 2

artefacto_explosivo
bomba
detonador
detonante
explosivo
fuegos_artificiales
fuegos_de_artificio
mecanismo_explosivo
mina_magnética
petardo
pirotecnia
trampa_explosiva

Sentido 3

barreno
boca
cantera
carbonera
charca
despensa
entrada
estanque
excavación
explotación
mina_de_carbón
mina_de_cobre
mina_de_oro
mina_de_plata
mina_de_sulfuro
pozo
salina
taladro
zanja

NATURALEZA

Sentido 1

accesibilidad
acrimonia
acritud
adustez
afabilidad
agitación
agrado
agresividad
aislamiento
alteración
amabilidad
amistosidad
amor_propio
angurria

animalidad
animalismo
antipatía
apartamiento
aplomo
apocamiento
aspereza
ataraxia
atención
austeridad
autoestima
azoramiento
benevolencia
buena_voluntad
calma
calmosidad
camaradería
campechanería
carácter
carácter_furtivo
cicatería
compañerismo
complacencia
comportamiento
comunicatividad
condescendencia
conducta
confiabilidad
confianza
confusión
congenialidad
cordialidad
cortesía
criterio
delicadeza
desabrimiento
desasosiego
desatino
desconfianza
desgana
determinación
dinamismo
discernimiento
disciplina
disposición
dulzura
dureza
egocentrismo
egoísmo
emotividad
empuje
encogimiento
entusiasmo
espiritualidad
espíritu_de_equipo
extroversión
familiaridad
femineidad
feminidad
firmeza
formalidad
franqueza
frialdad
frivolidad
generosidad
gregarismo
hospitalidad
hosquedad
humildad
humor_cambiadizo
identidad
impaciencia
inaccesibilidad
inclemencia
indisciplina
indisposición
individualidad
indolencia
indulgencia
informalidad
inquietud
inseguridad
insensatez
insociabilidad
intimidad
intolerancia
intranquilidad
intransigencia
introversión
irreflexión
irresolución
jovialidad
juicio
limpieza
longanimidad
mal_humor
malicia
manera_de_ser
maneras
masculinidad
misantropía
modales
moral
nerviosidad
nerviosismo
no_comunicación
noluntad
orgullo
paciencia
permisividad
placidez
privacidad
prontitud
prudencia
pugnacidad
puritanismo
quejumbridad
racaneo
racanería
rasgo
recelo
receptividad
reciedumbre
reluctancia
reserva
resistencia
resolución
reticencia
retiro
rigurosidad
risibilidad
roncería
roña
roñería
roñosería
sabiduría
sangre
sangre_fría
sensatez
serenidad
seriedad
seso
severidad
simpatía
sobresalto
sociabilidad
soledad
solvencia
sosiego
suciedad
sumisión
tacañería
talante
temperamento
tolerancia
tranquilidad
turbación
índole

Sentido 2

adecuación
apariencia
aptitud
aridez
aspecto
atracción
aura
bondad
calaña
calidad
candor
cara
carácter_agradable
carácter_desagradable
certeza
claridad
complejidad
comprensibilidad
constructividad
contrasentido
conveniencia
corrección
cotidianidad
cualidad
cualidad_de_los_padres
cualidades
demérito
deslucimiento
desmerecimiento
desnaturalidad
desprestigio
destructividad
diferencia
dificultad
distinción
divinidad
domesticidad
elegancia
esterilidad
etnicidad
exactitud
excepcionalidad

expresividad
extrañeza
facilidad
fecundidad
finitud
generalidad
gobernabilidad
humanidad
humanitarismo
igualdad
ilegalidad
impiedad
impopularidad
importancia
impotencia
inadecuación
inalterabilidad
inaptitud
incapacidad
incerteza
incomprensibilidad
inconveniencia
incorrección
incredibilidad
inelegancia
infinidad
infinitud
infructuosidad
inhumanidad
inmaterial
inmaterialidad
inmoralidad
inocuidad
inoperancia
insatisfacción
irregularidad
legalidad
lógica
madera
materialidad
mesurabilidad
moralidad
morbosidad
movilidad
mundología
naturalidad
negativismo
nitidez
objetividad
obligatoriedad
opacidad
originalidad
particularidad
popularidad
positivismo
potencia
precisión
presencia
propiedad
rareza
regularidad
rigor
romanticismo
santidad
satisfacción
semejanza
sencillez
similitud
simpleza
simplicidad
sofisticación
soltura
solubilidad
utilidad
ventaja
virtud
virtudes
volatilidad

