Construcción de una base de conocimiento léxico multilíngüe de amplia cobertura: Multilingual Central Repository

Building a wide coverage multilingual lexical knowledge base: Multilingual Central Repository

Aitor Gonzalez-Agirre
Universidad del País Vasco
aitor.gonzalez-agirre@ehu.es

German Rigau
Universidad del País Vasco
german.rigau@ehu.es

Resumen

El uso de recursos semánticos de amplia cobertura y dominio general se ha convertido en una práctica común y a menudo necesaria para los sistemas actuales de Procesamiento del Lenguaje Natural (PLN). WordNet es, con mucho, el recurso semántico más utilizado en PLN. Siguiendo el éxito de WordNet, el proyecto EuroWordNet ha diseñado una infraestructura semántica multilingüe para desarrollar wordnets para un conjunto de lenguas europeas. En EuroWordNet, estos wordnets están interconectados con enlaces interlingüísticos almacenados en el índice interlingual (en inglés, *interlingual-index* o ILI). Siguiendo la arquitectura de EuroWordNet, el proyecto MEANING ha desarrollado las primeras versiones del *Multilingual Central Repository* (MCR) usando un ILI basado en WordNet 1.6. Con ello, se mantiene la compatibilidad entre los wordnets de diferentes idiomas y versiones. Esta versión del MCR integra seis versiones diferentes de la WordNet inglés (de 1.6 a 3.0) y también wordnets en castellano, catalán, euskera e italiano, junto a más de un millón de relaciones semánticas entre conceptos así como propiedades semánticas de diferentes ontologías. Recientemente hemos desarrollado una nueva versión del MCR usando un ILI basado en WordNet 3.0. Esta nueva versión del MCR integra wordnets de cinco idiomas diferentes: inglés, castellano, catalán, euskera y gallego. La versión actual del MCR, al igual que la anterior, integra sistemáticamente miles de relaciones semánticas entre conceptos. Además, el MCR se ha enriquecido con cerca de 460.000 propiedades semánticas y ontológicas que incluyen *Base Level Concepts*, *Top Ontology*, *WordNet Domains* y *AdimenSUMO*, proporcionando coherencia ontológica a todos los wordnets y recursos semánticos integrados en ella.

Palabras clave

Semántica Léxica, Bases de Conocimiento Léxico, WordNet, EuroWordNet

Abstract

The use of wide coverage and general domain semantic resources has become a common practice and often necesary by existing systems Natural Language Processing (NLP). WordNet is by far the most widely used semantic resource in NLP. Following the success of WordNet, the EuroWordNet project has designed a multilingual semantic infrastructure to develop wordnets for a set of European languages. In EuroWordNet, these wordnets are interconnected with links stored in the Inter-Lingual Index (ILI). Following the EuroWordNet architecture, the MEANING project has developed the first versions of Multilingual Central Repository (MCR) using WordNet 1.6 as ILI. Thus, maintaining the compatibility between wordnets of different languages and versions. This version of the MCR integrates six different versions of the English WordNet (1.6 to 3.0) and wordnets in Spanish, Catalan, Basque and Italian, along with more than a million semantic relationships between concepts and semantic properties different ontologies. We recently developed a new version of MCR using WordNet 3.0 as ILI. This new version of the MCR integrates wordnets of five different languages: English, Spanish, Catalan, Basque and Galician. The current version of MCR, like the previous one, systematically integrates thousands of semantic relations between concepts. In addition, the MCR is enriched with about 460,000 semantic and ontological properties including *Base Level Concepts*, *Top Ontology*, *WordNet Domains* and *AdimenSUMO*, providing all ontological consistency the integrated semantic wordnets and resources on it.

