Extracción de relaciones semánticas entre nombres y verbos en EuroWordNet

Julio Gonzalo, Felisa Verdejo, Irina Chugur, Fernando López and **Anselmo Peñas**
UNED
Ciudad Universitaria, s.n.
28040 Madrid - España
{julio,felisa,irina,flopez,anselmo}@ieec.uned.es

Abstract

Una de las carencias del WordNet de Princeton (base de datos léxica con relaciones semánticas entre palabras de inglés americano) es la falta de relaciones semánticas entre palabras con distintas categorías gramaticales. En el proyecto EuroWordNet (cuyo objetivo es construir una base de datos léxica multilingüe al estilo de WordNet), una de las extensiones consideradas con respecto a WordNet es la inclusión de este tipo de relaciones semánticas. En este artículo se discuten algunas posibilidades de extraer estas relaciones de forma semiautomática, a partir del estudio manual de un conjunto de parejas nombre-verbo obtenidas por criterios morfológicos.

1 Introducción

WordNet (Miller et al., 1990) es una base de datos léxica para el inglés. Consiste en relaciones semánticas entre palabras inglesas, en la que las palabras con un significado similar se agrupan en los llamados *synsets* (synonym sets o conjuntos de sinónimos). Además de la sinonimia (implícita en la definición de synset), se establecen otras relaciones semánticas entre synsets (o, excepcionalmente, entre palabras): hiponimia/hiperonimia (relación IS-A), que dota a la red de una estructura jerárquica; meronimia/holonimia (relación HAS-A) en sus variantes *parte, miembro* y *sustancia*; y antonimia (entre palabras opuestas). Con estas relaciones, WordNet se configura como una red de 168.000 synsets (asimilables a conceptos) que contienen 126.000 palabras diferentes.

El objetivo del proyecto EuroWordNet (EWN) (Vossen, 1998) es desarrollar (semiautomáticamente) una base de datos léxica multilingüe, inspirada en WordNet, con relaciones semánticas entre palabras de varios idiomas de la comunidad europea: Italiano, Inglés, Holandés y Castellano (y, desde Abril de 1.998, también Alemán, Checoslovaco, Estonio y Francés). El proyecto comenzó en Marzo de 1.996 y tiene una duración de tres años.

La característica que distingue a la base de datos de EWN es, sobre todas las demás, su naturaleza multilingüe, que no discutiremos aquí. Pero, además, incorpora sobre la estructura original de WordNet algunos tipos de información que son relevantes para muchas aplicaciones y que WordNet no contempla: entre otros, las etiquetas de *dominio* y las relaciones semánticas entre distintas categorías léxicas.

En WordNet, cada categoría léxica abierta (nombres, verbos, adjetivos y adverbios) da lugar a una red semántica de synsets conectados entre sí. Sin embargo, no se contempla la posibilidad de establecer relaciones semánticas entre distintas partes de la oración, como entre verbos y nombres. Por ejemplo, entre el synset nominal

```
design, designing -- (the act of working
out the form of something)
```

y el synset verbal

```
design1 -- (create the design for;
"design a dress"; "design a house")
```

no puede establecerse, dada la estructura de WordNet 1.5, ninguna relación semántica. Es obvio que esa restricción obedece a criterios morfosintácticos y no semánticos, puesto que los significados de nombres y verbos están estrechamente ligados. De hecho, las relaciones semánticas entre nombres y verbos pueden ser muy útiles para muchas aplicaciones de una red semántica como WordNet o EuroWordNet.

Un ejemplo muy llamativo es de de la Recuperación de Información (Gonzalo et al., In press). Dado el problema de encontrar todos los documentos relevantes en una colección a partir de una interrogación formulada en lenguaje natural, una red semántica ofrece, al menos, dos ventajas evidentes:

- Ofrece la posibilidad de discriminar sentidos en preguntas y documentos. Así puede evitarse, por ejemplo, encontrar documentos sobre agricultura cuando se pregunta por abonos para asistir a todas las sesiones de un festival de música.
- Ofrece la oportunidad de buscar documentos que no contienen los términos originales de la pregunta. Por ejemplo, permite relacionar do-

cumentos sobre *estiércol* o *mantillo* con una pregunta sobre *abono*.