Sentido 3

-

Sentido 4

agente
agente_causal
alma
catalizador
causa
causa_última
destino
deus_ex_machina
fuerza
hado
humano
individuo
manipulador
mortal
motor
oculto
operador
operario
peligro
persona
primer_motor
principio_vital
ser_humano
ser_sobrenatural
sino
sobrenatural

Sentido 5

-

Sentido 6

acción
antagonismo
autonomia
cargo
circunstancias
condición
crudeza
delegación
dependencia
desempleo
desocupación
desorden
destreza
destrucción
empleo
enemistad
estado
estado_de_gracia
estado_de_la_materia
estado_de_ánimo
estado_emocional
estado_físico
estado_natural
estado_salvaje
estado_temporal
estatismo
estatus
etapa
existencia
fase
final
gracia
grosería
habilidad
humor
iluminación
imperfección
inacción
inminencia
integridad
isomería
libertad
liquidación
lugar
madurez
movimiento
muerte
nivel
ocupación
omnipotencia
omnisciencia
orden
paro
perfección
posición
primitivismo
punto
rango
relación
representación
situación
status
tierra_virgen
tosquedad
totalidad
trabajo
tramo
trato
unidad
unión
utopía

OPERACIÓN

Sentido 1

abastecimiento
abasto
activación
actividad
actividad_sensorial
actuación
adoración
agrupación
animación
apoyo
aprovisionamiento
asistencia
ayuda
busca
ceremonia
colocación

comportamiento
conato
conducta
continuación
control
costumbre
creación
deleitación
deleite
demanda
derroche
desarme
desarticulación
desmantelamiento
desmontaje
despilfarro
dilapidación
disimulo
diversión
educación
emplazamiento
empleo
enseñanza
escritura
formación
guarda
guardia
hábito
intento
interpretación
juego
juego_de_niños
liderato
liderazgo
mala_conducta
maldad
medición
mercado
murmullo
murmurio
ocultación
ocupación
organización
pan_comido
perturbación
placer
posición
precedencia
preparación
preparativo
presentación
proceso
protección
provisión
práctica
recreación
representación
rol
situación
solo
tentativa
turbación
ubicación
uso
utilización
variación

Sentido 2

acupuntura
alopatía
angioplastia
anticoagulación
aromaterapia
autoplastia
cauterio
cauterización
corrección
cura
cura_de_reposo
curativa
desinfección
desintoxicación
digitopuntura
ergoterapia
fisioterapia
galvanismo
hidroterapia
homeopatía
incisión
intervención
inyección
laparoscopia
laparotomía
legrado
leucotomía
lifting_facial
lobotomía
mamoplastia
medicación
naturopatía
osteopatía
psicoterapia
quimioterapia
quiromasaje
radioterapia
raspado
remedio
rinoplastia
ritidoplastia
sangría
shiatsu
sutura
terapia
terapia_ocupacional
termalismo
trasplante
tratamiento
tratamiento_autogénico
tratamiento_de_shock
tratamiento_digitálico
trepanación
vendaje
vivisección

Sentido 3

Operación_Tormenta_del_Desierto
acción
acción_militar
acometida
acometimiento
andanada
arremetida
asalto
ataque
ataque_aéreo
ataque_por_tierra
ataque_repentino
ataque_sorpresa
avanzada
batalla
bloqueo
bombardeo
bombardeo_en_picado
campaña
campaña_de_Petersburgo
campaña_naval
carga
cañoneo
censura
censura_civil
censura_de_las_fuerzas_armadas
censura_de_prisioneros_de_guerra
censura_militar
censura_nacional
censura_primaria
censura_secundaria
combate
contraataque
contrabombardeo
contraespionaje
contrafuego
contrainteligencia
contrasabotaje
contrasubversión
correría
cruzada
defensa
descarga
descarga_cerrada
descubierta
embestida
emboscada
escapada
espionaje
expedición
expedición_militar
exploración
fuego
fuego_antiaéreo
fuego_cruzado
fuego_de_artillería
fuego_de_mortero
fuego_directo
fuego_indirecto
fuego_preparado
fuego_supresivo
golpe
guerra
incursión
infiltración
inteligencia_de_seguridad
inteligencia_estratégica
inteligencia_táctica
invasión
irrupción
lucha
línea_de_fuego
maniobra
medida_defensiva
misión_de_combate
misión_militar
ofensiva
operación_clandestina

operación_de_inteligencia
operación_encubierta
operación_militar_anfibia
penetración
rebato
recogida_de_información
reconocimiento
reconocimiento_del_terreno
refuerzo
resistencia
ráfaga
salva
trabajo_clandestino
zalagarda