Keywords

Lexical Semantics, Lexical Knowledge Bases, WordNet, EuroWordNet

1 Introducción

A pesar del progreso realizado en los últimos años en el área del Procesamiento del Lenguaje Natural (PLN), aún estamos lejos de comprender automáticamente textos en lenguaje natural. El uso de bases de conocimiento de amplia cobertura es una práctica común en los sistemas de PLN avanzados. Sin duda, la base de conociminto más utilizada es WordNet[1] (Fellbaum, 1998). No obstante, la construcción de bases de conocimiento con cobertura suficiente pare procesar textos de dominio general requiere de un esfuerzo enorme. Este esfuerzo sólo pueden realizarlo grandes grupos de investigación durante largos periodos de desarrollo. Por ejemplo, en el caso del WordNet desarrollado en Princeton para el inglés, en más de diez años de construcción manual (desde 1995 hasta 2006, es decir, de la versión 1.5 a la 3.0) creció de 103.445 a 235.402 relaciones semánticas[2], lo que representa un crecimiento de aproximadamente mil nuevas relaciones por mes. Sin embargo, en 2008, el grupo de Princeton distribuyó un nuevo recurso con 458.825 palabras de las definiciones de WordNet, manualmente anotadas con el correspondiente sentido de WordNet[3]. Afortunadamente, en los ultimos años, la comunidad investigadora ha desarrollado un amplio conjunto de recursos semánticos de amplia cobertura vinculados a distantas versiones de WordNet. A lo largo de los últimos años, muchos de estos recursos han sido integrados en el *Multilingual Central Repository* (MCR) (Atserias et al., 2004; Gonzalez-Agirre, Laparra e Rigau, 2012a; Gonzalez-Agirre, Laparra e Rigau, 2012b). El MCR sigue el modelo propuesto por el proyecto europeo EuroWordNet[4] (LE-2 4003) (Vossen, 1998). EuroWorNet diseñó una base de la datos lexical multilingüe con wordnets de varios idiomas europeos, estructuras de forma analoga al WordNet inglés. La versión actual del MCR es el resultado del Proyecto Europeo MEANING[5] (IST-2001-34460) (Rigau et al., 2002), así como de los proyectos KNOW[6] (TIN2006-15049-C03), KNOW2[7] (TIN2009-14715-C04) y de varias acciones complementarias asociadas al proyecto KNOW2.

El artículo está estructurado como sigue: En la sección 2 realizamos un repaso de las bases de conocimiento léxico existentes, introduciendo también la primera versión del *Multilingual Central Repository* (MCR). A continuación la sección 3 presenta la última versión del MCR y el *Web EuroWordNet Interface* (WEI), incluyendo una detallada descripción de la estructura de la base de satos empleada para implementar el MCR. Por último, en la sección 4 redactamos algunas conclusiones, y marcamos el camino para trabajos futuros.

2 Bases de Conocimiento Léxicas

Este sección proporciona una revisión de las bases de conocimiento léxico para el Procesamiento de Lenguaje Natural (PLN). La sección esta dividida en tres partes. Primero, el apartado 2.1 revisa los conceptos más importantes relacionados con las tareas de PLN, y el uso de los *recursos semánticos* de amplia cobertura. El siguiente apartado presenta las principales metodologías, estrategias y técnicas empleadas para la *construcción manual de recursos semánticos de gran tamaño* (apartado 2.2). Finalmente, el apartado 2.3 presenta la primera versión del Multilingual Central Repository (MCR).

2.1 Conocimiento Léxico y PLN

En el contexto del Procesamiento del Lenguaje Natural (PLN), la semántica estudia el significado, y en concreto se centra en la relación entre significantes, tales como las palabras, las frases, los signos y símbolos. En particular, la *Semántica Léxica* estudia el significado individual de las palabras y sus relaciones. La semántica léxica también estudia como está organizado el léxico y como el significado léxico esta interrelacionado. Su mayor objetivo es estructurar un modelo de léxico a través de la categorización de tipos de relación entre palabras. La semántica léxica se centra en el estudio de las unidades léxicas. Las unidades léxicas son los elementos básicos de un léxico (el vocabulario) y pueden ser consideradas como la unidad mínima de significado.