- En el caso de EuroWordNet, permite plantearse la recuperación de información multilingüe mediante criterios semánticos independientes del lenguaje.

Curiosamente, si se utilizan synsets de WordNet para representar los documentos (en lugar de las palabras que en ellos aparecen), los resultados pueden ser mucho mejores (Gonzalo et al., 1998), pero la falta de relaciones nombre-verbo evita que se relacionen, por ejemplo, *design* como *diseño* y como *diseñar* (aparecen en lugares distintos e incomunicados de la red semántica), mientras que la representación convencional en Recuperación de Información, mediante palabras, no distingue entre los dos usos de *design* y, paradójicamente, tiene más posibilidades de relacionarlos con éxito. Creemos, por ello, que cualquier sistema de recuperación de información que se base en redes semánticas como WordNet o EuroWordNet debe tener la posibilidad de relacionar el significado de nombres y verbos sin depender de la categoría léxica a la que pertenecen.

La metodología con la que se está construyendo el wordnet para el castellano, dentro del proyecto EuroWordNet, se apoya principalmente en el enlace de palabras en castellano con los synsets de WordNet 1.5, para heredar así las relaciones semánticas entre synsets de este último (Atserias et al., 1997). En el caso de las relaciones nombre-verbo, que están ausentes de WordNet 1.5, debemos buscar otros medios para establecerlas. Para ello hemos adoptado la siguiente metodología: primero, establecer y evaluar técnicas automáticas para generar parejas nombre-verbo potencialmente relacionadas. Y segundo, establecer en qué forma se relacionan los distintos sentidos de nombre y verbo, y cómo se etiquetan esas relaciones.

Para la selección de parejas candidatas disponemos de varios criterios:

- Derivación morfológica de nombres a partir de la raiz verbal, o viceversa, en castellano. Esta es la variante que analizamos en este artículo.

- Análisis de las definiciones de diccionarios, en busca de patrones. Por ejemplo, *alicatador* definido como *Instrumento para alicatar* nos sugiere una relación INVOLVED_INSTRUMENT (instrumento involucrado en la acción del verbo) entre alicatador y alicatar.

- Aplicación de las técnicas anteriores sobre WordNet 1.5 y extrapolación al castellano. Esta última técnica es posible gracias a que la estrategia básica para crear el wordnet castellano es insertar palabras del castellano en synsets del inglés.

- Utilización del índice interlingua de EuroWordNet para extrapolar relaciones de otros idiomas incluidos en el proyecto.

En este artículo nos hemos centrado en el primero de los criterios. Para ello hemos generado, a partir de los verbos disponibles actualmente en el wordnet español, todas las parejas nombre-verbo correspondientes a cinco sufijos. De entre ellas, nos hemos quedado con las que el nombre generado aparece también en el wordnet español. La lista final de parejas candidatas tiene 580 elementos.

Para analizar los vínculos semánticos entre cada pareja, hemos construido una interfaz que permite establecer todas las relaciones semánticas entre los distintos sentidos del nombre, por un lado, y del verbo, por el otro. El objetivo es doble: por un lado, enriquecer la base de datos con relaciones nombre-verbo establecidas manualmente. Por otro, disponer de datos fiables a partir de los cuales extrapolar criterios de asignación automática que puedan ser aplicados sobre el wordnet final, del que dispondremos en la primavera de 1.999.

En la sección siguiente resumimos las relaciones nombre-verbo consideradas en EuroWordNet, así como los procesos morfológicos de derivación y su relación con la semántica de los nombres derivados. A continuación describimos la metodología de adquisición en más detalle. Finalmente, discutimos los resultados obtenidos y sus implicaciones para nuestro trabajo en curso.