Sentido 4

negocio

Sentido 5

clasificación
función_de_control
interlínea
multiprocesamiento
operación_asincrónica
operación_auxiliar
operación_binaria
operación_booleana
operación_de_control
operación_de_impresión
operación_diádica
operación_en_paralelo
operación_lógica
operación_monádica
operación_múltiplex
operación_off-line
operación_secuencial
operación_simultánea
operación_sincrónica
procesamiento_de_datos
procesamiento_de_datos_autómatico
procesamiento_de_palabras
procesamiento_en_tiempo_real
procesamiento_prioritario
procesamiento_serial
teleproceso
teletratamiento

Sentido 6

atención
cuidado
deber
faena
faenas_de_la_casa
funcionalidad
funcionamiento
función
investigación
labor
marcha
operatividad
papeleo
prestación
proyecto
quehaceres_domésticos
servicio
tarea

ÓRGANO

Sentido 1

cañón
cornamusa
corneta
gaita
instrumento_de_viento
instrumento_de_viento_de_madera
kazoo
metal
ocarina
pedal
tecla
teclado
zampoña

Sentido 2

abdomen
abductor
aductor
aductor_largo
aductor_mayor
agalla
almeja
ambulacro
anca
aparato
articulación
bonete
branquia
buche
bucinador
bíceps
cadera
caja_craniana
calavera
canilla
caracol_óseo
cardias
centriolo
chirri
chocho
chueca
cilio
clítoris
cojón
colon
colon_ascendente
colon_descendente
colon_sigmoideo
colon_transversal
condriosoma
conducto_eyaculador
conejo
coracobraquial
corazón
corazón_biauricular
corpus_luteum
coño
cristalino
cráneo
cuarto_estómago
culata
culo
cuádriceps
cóclea
dedo
dedo_del_pie
deltoides
depresor
dorso
duodeno
dídimo
entrañas
esfínter
esfínter_de_la_pupila
esfínter_de_la_uretra
esfínter_de_la_vejiga
esfínter_del_ano
esfínter_del_conducto_de_la_bilis
esfínter_del_páncreas
esfínter_del_píloro
esfínter_externo_del_ano
esfínter_fisiológico
esfínter_interno_del_ano
espalda
espinilla
esplenio_de_la_cabeza
esplenio_del_cuello
estructura_anatómica
estría
estómago
facción
falo
fascia_temporal
fascia_toracolumbar
galea_aponeurótica
gastrocnemio
genitales
genitales_femeninos
genitales_masculinos
glotis
glándula
glándula_adrenal
glándula_aprocrina
glándula_bulbouretral
glándula_de_Bartolino
glándula_de_Brunner
glándula_ecrina
glándula_endocrina
glándula_exocrina
glándula_lagrimal
glándula_paratiroides
glándula_paratiroides_inferior_derecha
glándula_paratiroides_superior_derecha
glándula_pineal
glándula_pituitaria
glándula_pituitaria_anterior
glándula_pituitaria_posterior
glándula_pulmonar
glándula_salival
glándula_sebácea
glándula_sublingual
glándula_submandibular
glándula_submaxilar
glándula_sudorípara
glándula_suprarrenal
glándula_timo
glándula_tiroides
glándula_vestibular_mayor
glúteo
grupa
gándula_digestiva
gónada
hombro