2.2 Construcción manual de Bases de Conocimiento

La tarea de Procesamiento de Lenguaje Natural (PLN) requiere de enormes bases de conocimiento semántico como respaldo de procesos semánticos intensos. Por ello, en los últimos años el desarrollo de recursos léxicos y semánticos de amplia cobertura ha sido un objetivo prioritario de investigación.

[1] `http://wordnet.princeton.edu`
[2] Las relaciones simétricas sólo se contabilizan una vez.
[3] `http://wordnet.princeton.edu/glosstag.shtml`
[4] `http://www.illc.uva.nl/EuroWordNet`
[5] `http://nlp.lsi.upc.edu/projectes/meaning`
[6] `http://ixa.si.ehu.es/know`
[7] `http://ixa.si.ehu.es/know2`

La construcción de estas bases de conocimiento requiere del esfuerzo de grandes grupos de investigación a lo largo de periodos de desarrollo prolongados. Sin embargo, estas bases de conocimiento, aún hoy en día, no parecen ser lo suficientemente ricos como para se empleados directamente en aplicaciones semánticas avanzadas. Parece que las aplicaciones de PLN no podrán mejorar sin la incorporación de conocimiento más detallado, rico y de propósito general.

Es más, todos los idiomas encapsulan el conocimiento de modos distintos. Esta variación entre idiomas es uno de los problemas principales que impide el uso extendido de las tecnologías de PLN. Una de las soluciones propuestas es la de adoptar una representación conceptual común que factorice la variación dentro de un idioma y también con el resto de los idiomas.

La necesidad de grandes bases de conocimiento semántico se puede vislumbrar observando la cantidad de proyectos que actualmente construyen recursos de este tipo. Proyectos como WordNet[8] (Fellbaum, 1998), FrameNet[9] (Baker, Fillmore e Lowe, 1998), VerbNet[10] (Kipper et al., 2006), SUMO[11] (Niles e Pease, 2001) o Cyc[12] (Lenat, 1995) han dedicado décadas y miles de horas de trabajo a la construcción manual de estos recursos de conocimiento semántico. Por desgracia, la construcción manual de estos recursos limita de forma severa su cobertura y escala.

Por ejemplo, la gran mayoría de ontologías formales se han desarrollado para dominios particulares[13]. Las ontologías son representaciones formales de un conjunto de conceptos dentro de un dominio, y de la relaciones entre dichos conceptos, normalmente incluyendo una taxonomía y un conjunto de relaciones semánticas. Las ontologías suelen ser empleadas para razonar sobre las propiedades de dichos dominios, y también pueden ser usadas para definir el dominio en sí (Álvez, Lucio e Rigau, 2012).

2.2.1 WordNet

WordNet[14] (Miller et al., 1991; Fellbaum, 1998) es una base de conocimiento léxica para el idioma inglés. Esta inspirada por teorías psicolingüisticas y computacionales sobre la memoria léxica humana. Contiene información codificada manualmente sobre nombres, verbos, adjetivos y adverbios del inglés, y esta organizada entorno a la noción de *synset*. Un synset es un conjunto de palabras de la misma categoría morfosintáctica que pueden ser intercambiados en un contexto dado. Por ejemplo, $\langle student, pupil, educatee \rangle$ forman un synset porque pueden ser utilizados para referirse al mismo concepto. Un synset es comúnmente descrito por una *gloss* o definición, que en el caso del synset anterior es "*a learner who is enrolled in an educational institution*", y además, por un conjunto explícito de relaciones semánticas con otros synsets. Cada synset representa un concepto que está relacionado con otros conceptos mediante una gran variedad de relaciones semánticas, incluyendo hiperonimia/hiponimia, meronimia/holonimia, antonimia, etc. Los synsets están enlazados entre ellos mediante relaciones léxicas y semántico-conceptuales. WordNet también codifica 26 tipos diferentes de relaciones semánticas. WordNet está disponible de modo público y gratuito para su descarga. La versión actual de WordNet es la 3.1. Su estructura lo convierte en una herramienta útil para la lingüística computacional y el procesamiento de lenguaje natural. Resulta evidente que WordNet se ha convertido en un estándar en el PLN. De hecho, WordNet es usado en todo el mundo como base para anclar distintos tipos de conocimiento semántico, incluyendo wordnets de otros idiomas (Vossen, 1998), conocimiento de dominios (Magnini e Cavaglià, 2000) u ontologías como la Top Ontology (Álvez et al., 2008) o AdimenSUMO (Álvez, Lucio e Rigau, 2012).