2 Tipología de las relaciones nombre-verbo en EuroWordNet

Las relaciones que EuroWordNet define para establecer enlaces entre synsets de distintas categorías gramaticales son:

- XPOS_NEAR_SYNONYM (Quasisinónimo : *caza* XPOS_NEAR_SYNONYM *cazar*. XPOS, aquí y en adelante, hace referencia al carácter intercategorial de la relación; es decir, se establece entre partes de la oración diferentes).

- HAS_XPOS_HYPERONYM (Hiperónimo : *eclipse* HAS_XPOS_HYPERONYM *oscuro*).

- HAS_XPOS_HYPONYM (Hipónimo : *oscuro* HAS_XPOS_HYPONYM *eclipse*).

- XPOS_NEAR_ANTONYM (Quasiantónimo : *desconexión* XPOS_NEAR_ANTONYM *conectar*).

- ROLE & INVOLVED (Papel temático & Implicado : *libro* ROLE *leer* / *leer* INVOLVED *libro*).

- CAUSES & IS_CAUSED_BY (Causa & efecto : *matar* CAUSES *muerte* / *muerte* IS_CAUSED_BY *matar*).

- HAS_SUBEVENT & IS_SUBEVENT_OF (tiene un subevento & es subevento de : *lavar* HAS_SUBEVENT *mojar* / *mojar* IS_SUBEVENT_OF *lavar*).

- XPOS_FUZZYNYM (relación semántica sin precisar : *especiero* XPOS_FUZZYNYM *especiar*).

La especificidad del criterio para establecer el subconjunto de nombres y verbos seleccionado - la existencia de lazos morfológicos entre los componentes de cada pareja - ha determinado la exclusión de esta lista de las siguientes relaciones: HAS_SUBEVENT, HAS_XPOS_HYPONYM, y HAS_XPOS_HYPERONYM. De esta forma, las relaciones que consideramos en el experimento son XPOS_NEAR_SYNONYM, XPOS_NEAR_ANTONYM, ROLE & INVOLVED, CAUSES & IS_CAUSED_BY, y XPOS_FUZZYNYM.

A continuación, se expondrán los criterios que se han tenido en cuenta para codificar los enlaces detectados. En caso de relaciones que cuentan con tests, éstos también se incluirán. Los tests o marcos diagnósticos han sido elaborados para verificar relaciones. Constituyen patrones generales y están basados en juicios de normalidad. Insertando dos palabras en el marco propuesto para una relación habitualmente se provoca una respuesta de tipo sí/no que refleja la aceptabilidad construcción obtenida. Algunas veces se admiten respuestas de tipo "probable", "no está claro", "poco probable". Teóricamente, deberíamos disponer de tests para todas las relaciones en cada una de las lenguas participantes. Hay que subrayar que estos marcos han sido concebidos para detectar exclusivamente relaciones semánticas y, por lo tanto, no cubren diferencias de registro, estilo o dialecto que se establecen entre palabras.

XPOS_NEAR_SYNONYM: La relación que se codifica como XPOS_NEAR_SYNONYM se basa en una noción débil de sinonimia. Se consideran sinónimos dos palabras intercambiables en un contexto determinado, haciendo las debidas concesiones para los reajustes gramaticales necesarios. De esta manera, la sinonimia verbo-nombre queda reducida a las nominalizaciones de verbos (verbo→nombre), ya que los verbos denominales y los nombres que han servido de base para su derivación establecen otro tipo de relaciones semánticas. Partiendo de esta base, consideraremos sinónimos verbos y nombres que denotan procesos, eventos o estados.

El test básico para este tipo de relaciones se puede resumir en el siguiente marco:

$$\left.\begin{array}{l}\textsf{Si se da el caso de (el estado de) X}\\ \textsf{Si tiene lugar un X}\end{array}\right\}\ \textsf{entonces}$$
alguien/algo (se) Y

(donde X es un nombre sustantivo e Y es un verbo).

Los patrones léxicos que permiten extraer este tipo de parejas de forma automática de los diccionarios responden a tres tipos fundamentales, señalados por Irene Castellón (Castellón, 1997) en los que el nombre se define como : a) acción del verbo (*caza-cazar*); b) efecto (*enamoramiento-enamorar*); c) acción y efecto (*abaratamiento-abaratar*).