huevo
hígado
ijada
intestino
intestino_delgado
intestino_grueso
laberinto
laringe
lengua
librillo
libro
lomo
lóbulo
madre
mama
mamella
matriz
miembro
mitocondria
muñón
músculo
músculo_abductor
músculo_abductor_del_dedo_meñique
músculo_abductor_del_dedo_meñique_del_pie
músculo_abductor_del_pulgar
músculo_abductor_del_pulgar_del_pie
músculo_aductor
músculo_aductor_del_muslo
músculo_aductor_del_pulgar_del_pie
músculo_aductor_largo
músculo_aductor_mayor
músculo_angular_del_omóplato
músculo_antagonista
músculo_articular
músculo_articular_de_la_rodilla
músculo_articular_del_cúbito
músculo_auricular_anterior
músculo_auricular_posterior
músculo_auricular_superior
músculo_axial
músculo_bucinador
músculo_bíceps
músculo_bíceps_femoral
músculo_canino
músculo_cigomático_mayor
músculo_cigomático_menor
músculo_complexo_mayor
músculo_coracobraquial
músculo_cuadrado_del_mentón
músculo_cutáneo_del_cuello
músculo_de_la_pared_torácica
músculo_del_cráneo
músculo_del_cuello
músculo_del_oído
músculo_del_oído_externo
músculo_del_oído_medio
músculo_del_sastre
músculo_deltoides
músculo_depresor
músculo_dorsal_ancho
músculo_dorsal_largo
músculo_elevador_común_del_ala_de_la_nariz_y_el_labio_superior
músculo_elevador_propio_del_labio_superior
músculo_epicráneo
músculo_espinal
músculo_espinoso_dorsal
músculo_esquelético
músculo_estapedio
músculo_esternocleidomastoideo
músculo_esternohioideo
músculo_extensor
músculo_facial
músculo_flexor
músculo_frontal
músculo_iliocostal_de_la_espalda
músculo_infraespinoso
músculo_intercostal
músculo_involuntario
músculo_masetero
músculo_mentoniano
músculo_nasal
músculo_oblicuo_externo
músculo_oblicuo_interno
músculo_occipital
músculo_omohioideo
músculo_orbicular_de_los_labios
músculo_orbicular_de_los_párpados
músculo_pectoral
músculo_pectoral_mayor
músculo_pectoral_menor
músculo_piramidal_de_la_nariz
músculo_redondo_mayor
músculo_redondo_menor
músculo_risorio
músculo_sartorio
músculo_serrato
músculo_serrato_anterior
músculo_serrato_posterior
músculo_serrato_posterior_inferior
músculo_serrato_posterior_superior
músculo_superciliar
músculo_supraespinoso
músculo_temporal
músculo_tibial
músculo_tibial_anterior
músculo_tibial_posterior
músculo_transverso_del_abdomen
músculo_triangular_de_los_labios
músculo_tríceps
músculo_voluntario
nabo
nalga
nalgas
napia
nariz
narizota
nucléolo
núcleo
occipucio
ojo
olfato
oreja
oviscapto
oído
oído_interno
pabellón_del_oído
pandero
panza
pars_anterior
pars_distilis
pars_intermedia
parte_del_cuerpo
partes
partes_naturales
partes_pudentas
partes_vergonzosas
parótida
patas
patata
pechos
pectoral
pectoral_mayor
pectoral_menor
pene
picota
pierna
pilila
polla
pompis
poto
primer_estómago
proceso
protoplasto
próstata
pulmón
pulmón_derecho
pulmón_izquierdo
páncreas
rasgo
receptor
recto
redecilla
región
riñón
romboides_mayor
romboides_menor
rudimento
segundo_estómago
seno
senos
serrato
serrato_anterior
serrato_posterior
serrato_posterior_inferior
serrato_posterior_superior
sexo
sistema
sóleo
tejido
tendón_de_la_corva
tercer_estómago
testículo
teta
tetilla
tetina
tiroides
trapecio
trasero
tripa
trompa
tríceps
tubo_uterino
turma
tórax
ubre
ventosa
verga
vesícula_biliar

vientre
vulva
vísceras
yeyuno
zona
área
íleon
órgano_auditivo
órgano_contráctil
órgano_de_Corti
órgano_del_habla
órgano_efector
órgano_eréctil
órgano_excretor
órgano_excretorio
órgano_externo
órgano_interno
órgano_olfactivo
órgano_reproductivo_interno_femenino
órgano_reproductivo_interno_masculino
órgano_respiratorio
órgano_secretatorio
órgano_segregatorio
órgano_sensorial
órgano_vital
órganos_genitales
órganos_reproductores
órganos_sexuales
útero

Sentido 3

Banco_Central
F.B.I.
FBI
Proyecto_Manhattan
aduanas
agencia
agencia_de_la_ONU
agencia_de_las_Naciones_Unidas
agencia_ejecutiva
agencia_meteorológica
banco_central
departamento
oficina
oficina_de_aduanas
oficina_de_la_seguridad_social
oficina_de_patentes
seguridad_social
servicio_meteorológico
servicio_secreto
servicios_secretos