WordNet ha sido creado y está siendo mantenido por el *Cognitive Science Laboratory* de la Universidad de Princeton inicialmente bajo la dirección del profesor George A. Miller y actualmente por la profesora Christiane D. Fellbaum. Su desarrollo comenzó en 1985. A lo largo de los años el proyecto ha recibido financiación de diferentes agencias del gobierno americano. WordNet ha sido empleado en una amplia variedad de tareas de PLN, tales como *Information Extraction* (Stevenson e Greenwood, 2006), *Automatic Summaritzation* (Chaves, 2001), *Question Answering* (Moldovan e Rus, 2001), *Lexical Expansion* (Parapar, Barreiro e Losada, 2005), etc.

La difusión y el éxito de WordNet ha provocado la aparición de multitud de proyectos con el objetivo de construir wordnets para otros idiomas, tomando como referencia la versión inglesa. Por ejemplo, catalán (Benítez et al., 1998), castellano (Atserias et al., 1997), euskera (Agirre et

[8] http://wordnet.princeton.edu

[9] http://framenet.icsi.berkeley.edu

[10] http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

[11] http://www.ontologyportal.org/

[12] http://www.cyc.com

[13] http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library

[14] http://wordnet.princeton.edu/

al., 2002), árabe (Rodríguez et al., 2008), etc. [15]

Algunos esfuerzos se han centrado en el desarrollo de wordnets multilingües, como EuroWordNet[16] (Vossen, 1998), MultiWordNet[17] (Pianta, Bentivogli e Girardi, 2002), Balkanet (Stamou et al., 2002b) o más recientemente, el WordNet Asiático[18] (Sornlertlamvanich, Charoenporn e Isahara, 2010), o wordnets de dominios particulares, como EuroTerm (Stamou et al., 2002a) o JurWordNet[19] (Sagri, Tiscornia e Bertagna, 2004).

La *Global WordNet Association*[20] es una organización sin fines lucrativos que proporciona un marco para establecer contactos, compartir y discutir sobre los wordnets que se desarrollan en todos los idiomas del mundo.

2.2.2 EuroWordNet

El proyecto EuroWordNet[21] (Vossen, 1998) diseñó una arquitectura completa para el desarrollo de una base de conocimiento multilingüe que incluyera varios wordnets de idiomas europeos (entre ellos, holandés, italiano, castellano, alemán, francés, checo y estonio). En EuroWordNet, cada WordNet representa un único sistema interno de lexicalizaciones siguiendo la estructura del wordnet inglés. Los wordnets de los distintos idiomas están ligados mediante el *Inter-Lingual Index* (abreviado como ILI). Estas conexiones permiten acceder a palabras similares en cualquiera de los idiomas integrados en el arquitectura EuroWordNet. Además, el ILI dá acceso a una ontología lingüística compuesta por 63 relaciones semánticas distintas. Esta ontología proporciona una categorización común para todos los idiomas, mientras que las distinciones específicas de cada idioma están en cada uno de los wordnets locales.

Aunque el proyecto EuroWordNet se concluyó en el verano de 1999, muchos de sus principios siguen aún vigentes. Por ejemplo, el diseño de la arquitectura multilingüe, los *Base Concepts*, las relaciones, la ontología, etc. se han seguido usando por grupos de investigación que están desarrollando wordnets en otros idiomas (como por ejemplo, el castellano, el euskera, el catalán y el gallego) usando buena parte de la especificación de EuroWordNet. Si los wordnets son compatibles con la especificación, pueden ser añadidos a una base de datos común, y mediante el ILI, ser conectados con otros wordnets, permitiendo el uso aplicaciones multilingües de lenguaje natural.