En el presente trabajo, hasta el momento, no hemos explotado esta posibilidad. En cambio, nos hemos apoyado en la morfología, puesto que una serie de sufijos (*-ción*, *-miento*, *-o/a*) se asocian con el significado de "acción y/o efecto del V" de una forma bastante regular, aunque no exenta de excepciones, ya que se pueden dar ejemplos como *aparcar/aparcamiento* o *establecer/establecimiento* en los que junto con el sentido esperado de "acción y/o efecto" se halla la acepción referida al "lugar".

XPOS_NEAR_ANTONYM: La relación de antonimia que se establece entre synsets que pertenecen a distintas partes de la oración se codifica en EWN como XPOS_NEAR_ANTONYM.

De momento, esta relación no ha sido incluida en nuestros experimentos. Sin embargo, centrándonos en el subconjunto seleccionado de verbos y nombres que comparten un mismo lexema, en un futuro se podrían establecer vínculos antonímicos, utilizando la lista de las parejas marcadas con la relación de XPOS_NEAR_SYNONYM y trabajando con la morfología: buscando sufijos negativos tanto en las entradas nominales, como en las verbales. De esta forma, se podría obtener parejas de antónimos en ambas direcciones:

1) *desintegración/integrar*
integración/desintegrar
2) *desaprovechamiento/aprovechar*
aprovechamiento/desaprovechar
3) *desatascar/atasco*

ROLE/INVOLVED: La relación de implicación hace referencia a la telicidad y a otros aspectos de implicatura semántica. En EWN este tipo de vínculos entre los significados está cubierto por las relaciones ROLE e INVOLVED. Mediante ROLE se codifica una relación semántica entre un nombre y un verbo que se refiere a una situación en la que la entidad designada por el nombre desempeña un papel temático típico seleccionado por el verbo. La relación INVOLVED si el significado del verbo está fuertemente caracterizado por un nombre (se puede tratar de los télicos, así como de la acción básica asociada a un verbo). Es decir, si la relación va del nombre al verbo, se codificará como ROLE y, en el caso contrario,hablaremos de un INVOLVED.

De acuerdo con el papel temático que desempeña el nombre, cada una de estas relaciones se puede subespecificar en AGENT (agente típico, experimentante o causante), PATIENT (paciente) , INSTRUMENT (instrumento, objeto que se utiliza típicamente para llevar a cabo cierta acción), LOCATION (lugar asociado con la acción), DIRECTION (ruta o dirección). Para aquellos casos en los que resulte difícil identificar la subespecificación se reserva la relación general de

ROLE/INVOLVED . De esta forma, se obtiene una amplia gama de enlaces que refleja diversos matices que puede adquirir la relación de implicación o involucramiento:

ROLE_AGENT	INVOLVED_AGENT
ROLE_PATIENT	INVOLVED_PATIENT
ROLE_INSTRUMENT	INVOLVED_INSTRUMENT
ROLE_LOCATION	INVOLVED_LOCATION
ROLE_DIRECTION	INVOLVED_DIRECTION
ROLE_SOURCE_DIRECTION	INVOLVED_SOURCE_DIRECTION
ROLE_TARGET_DIRECTION	INVOLVED_TARGET_DIRECTION

Los ejemplos estudiados hasta ahora sugieren la posibilidad de interpretar estos dos tipos de enlace, ROLE e INVOLVED, como dos caras de una única relación, ROLE, automáticamente reversible para dar lugar al correspondiente INVOLVED.

Los tests generales elaborados para verificar la existencia de esta relación ofrecen una serie de patrones léxicos que se pueden sintetizar en el siguiente marco:

1) "La entidad que designa el N sirve/se usa/es un lugar típico para realizar la acción expresada por el V".
2) "N es un N porque sirve/se usa para V o puede ser V-ado(-ido)".

Los sufijos que prometen ser productivos para esta clase de enlaces son *-dor*, *-ante*, *-dero*, *-torio* y sus alomorfos. *-Dor* se asocia normalmente con el papel de agente, instrumento; *-ante* se relaciona típicamente con la noción de agente, causante; *-dero* suele aportar el significado de instrumento o lugar; *-torio* participa, habitualmente, en formaciones que designan lugares.