Sentido 4

abajera
acimboga
acícula
adormidera
aguacate
albaricoque
almendra
almendra_garrapiñada
almez
amento
amplexicaule
amplexicaulo
anacardo
ananá
ananás
anillo
antera
anteridio
anís
aparato_reproductor
apotecio
arándano
asca
ascocarpo
ascospora
ascóspora
avellana
azufaifa
banana
basidio
basidiocarpo
basidiospora
basidióspora
baya
baya_de_cambrón
baya_de_enebro
baya_de_espina_cerval
baya_del_saúco
bellota
boniato
bulbo
cabezuela
cabillo
cacahuate
cacahuete
cacahuey
café
calabaza
camuesa
candelilla
capullo
caqui
carambola
cardamomo
cariópside
carpelo
carpospora
carpóforo
castaña
castaña_de_Indias
caña_de_azúcar
cebolla
cepo
cereza
cereza_picota
chalaza
chalota
chalote
chirimoya
chirivía
cidra
ciruela
ciruela_claudia
ciruela_damascena
ciruela_pasa
clamidospora
coco
comino
conidio
corimbo
corola
cotiledón
crisantemo
cálamo
cáliz
cáudice
cítrico
delicia
diente_de_león
drupa
dátil
endosperma
enebrina
eneldo
episperma
epispermo
espiga
espora
esporada
esporangio
esporocarpio
esporocarpo
esporofilo
esqueje
estambre
estigma
estilo
estipe
estróbilo
estípite
eusporangio
falso_fruto
filamento
filodio
flor
flor_apétala
follaje
folíolo
frambuesa
fresa
fresón
fronda
fruta
fruta_cítrica
fruta_seca
fruto
fruto_del_árbol_del_pan
funículo
gajo
gametangio
garbanzo
gema
gineceo
gleba
granada
granadilla
grano
grosella
grosella_espinosa
grosella_negra
grosella_roja
guanábana
guayaba
guinda
guisante
haba
hayuco
higo
higo_chumbo

higo_de_pala
higo_de_tuna
himenio
hinojo
hinojo_hediondo
hoja
hoja_acorazonada
hoja_acuminada
hoja_aovada
hoja_aserrada
hoja_astada
hoja_bipinnada
hoja_compuesta
hoja_de_la_higuera
hoja_de_nenúfar
hoja_deltoide
hoja_dentada
hoja_denticulada
hoja_elíptica
hoja_ensiforme
hoja_entera
hoja_espatulada
hoja_festoneada
hoja_floral
hoja_imparipinada
hoja_lanceolada
hoja_linear
hoja_lirada
hoja_lobulada
hoja_oblonga
hoja_obovada
hoja_obtusa
hoja_orbicular
hoja_palmadocompuesta
hoja_palmeada
hoja_panduriforme
hoja_paripinnada
hoja_partida
hoja_peltada
hoja_perfoliada
hoja_pinnaticompuesta
hoja_quinquefoliada
hoja_reniforme
hoja_runcinada
hoja_sagitada
hoja_simple
hoja_trifoliada
hueso
inflorescencia
judión
judía_blanca
judía_pinta
judía_seca
kiwi
legumbre
lenteja
leptosporangio
lichi
lima
limbo
limón
linaza
lámina
macrospora
mandarina
mandrágora
mango
manzana
manzana_camuesa
manzana_delicia
manzana_golden
manzana_reineta
manzana_silvestre
manzana_ácida
maní
maracuyá
margen
megaspora
melocotón
melón
membrillo
micelio
microspora
mimbre
mirtilo
moniato
mora
moscatel
mostaza
nabina
naranja
naranja_california
naranja_dulce
naranja_nável
naranja_valenciana
nuez_de_cedro
nuez_de_cola
nuez_de_macadamia
piña_de_abeto
piñón
placenta
plátano
pomelo
poro
pseudocarpo
pseudofruto
pétalo
píleo
quivi
rabillo
racimo
rama
ramita
ramo
raquis
raíz
receptáculo
retoño
rizoma
rábano
rábano_blanco
rábano_picante
salsifí
sandía
sapote
semilla_de_apio
semilla_de_ricino
semillas_oleaginosas
serba
seta
seudocarpo
seudofruto
sombrerete
sombrerillo
sombrerito
sombrero
soro
sépalo
sésamo
tallo
talo
tamarindo
tetraspora
tetrasporangio
tijereta
tocón
troncho
tubérculo
tálamo
túbulo
uva
vaina
vara
vareta
varilla
varita
velo_parcial
velo_universal
verdor
verdugo
verdugón
verdura
vilano
vimbre
volva
vástago
yema
zanahoria
zarcillo
zarza
zarzamora
zoospora
ánulo
ñame
órgano_de_una_planta
órgano_vegetal