2.2.3 Base Concepts

The noción de los *Base Concepts* (a partir de ahora BC) fue introducida en EuroWordNet. Se supone que los BC son conceptos que juegan un papel importante en los diversos wordnets de diferentes idiomas. Este rol puede ser definido mediante dos criterios principales:

- Una posición elevada en la jerarquía semántica.
- Tener muchas relaciones con otros conceptos.

Por lo tanto, los BC son los bloques fundamentales para el establecimiento de relaciones en un wordnet y dar información acerca de los patrones dominantes de lexicalización en los idiomas. De este modo, los *Lexicografic Files* (o Supersentidos) de WordNet pueden ser considerados como el conjunto más básico de BC. Siguiendo estos criterios, en EuroWordNet se seleccionó un conjunto de BC para que se alcanzara un máximo de cobertura y compatibilidad durante el desarrollo de los wordnets de los distintos idiomas. Inicialmente, se seleccionó un conjunto de 1.024 *Common Base Concepts* extraídos de WordNet 1.5 (conceptos que actúan como BC en al menos dos idiomas), considerando solamente los wordnets en inglés, holandés, español e italiano.

Los *Basic Level Concepts* (Rosch e Lloyd, 1978) (a partir de ahora BLC) son el resultado de un compromiso entre dos principios de caracterización opuestos:

- Representar tantos conceptos como sea posible.
- Representar tantas características como sea posible.

Así, los BLC típicamente deberían ocurrir en niveles de abstracción medios, es decir, en posiciones intermedias de las jerarquías. Con esta idea en mente, diseñamos un algoritmo que utiliza propiedades estructurales básicas de cualquier versión de WordNet para obtener un conjunto completo de BLC que represente a todos sus sustantivos y verbos (Izquierdo, Suárez e Rigau, 2007)[22]. Para seleccionar los BLCs de forma

[15] Una lista de wordnets actualmente en desarrollo puede encontrarse en `http://www.globalwordnet.org/gwa/wordnet_table.html`

[16] `http://www.illc.uva.nl/EuroWordNet/`

[17] `http://multiwordnet.fbk.eu/`

[18] `http://www.asianwordnet.org`

[19] `http://www.ittig.cnr.it/Ricerca/materiali/JurWordNet/JurWordNetEng.htm`

[20] `http://www.globalwordnet.org`

[21] `http://www.illc.uva.nl/EuroWordNet/`

[22] `http://adimen.si.ehu.es/web/BLC`

automática, el programa calcula el número total de relaciones del *synset* o el número de relaciones de hiponímia y descarta los BLCs que no representan al menos un número determinado de *synsets* descendientes. Estos BLCs automáticos se han utilizado para *Word Sense Disambiguation* basada en clases semánticas (Izquierdo, Suárez e Rigau, 2009; Izquierdo, Suárez e Rigau, 2010) y para facilitar la conexión de WordNet con la ontología del proyecto KYOTO (Laparra, Rigau e Vossen, 2012).

2.2.4 Top Ontology

Para maximizar un desarrollo uniforme y consistente de los wordnets, el proyecto EuroWordNet categorizó los *Base Concepts* usando la *Top Ontology*, que fue diseñado específicamente para este propósito. La **Top Ontology**[23] (Rodríguez et al., 1998) esta basada en clasificaciones lingüísticas ya existentes y adaptada para representar la diversidad de los *Base Concepts*. Es importante tener en cuenta que los *Top Concepts* representan características semánticas que puede ser aplicadas de forma conjuntiva. Por ejemplo, es posible obtener grupos complejos de caracteristicas, como *Container+Living+Part+Solid*, que puede ser aplicado, por ejemplo, a un "vaso sanguíneo".