CAUSES & IS_CAUSED_BY constituye una relación inversa de causa/efecto. Teóricamente, se puede establecer entre distintas categorías léxicas, incluidos nombres y verbos. Sin embargo,tratándose de verbos y sus nominalizaciones, resulta difícil distinguir claramente entre CAUSES y XPOS_NEAR_SYNONYM. Considérese ejemplos de tipo *aprovechamiento-aprovechar*, *enamoramiento-enamorar*. Para el sustantivo WN español no ofrece más que un solo sentido : el proceso mismo y su resultado quedan íntimamente entrelazados. Así, la relación de esta clase de nombres y los verbos correspondientes admite ambas interpretaciones (en términos de sinonimia y causa-efecto). Dado que, de acuerdo con el principio de homogeneidad adoptado para EWN, dos palabras (o dos sentidos, si se trata de palabras polisémicas) pueden ser vinculadas entre sí solamente por medio de una única relación, y que la denotación de un proceso o acción y la de su efecto están estrechamente relacionadas, hemos optado por marcar este tipo de parejas nombre-verbo como XPOS_NEAR_SYNONYM.

XPOS_FUZZYNYM: Finalmente, el enlace que recibe el nombre de XPOS_FUZZYNYM sirve para marcar las parejas cuya relación no se corresponde con ninguno de los vínculos definidos en EWN.

3 Metodología de adquisición

Los datos de partida de nuestro wordnet castellano son los del *subset 1*, segunda de las tres fases de producción de que consta el proyecto EuroWordNet. La red de nombres tiene 23.216 (para un total de 41292 sentidos agrupados en 18577 synsets). La red de verbos corresponde, principalmente, a anotaciones manuales basadas en los resultados del proyecto PIRÁPIDES (Castellón et al., 1997), y consta de 3086 formas verbales (para un total de 7953 sentidos agrupados en 3294 synsets). Descartando los verbos reflexivos y los multipalabra, nos quedan 2231 para aplicar las derivaciones morfológicas nombre-verbo relacionadas en la sección anterior.

A partir de esa lista de verbos, hemos generado automáticamente todos los derivados verbales a partir de los sufijos considerados, y después hemos filtrado los que no aparecían en el wordnet castellano. La mayoría de los nombres descartados corresponden a construcciones inexistentes en castellano, como *amar / amación*, aunque algunos son nombres válidos que, simplemente, no han sido incorporados todavía a la base de datos, como *abrasadero* o *adoratorio*. Estos últimos se almacenan para ser revisados en próximas ampliaciones de la base de datos.

Obtuvimos una lista de 1.273 parejas nombre/verbo. De ellas, seleccionamos los sufijos *-ante*, *-dero*, *-ero*, *-dor* y *-torio* (580 parejas) para tratar manualmente. El análisis del resto de los sufijos y alomorfos (*-ción*, *-mento*, *-miento*, *-imiento*, *-ón*, *-ión*, *-ica*) está actualmente en curso. La lista de 580 parejas se dividió en no-ambiguas (cuando verbo y nombre sólo tienen un significado en el wordnet castellano) y ambiguas (cuando alguno de los dos tiene más de un significado). La lista de parejas no-ambiguas también se revisó manualmente, puesto que desconocíamos, a priori, si los sentidos ya observados en la base de datos eran los adecuados, o incluso si la relación semántica realmente existía entre nombre y verbo.