Sentido 5

concertina
harmonio
órgano_americano

PARTIDO

Sentido 1

campeonato
carrera
certamen
combate_de_boxeo
combate_de_boxeo_profesional
combate_de_lucha
competición
competición_atlética
concurso
concurso_atlético
cuartos
cuartos_de_final
final
final_de_copa
final_de_liga
jabalina
juego

lanzamiento_de_jabalina
lanzamiento_de_martillo
lanzamiento_de_peso
martillo
pelea_de_gallos
playoff
prueba_de_buceo
prueba_de_campo
prueba_de_natación
salto_con_pértiga
salto_de_altura
salto_de_longitud
salto_de_pértiga
semifinal
torneo

Sentido 2

C.D.C.
CDC
Convergència_Democràtica_de_Catalunya
Cruz_Roja
E.R.C.
ERC
Ejército_de_Salvación
Esquerra_Republicana_de_Catalunya
F.I.S.
F.N.
FIS
FN
Frente_Islámico_de_Salvación
Frente_Nacional
P.C.E.
P.N.V.
P.P.
P.S.A.
P.S.C.
P.S.E.
P.S.O.E.
P.S.U.C.
PCE
PNV
PP
PSA
PSC
PSE
PSOE
PSUC
Partido_Comunista
Partido_Comunista_de_España
Partido_Comunista_de_los_Estados_Unidos
Partido_Conservador
Partido_Demócrata
Partido_Laborista
Partido_Liberal
Partido_Nacional_Socialista_Alemán
Partido_Nacionalista_Vasco
Partido_Popular
Partido_Republicano
Partido_Socialista_Obrero_Español
Partido_Socialista_de_Andalucía
Partido_Socialista_de_Euskadi
Partido_nazi
Partit_Socialista_Unificat_de_Catalunya
Partit_dels_Socialistes_de_Catalunya
U.C.D.
U.D.C.
U.P.N.
UCD
UDC
UPN
Unió_Democràtica_de_Catalunya
Unión_de_Centro_Democrático
Unión_del_Pueblo_Navarro
alianza
asociación
autarquía
autocracia
coalición
comisión
comité
compañía
defensa
delegación
democracia
diputación
dotación
encomienda
grupo_musical
guardia
hegemonía
institución
instituto
jerarquía
misión
misión_religiosa
oligarquía
orden
organización
partido_político
personal
plantilla
plutocracia
república
sección
sindicato
tecnocracia
teocracia
tories
unidad

Sentido 3

-

Sentido 4

-

Sentido 5

-

Sentido 6

-

Sentido 7

-

PASAJE

Sentido 1

acceso
acueducto
aerovía
alcantarilla
alcantarillado
algalia
aliviadero
bajante
boca
boca_de_incendio
boca_de_riego
boquilla
bureta
caja_de_la_escalera
camino
canal
canalón
capilar
carretera
catacumbas
catéter
caz
cañería
caño
cañón
cañón_de_agua
cañón_de_chimenea
cerbatana
chimenea
colector_de_escape
conducto
conducto_de_admisión
conducto_de_entrada_del_aire
corredor
cuello
cuentagotas
cámara
cánula
derramadero
desaguadero
desagüe
difusor
embornal
escalera
escalinata
fuste_de_chimenea
galería
garganta
gasoducto
hueco
hueco_de_la_escalera
humero
imbornal
lanzatorpedos
manguera
manguera_de_bomberos
narguile
oleoducto
pasarela
pasillo
paso_subterráneo
pipa
pipa_de_brezo
pipa_de_cerámica
pipa_de_la_paz
pipeta
prioridad
prioridad_de_paso
probeta
rampa_de_caída_de_carbón
saetín
sifón
sumidero
tobogán
tobogán_de_agua
trayectoria

tubería
tubería_de_distribución
tubería_general
tubería_general_del_agua
tubería_general_del_gas
tuberías
tubo
tubo_acústico
tubo_de_admisión
tubo_de_desagüe
tubo_de_escape
tubo_de_estufa
tubo_de_lámpara
tubo_en_espiral
tubo_snorkel
túnel
vertedor
vestíbulo
vía
vía_aérea
vía_fluvial