El primer nivel de la *Top Ontology* está dividido en tres tipos:

- 1stOrderEntity (corresponde a objetos y sustancias concretas y perceptibles)
- 2ndOrderEntity (estados, situaciones y eventos)
- 3rdOrderEntitiy (entidades mentales como las ideas, conceptos y conocimientos)

Así, la *Top Ontology* de EuroWordNet está organizada mediante 63 características que pueden ser combinadas. La ontología esta especialmente diseñada para ayudar en la codificación de las relaciones léxico-semánticas en WordNet. Sin embargo, durante el proyecto EuroWordNet sólo se pudieron caracterizar con etiquetas de la TO los Base Concepts (BC) seleccionados en el proyecto.

Muchas de las subdivisiones de la TO son disjuntas. Por ejemplo, un concepto no puede ser a la vez *Natural* y *Artifact*. Explotando estas incompatibilidades entre las características de la TO podemos localizar inconsistencias ontológicas en la jerarquía de WordNet. Para ello, simplemente debemos heredar las características de la TO asignadas a los BC a través de la jerarquía de hiponimia de WordNet. Para evitar la herencia de categorías incompatibles podemos incluir algunos puntos de bloqueo en la jerarquía de hiponimia. De esta forma, hemos desarrollado un conjunto de herramientas para el control de la consistencia de la anotación y obtener su expansión. Para demostrar la consistencia de la anotación, hemos comprobado que no hay incompatibilidad en la anotación de la parte nominal de WordNet 1.6 cuando se utilizan los puntos de bloqueo. La expansión de la anotación se puede obtener cuando la anotación es consistente. Siguiendo este proceso hemos obtenido una anotación consistente de la parte nominal de WorNet[24] (Álvez et al., 2008).

2.2.5 WordNet Domains

Uno de los problemas de WordNet es su nivel de granularidad. Hay conceptos cuyas diferencias a nivel semántico son virtualmente indetectables. **WordNet Domains**[25] (WND) (Magnini et al., 2002) es un recurso léxico desarrollado en el ITC-IRST por (Magnini e Cavaglià, 2000) donde los synsets han sido anotados de un modo semi-automático con una o más etiquetas de dominio, escogidas de un conjunto de 165 etiquetas organizadas jerárquicamente. Los usos de WND incluyen el poder de reducir el nivel de polisemia de las palabras y agrupar aquellos sentidos que pertenecen al mismo dominio. Por ejemplo, para la palabra *bank* (banco, en inglés), siete de sus diez sentidos en WordNet no comparten dominio, reduciendo de este modo la polisemia. Además, un dominio puede incluir *synsets* de diferentes categorías morfosintácticas. Por ejemplo, *MEDICINE* puede contener sentidos de nombres y de verbos. Un dominio también puede incluir sentidos de diferentes sub-jerarquias de WordNet. Por ejemplo *SPORTS* tiene conceptos subclase de *lifeform*, *physical-object*, *act*, *location*, etc. Sin embargo, la construcción de WND ha seguido un proceso semi-automático y aunque su anotación ha sido revisada (Bentivogli et al., 2004), aún podemos encontrar fácilmente muchas inconsistencias en su anotación (Castillo, Real e Rigau, 2004). Es por ello que se han propuesto y desarrollado métodos más robustos para la asignación de etiquetas de dominio a través de WordNet (González, Rigau e Castillo, 2012; Gonzalez-Agirre, Castillo e Rigau, 2012).

[23] `http://www.illc.uva.nl/EuroWordNet/corebcs/ewnTopOntology.html`

[24] `http://adimen.si.ehu.es/web/WordNet2TO`

[25] `http://wndomains.fbk.eu/`

2.2.6 SUMO y AdimenSUMO

SUMO[26] (Niles e Pease, 2001) fue creado por el *IEEE Standard Upper Ontology Working Group*. Su objetivo era desarrollar una ontología estándard de alto nivel para promover el intercambio de datos, la búsqueda y extracción de información, la inferencia automática y el procesamiento del lenguaje natural. SUMO provee definiciones para términos de propósito general resultantes de fusionar diferentes ontologías libres de alto nivel (ej. la ontología de alto nivel de Sowa, axiomas temporales de Allen, mereotopología formal de Guarino, etc.).