Para la revisión manual hemos construido una interfaz genérica que permite establecer y verificar relaciones semánticas entre parejas de palabras arbitrarias dentro de un wordnet monolingüe. La funcionalidad es doble: por un lado, permite relacionar pares de synsets que contienen a una y otra palabra. Por otro lado, cuando el sentido adecuado existe, pero no se encuentra en la base de datos, se aprovecha la revisión manual para incluirlo, mediante su identificación con un synset del WordNet 1.5. La forma de operar de la interfaz es muy sencilla: se muestran, por un lado, los sentidos de una de las palabras (en

Sufijo	#parejas	no ambiguas	ambiguas	#rel no ambiguas	#rel ambiguas	# rel
ante	101	20	81	16	171	187
dero	67	6	61	5	124	129
dor	354	59	295	56	679	735
ero	40	8	32	3	36	39
torio	18	3	15	3	23	26
Total	580	96	484	83	1033	1116

Tabla 1: Estadísticas sobre los sufijos tratados manualmente

nuestro caso, el nombre) y, por otro, los sentidos de la otra (el verbo en este caso). Para establecer una relación, basta con pinchar un sentido de cada lista y seleccionar de un menú el tipo de relación semántica que se da entre ellos. Si falta el sentido adecuado, se muestran los synsets de WordNet 1.5 de la palabra inglesa equivalente, y se permite seleccionar aquél más adecuado.

4 Resultados

En la tabla 1 pueden verse las estadísticas sobre las relaciones obtenidas para el subconjunto de 580 parejas candidatas correspondientes al primer conjunto de sufijos. Se han obtenido un total de 1116 relaciones, lo que constituye aproximadamente 1/2 del número de formas verbales en la base de datos y una productividad de casi dos relaciones semánticas por pareja. Si se mantiene ese ratio sobre el subconjunto restante (693 parejas), se obtendrá una relación semántica por verbo. Este es un dato alentador, pues muestra que sólo mediante derivaciones morfológicas se puede dotar al wordnet castellano de una densidad de enlaces muy alta entre la red de verbos y la red de nombres. Además, faltan todavía por considerar las derivaciones de nombres a partir de verbos.

De las 580 parejas consideradas, se establecieron relaciones para 503, y 77 fueron descartadas como derivaciones incorrectas (como, por ejemplo, *turbante/turbar*). Por tanto, un 87% de parejas fueron productivas, lo que constituye un margen suficiente para intentar establecer enlaces automáticamente. Sin embargo, en total sólo 485 verbos quedaron relacionados directamente con nombres, lo que hace un 22% de los verbos considerados para el experimento. El resto de los verbos, o bien heredarán las relaciones por sinonimia, o podrán entrar en alguno de los métodos restantes (patrones de definiciones en el Vox, enlace indirecto a través de WordNet 1.5, etc).

Las relaciones predichas incorrectamente por el proceso de derivación morfológica se agrupan en tres fenómenos:

- La presencia de falsos derivados: *librero/librar*, *montero/montar*, *ojeador/ojear*.
- Derivados cuya relación se puede explicar etimológicamente y que, sin embargo, no existe en la lengua actual: *dominante/dominar*; préstamos y palabras patrimoniales que provienen de la misma raíz latina : *restaurante/restaurar*.
- Palabras que pertenecen a la misma familia pero no están relacionadas directamente, el vínculo semántico existe aunque resulta ser bastante remoto : *aceitero/aceitar* (la derivación a partir de *aceite*), *especiero/especiar* (relacionadas ambas con *especias*).

Respecto a las relaciones válidas pero entre sentidos que no aparecían en el wordnet castellano, hubo que añadir manualmente 116 nuevos sentidos (de palabras que ya tenían uno o más sentidos en el wordnet) para establecer todas las relaciones semánticas potenciales entre las parejas de candidatos. Esto constituye un 10% del número de relaciones establecidas, que es una cifra no despreciable a la hora de considerar procedimientos de enlace automático (que sólo intentarán encontrar relaciones entre los sentidos que existen previamente en el wordnet).

Finalmente, nos interesa saber si el tipo de relación entre un nombre y un verbo puede predecirse a partir de la derivación morfológica de la que proviene, con vistas a automatizar el proceso de enlazar synsets de verbos y nombres. En la tabla 2 se desglosan las relaciones obtenidas para cada sufijo según el tipo de enlace que se establece.