Sentido 2

apartado
artículo
borrador
capítulo
carta
cierre
cláusula
corte
encadenado
entrega
episodio
estancia
estrofa
final
flashback
fundido
introducción
letra
libro
retrovisión
salto
sección

Sentido 3

adagio
adaptación
cadencia
canción
composición
cuarteto
divertimento
dúo
estudio
fantasía
frase
frase_musical
impromptu
intermezzo
mezcla
modulación
movimiento
nocturno
obra
octava
pastiche
pastoral
pieza
pieza_musical
poema_sinfónico
popurrí
quinteto
recitado
recitativo
septeto
serenata
sexteto
solo
suite
tocata
trío

Sentido 4

abono
billete
billete_de_autobús
billete_de_avión
billete_de_ida_y_vuelta
billete_de_tren
bono
bonocity
bonotren
comprobante
cupón
declaración_financiera
estado_de_cuentas
evaluación
inter-rail
interrail
justificante
orden_de_compra
pase
resguardo
tasa
tasación
ticket
tique
vale
valoración

Sentido 5

-

Sentido 6

aliño
anulación
apertura
cambio_de_color
cambio_de_estado
cocción
comienzo
decoración
defoliación
degradación
embellecimiento
ensuciamiento
especialización
esterilización
iniciación
inicio
inversión
masticación
mejora
mejoramiento
occidentalización
principio
transferencia
trasferencia
urbanización

PENDIENTE

Sentido 1

alhaja
anillo
brazalete
collar
cuenta
gargantilla
gema
gemelos
joyas
joyería
piedra_preciosa
pinza_de_corbata
pulsera
zarcillo

Sentido 2

gradualidad
precipitación

Sentido 3

caída

Sentido 4

abajadero
acantilado
acuífero
ascenso
bajada
brollador
cadena_montañosa
colina
collado
cordillera
cresta
cuesta
declive
delta
depresión
descenso
desembocadura
despeñadero
elevación
emanación
farallón
formación_geológica
fuente
fuente_natural
ladera
ladera_de_montaña
ladera_de_monte
loma
macizo
manantial
margen
masa_de_hielo
montaña
monte
nivel_freático
orilla
orilla_del_agua
otero
peralte
peñasco
playa

precipicio
promontorio
ribera
risco
sierra
subida
surtidor
talud
vera
veta

Sentido 5

-

Sentido 6

-

Sentido 7

-

Sentido 8

-

PROGRAMA

Sentido 1

activismo
alboroto
animación
antojo
ardid
artimaña
bloqueo
bullicio
complot
concepción
concepto
confabulación
conjuración
conspiración
contracomplot
creencia
dibujo
diseño
error
esquema
estratagema
estrategia
fiscalidad
generalidad
generalización
idea
ideal
idealización
impresión
inspiración
intriga
itinerario
jaleo
juerga
jugada
maquinación
megaproyecto
melé
menú
motivo
noción
obscurantismo
pase
pase_adelantado
pase_lateral
pensamiento
plan
plan_de_acción
plan_de_actuación
planing
política
preocupación
proyecto
reacción
revuelo
ruta
régimen
sentido
sentimiento
significado
sugerencia
tema
teorema
trampa
trapisondeo
treta
táctica

Sentido 2

biorretroacción
contabilidad
código_ético
disciplina
lógica
teología
ética

Sentido 3

carta_adjunta
certificación
confesión
copyright
cédula
derechos_de_autor
dimisión
documentación
documento
documento_adjunto
documento_judicial
documento_oficial
fax
formulario
fuero
papeleta
papiro
patente
programa_político
resolución
telecopia
telefax

Sentido 4

compilador
editor
ensamblador
instrucción
intérprete
orden
programa_de_servicio
programa_fuente
programa_informático
programa_objeto
programas_compatibles
scandisk
sistema_operativo
software
software_compatible
subrutina
supervisor
traductor

Sentido 5

actuación
atracción
concurso
culebrón
debate_televisivo
documental
emisión
episodio
episodio_piloto
espectáculo
espectáculo_de_sombras
espectáculo_de_striptease
espectáculo_de_títeres
espectáculo_de_variedades
film
filme
función
noticiario
obra
película
producción
programa_de_entrevistas
programa_de_televisión
programa_piloto
reposición
representación
serial
serie
sesión
show
sitcom
striptease
telecomedia
telediario
telenovela
variedades