SUMO consiste en un conjunto de conceptos, relaciones y axiomas que formalizan una ontología de alto nivel. Una ontología de alto nivel está limitada a conceptos que son meta, genéricos o abstractos. Por tanto, estos conceptos son suficientemente genéricos como para caracterizar un amplio rango de dominios. Aquellos conceptos que son de dominios especificos o particulares no están incluidos en una ontología de alto nivel.

SUMO está organizada en tres niveles. La parte superior y la parte central consisten en aproximadamente 1.000 términos y 4.000 axiomas, dependiendo de la versión. El tercer nivel contiene ontologías de dominio. En total, cuando todas las ontologías de dominio son combinadas, SUMO consiste en aproximadamente 20.000 términos y cerca de 70.000 axiomas.

Además, los desarrolladores de SUMO han creado un enlace completo a WordNet (Niles e Pease, 2003).

AdimenSUMO[27] (Álvez, Lucio e Rigau, 2012) es una reconversión de SUMO a una ontología de primera orden operativa. Así, AdimenSUMO puede ser utilizada para el razonamiento formal por demostradores de teoremas de lógicas de primer orden (como E-prover o Vampire). Al estar también enlazado a WordNet, AdimenSUMO se convierte en una herramienta muy potente para realizar razonamiento avanzado. Por ejemplo, utilizando demostradores de teoremas avanzados, es fácil inferir de AdimenSUMO que ninguna planta tiene cerebro (ni otras partes de animal).

2.3 Multilingual Central Repository

Uno de los principales resultados del proyecto MEANING[28] fue el desarrollo de la primera versión del Multilingual Central Repository (MCR)[29] (Atserias et al., 2004) para mantener la compatibilidad entre wordnets de distintos idiomas y versiones, tanto nuevos como anteriores, así como el nuevo conocimiento que se fuera adquiriendo.

Todo el diseño del MCR sigue la arquitectura propuesta por EuroWordNet. Esta arquitectura hace posible desarrollar wordnets locales de forma relativamente independiente, garantizando al mismo tiempo un alto nivel de compatibilidad. Esta estructura multilingüe permite transportar el conocimiento de un wordnet al resto de wordnets a través del ILI (*Inter-Lingual Index*), manteniendo la compatibilidad entre todos ellos. De esta forma, la estructura del ILI (incluyendo la *Top Ontology* (Vossen et al., 1997), *Wordnet Domains* (Magnini e Cavaglià, 2000) y la ontología *SUMO* (Niles e Pease, 2001)) actúa como la columna vertebral que permite transferir el conocimiento adquirido de cada uno de los wordnets locales al resto. Del mismo modo, los diferentes recursos (ej. las diferentes ontologías) están relacionadas mediante el ILI, y en consecuencia también pueden ser validados entre ellos (Álvez et al., 2008; Álvez, Lucio e Rigau, 2012).

El MCR sólo incluye conocimiento conceptual. Esto significa que tan solo las relaciones semánticas entre *synsets* pueden ser integradas y transportadas entre los diferentes wordnets. Aún así, cuando sea necesario, las relaciones adquiridas pueden mantenerse sub-especificadas. En ese sentido, pueden integrarse y transportarse a otros idiomas o procesos. Por ejemplo, la relación *<gain> involved <money>* capturada como un objeto-directo típico, más tarde puede detallarse como *<gain> involved-patient <money>* y ser portada al wordnet en castellano como *<ganar> involved-patient <dinero>*.

La versión del MCR desarrollada en el marco del proyecto MEANING contiene seis versiones distintas del WordNet inglés (Fellbaum, 1998) (de la 1.5 a la 3.0) junto con más de un millon de relaciones semánticas entre *synsets* adquiridas de WordNet, eXtended WordNet (Mihalcea e Moldovan, 2001), y preferencias de selección adquiridas de *SemCor* (Agirre e Martínez, 2001; Agirre e Martínez, 2002) y del *British National Corpus* (BNC) (McCarthy, 2001). Esta versión del MCR también incluye wordnets del castellano (Atserias et al., 1997), italiano (Bentivogli, Pianta e Girardi, 2002), euskera (Agirre et al., 2002) y catalán (Benítez et al., 1998). Esta versión usa un ILI basado en WordNet 1.6.