Entre los sufijos seleccionados para elaborar las listas de parejas hay algunos que, a pesar de no mostrar un comportamiento regular, restringen los significados que podrían aportar a la base. Se da una serie de regularidades, en cuanto a la distribución de los sufijos y las relaciones. Así, a pesar de no existir una sistemática clara, las posibles relaciones que acompañan la aparición de los sufijos seleccionados se organizan fundamentalmente en torno a las nociones de agente, instrumento (*-dor*, *-ante*, *-dero*), lugar(*-torio*) :

Sufijo	ROLE	AGENT	PATIENT	INSTRUMENT	LOCATION	XPOS_FUZZYNYM	XPOS_NEAR_SYN	Total
ante	35	126	1	22	0	2	1	187
dero	12	2	4	47	55	9	0	129
dor	18	460	1	224	19	12	0	735
ero	1	22	0	4	0	12	0	39
torio	4	0	0	4	15	0	3	26
Total	70	610	6	301	89	35	4	1116

Tabla 2: Distribución por tipo de relación y sufijo

decorador	ROLE_AGENT	*decorar*
aspirante	ROLE_AGENT	*aspirar*
remero	ROLE_AGENT	*remar*
asador	ROLE_INSTRUMENT	*asar*
aislante	ROLE_INSTRUMENT	*aislar*
comedero	ROLE_INSTRUMENT	*comer*
comedor	ROLE_LOCATION	*comer*
vertedero	ROLE_LOCATION	*verter*
dormitorio	ROLE_LOCATION	*dormir*

Se puede apreciar una mayor ambigüedad en los sufijos -dor y -dero que pueden aparecer ligados a ROLE_AGENT, ROLE_INSTRUMENT y ROLE_LOCATION. El sufijo *-torio* se muestra más predecible, dando lugar casi exclusivamente a formaciones que establecen la relación de ROLE_LOCATION con el verbo base.

Los casos de sinonimia son muy aislados: *lavatorio* XPOS_NEAR_SYNONYM *lavar* (*lavatorio* en el sentido de *acto de lavar*)

En lo que respecta a los sufijos *-ción*, *-miento*, éstos muestran un comporatamiento bastante regular. Dentro del conjunto analizado funcionan sistemáticamente como marcadores de sinonimia. Aunque, teóricamente, queda la posibilidad de que aparezcan algunos de los contraejemplos mencionados en el apartado dedicado a la relación de XPOS_SYNONYM. De momento queda por comprobar si el número de interferencias de este tipo es significativo para la base de datos de EWN.

El trabajo realizado sobre la lista de parejas ambiguas permite hacer una serie de observaciones acerca de la compatibilidad entre distintos tipos de enlaces. Si un nombre tiene varios sentidos, las relaciones que éstos entablan con las correspondientes acepciones del verbo pueden ser: a) de un mismo tipo (*bogador/bogar*, *visitante/visitar*) o b) de naturaleza distinta. Si se da este último caso, las combinaciones que se producen entre distintos enlaces dentro de una palabra polisémica muestran cierta regularidad. Generalmente aparecen juntas ROLE_AGENT y ROLE_INSTRUMENT (*apuntador/apuntar*). Se pueden encontrar igualmente, aunque con una frecuencia menor, ROLE_INSTRUMENT y ROLE_LOCATION (*achicharradero/achicharrar*, *bailadero/bailar*). Estas dos combinaciones prácticamente agotan todas las posibilidades para el conjunto de sufijos con que hemos trabajado, exceptuando el caso de *lavatorio/lavar* comentado más arriba.

Las estadísticas de la tabla 2 muestran estas tendencias, pero, desafortunadamente, ninguna estadística es suficientemente rotunda para predecir con un margen de confianza suficiente el tipo de relación a partir del sufijo:

- Para el sufijo *-ante*, ROLE_AGENT es la relación predominante, pero ROLE también aparece en un 19% de los casos, y ROLE_INSTRUMENT en un 12%.
- Para el sufijo -dero, ROLE_INSTRUMENT y ROLE_LOCATION son los más frecuentes, aunque también aparecen ROLE y XPOS_FUZZYNYM.
- Para el sufijo *-dor*, ROLE_AGENT es la relación predominante (y muy productiva, con 460 relaciones, es decir, un 40% del número total de relaciones establecidas manualmente). Pero también aparecen 224 ROLE_INSTRUMENT, y cifras no despreciables de ROLE, ROLE_LOCATION y XPOS_FUZZYNYM.
- Para el sufijo *-ero*, ROLE_AGENT es la relación predominante, pero no puede despreciarse XPOS_FUZZYNYM.
- Finalmente, para *-torio* la predominante es ROLE_LOCATION, pero también aparecen ROLE, ROLE_INSTRUMENT y XPOS_NEAR_SYNONYM de forma significativa.