Sentido 6

arcano
base_de_datos
confirmación
currículo
datos
estudio
evidencia
formato
fuente
fuente_de_información
hecho
información
informe
material
memoria
nuevas
plan_de_estudios
programa_de_estudios
programa_de_lectura
propaganda
ratificación
secreto

Sentido 7

amonestaciones

anunciación
anuncio
aviso
banas
boletín
cartel
comunicación
comunicado
despido
necrología
notificación
obituario
pasaporte
programa_de_carreras

TABLA

Sentido 1

capa
cartelera
chapa
chapa_de_metal
encerado
escurridero
esquí_acuático
fibra_prensada
hoja_de_vidrio
laminado
lámina
marcador
membrana
mesa_de_dibujo
monopatín
paleta
panel
pizarra
plancha
plancha_de_surf
plato_fotográfico
tabla_de_cortar
tabla_de_picar
tabla_de_planchar
tabla_de_surf
tabla_del_pan
tabla_del_suelo
tablón_de_anuncios
tajadero
trampolín
trampolín_de_saltos

Sentido 2

ábaco

Sentido 3

carta_astral
cuadro
diagrama_de_barras
diagrama_de_pastel
figura
gráfica
gráfico
histograma
ilustración
ordinograma
organigrama
perfil
pirámide_de_población
rueda_cromática

Sentido 4

batería
calendario
columna
espectro
fila
formación
hilera
matriz
orden
panoplia
renglón
tabla_de_mortalidad
tabla_estadística
tabla_periódica

Sentido 5

aglomerado
conglomerado
conglomerado_de_madera
madera
nudo
tabla_de_abeto
tabla_de_pino
tablón

TIMBRE

Sentido 1

-

Sentido 2

adaptador
adminículo
aguja
alarma
alcoholímetro
alcohómetro
aldaba
anticonceptivo
apagador
aparato
aparato_acústico
aparato_dental
aparato_electrónico
aparato_eléctrico
aplicador
apoyo
armatoste
artefacto_explosivo
artilugio
autocue
autodirección
autopiloto
aventador
bomba_de_cobalto
botón_de_arranque
bula
calentador
campana_de_salvamento
cañón_de_nieve
cerbatana
chupete
cimbel
cinetoscopio
conductor
contraceptivo
conversor
convertidor
correctivo
deflector
depresor
depuradora
descodificador
descremadora
deshumidificador
desnatadora
detector
dispositivo
dispositivo_acústico
encendedor
estabilizador
estampilla
estereocomparador
extintor
extractor
filtro
fotocomponedora
freno
geófono
hidrojardinera
huella
humidificador
impresión
impronta
imán
indicador
ingenio
instrumento
interruptor
kinetoscopio
ladrón
limitación
llamador
llave
luz
lápiz_electrónico
machacadora
maneta
marchamo
mecanismo
mecanismo_acústico
mecanismo_explosivo
mechero
moledora
máquina
máscara
palpador
pasterizador
pasteurizador
peine
picaporte
piloto_automático
pinzas
plantilla
plectro
posquemador
prompter
protección
prótesis_dental
púa
rascaespalda
reflector
reset
restricción
sacabotas
sacamuestras

sacatestigos
sello
sensor
sistema_de_alarma
sistema_de_seguridad
soporte
starter
tabla_de_lavar
teclado
teleapuntador
terraja
tetina
tomamuestras
trampa
trituradora
ventilador
válvula

Sentido 3

estrépito
jaleo
ruido
unísono
voz

Sentido 4

afonía
aspereza
bronquedad
cadencia
color
cualidad
disonancia
dureza
enronquecimiento
estridencia
estridor
fuerza
gangosidad
intensidad
madurez
musicalidad
nasalidad
plenitud
resonancia
riqueza
ronquedad
ronquera
silencio
sonoridad
suavidad
tajada
tono
volumen

Sentido 5

IRPF
arancel
canon
capitación
carga
censo
contribución
derechos_de_sucesión
gabela
gravamen
imposición
impuesto
impuesto_adicional
impuesto_de_plusvalía
impuesto_directo
impuesto_indirecto
impuesto_sobre_el_consumo
impuesto_sobre_la_renta
pecho
plusvalía
póliza
recargo
sobretasa
tarifa
tasa
tributación
tributo

Sentido 6

-