5 Conclusiones

La derivación morfológica verbo → nombre resulta ser altamente productiva para establecer relaciones semánticas entre nombres y verbos. Hemos producido manualmente 1.116 relaciones a partir de 2.247 verbos, y estimamos que con los sufijos que faltan por tratar se puede llegar a tantas relaciones como verbos.

Esas 1.116 relaciones verificadas manualmente ofrecen una colección de prueba interesante para evaluar posibles estrategias que determinen relaciones nombre-verbo de forma automática. Sin embargo, la posibilidad de que esas estrategias sean precisas para el wordnet castellano no es muy alta, debido a la combinación de varios factores: por un lado, en un 10% de ocasiones alguno de los sentidos adecuados no estaba presente en la base de datos (este problema debería haber mejorado sustancialmente cuando se produzca la versión final de la base de datos). Por otro lado, la información contextual de la que se dispone con este método para hacer la desambiguación es escasa. Y con ella es necesario tanto elegir los sentidos entre los que existe conexión semántica, como decidir qué tipo de relación es la que existe entre ellos. Entre las posibles estrategias están la trasladar la desambiguación al plano del WordNet 1.5 mediante el índice interlingua de EuroWordNet, para aprovechar la riqueza de las jerarquías y la información contextual que pueden ofrecernos las glosas de éste. Entonces pueden aplicarse restricciones selectivas y criterios de proximidad semántica entre las glosas para predecir las relaciones correctas.

Nuestro trabajo en curso consiste en completar el estudio manual para todos los sufijos considerados, extenderlo a las derivaciones nombre → verbo (aunque, como hemos comentado, son menos predecibles) y, muy especialmente, utilizar como fuente complementaria los patrones de definiciones en el diccionario VOX del castellano. Por ejemplo, el VOX contiene más de siete mil definiciones para nombres con el patrón "Acción y efecto de", que puede ser asociado directamente a unas relaciones nombre/verbo determinadas.

Agradecimientos

Este trabajo ha sido financiado parcialmente por la CE, proyecto LE #4003, y también parcialmente por la CICyT, proyecto TIC-96-1243-CO3-O1. Estamos en deuda con Irene Castellón y Toni Martí por su orientación en los comienzos de este trabajo.

Referencias

J. Atserias, S. Climent, J. Farreres, G. Rigau, and H. Rodríguez. 1997. Combining multiple methods for the automatic construction of multilingual wordnets. In *Proceedings of the Conference on Recent Advances on NLP (RANLP'97)*.

I. Castellón, M.A. Martí, R. Morante, and G. Vázquez. 1997. Tipología de diátesis para el español y el catalán. *Revista de la SEPLN*.

I. Castellón. 1997. Extracción de información de fuentes diccionariales. Technical Report UB-LG 1997#3, UB.

J. Gonzalo, M. F. Verdejo, I. Chugur, and J. Cigarrán. 1.998. Indexing with Wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*.

J. Gonzalo, M. F. Verdejo, C. Peters, and N. Calzolari. In press. Applying EuroWordnet to multilingual text retrieval. *Journal of Computers and the Humanities, Special Issue on EuroWordNet*.

G. Miller, C. Beckwith, D. Fellbaum, D. Gross, and K. Miller. 1990. Five papers on Wordnet, CSL report 43. Technical report, Cognitive Science Laboratory, Princeton University.

P. Vossen. 1998. Introduction to EuroWordnet. *Computers and the Humanities, Special Issue on EuroWordNet (this volume)*.