Como estos recursos han sido desarrollados usando diferentes versiones de WordNet (de la

[26] `http://www.ontologyportal.org`
[27] `http://adimen.si.ehu.es/web/AdimenSUMO`
[28] `http://nlp.lsi.upc.edu/projectes/meaning`
[29] `http://adimen.si.ehu.es/web/MCR`

1.5 a la 3.0), hemos tenido que aplicar una tecnología que alineara los wordnets automáticamente **WordNet Mappings**[30] (Daudé, 2005). Esta tecnología proporciona enlaces entre synsets de diferentes versiones de WordNets, manteniendo la compatibilidad de todos los recursos que usan una determinada versión de WordNet. Además, esta tecnología permite realizar el transporte de todo el conocimiento asociado a una version de WordNet al resto de versiones.

Al término del proyecto MEANING, el MCR ha continuado su desarrollo y mejora en los proyectos nacionales KNOW[31] y KNOW2[32], así como varias acciones complementarias, con especial énfasis en los idiomas inglés, castellano, catalán, euskera y gallego.

La versión actual del MCR integra, siguiendo la arquitectura EuroWordNet, wordnets de cinco idiomas diferentes: inglés, castellano, catalán, euskera y gallego. El *Inter-Lingual-Index* (ILI) permite la conectividad entre las palabras en un idiomas con las traducciones equivalentes en cualquiera de las otras lenguas gracias a los enlaces generados automáticamente. El ILI actual corresponde a la versión 3.0 de WordNet.

Por ello, el MCR constituye un recurso multilingüe de amplia cobertura que puede ser de gran utilidad para un gran número de procesos semánticos que requieren de conocimientos lingüístico-semánticos ricos y complejos (por ejemplo, ontologías para la web semántica). Así, el MCR está siendo utilizado en múltiples proyectos y desarrollos. Por ejemplo, los proyectos europeos KYOTO[33], PATHS[34], OpeNER[35] y NewsReader[36], y el proyecto nacional SKaTer[37].

2.3.1 MCR usando ILI 1.6

La versión del MCR que usa un ILI basado en WordNet 1.6 tiene los siguientes componentes:

- ILI (versión WordNet 1.6):
  - WordNet 1.6 (Fellbaum, 1998)
  - Base Concepts (Izquierdo, Suárez e Rigau, 2007)
  - Top Ontology (Álvez et al., 2008)
  - WordNet Domains (Bentivogli et al., 2004)
  - AdimenSUMO (Álvez, Lucio e Rigau, 2012)
- Wordnets locales:
  - WordNet inglés: versiones 1.5, 1.6, 1.7.1, 2.0, 2.1, 3.0 (Fellbaum, 1998)
  - wordnets castellano y catalán (Benítez et al., 1998), italiano (Bentivogli, Pianta e Girardi, 2002) y euskera (Agirre et al., 2002).
  - eXtended WordNet (Mihalcea e Moldovan, 2001)
- Preferencias semánticas:
  - Adquiridas de SemCor (Agirre e Martínez, 2002)
  - Adquiridas del BNC (McCarthy, 2001)
- Instancias
  - Entidades nombradas (Alfonseca e Manandhar, 2002)

Inicialmente, la mayor parte del conocimiento que pretendíamos integrar en el MCR estaba alineado a WordNet 1.6, el WordNet italiano o de *MultiWordNet Domains*, éstos últimos desarrollados usando un ILI basado en WordNet 1.6 (Bentivogli, Pianta e Girardi, 2002; Magnini e Cavaglià, 2000). Por tanto, el MCR usando un ILI basado en WordNet 1.6 minimizaba efectos secundarios con otras iniciativas europeas (proyectos Balkanet, EuroTerm, etc.) y otros wordnets desarrollados alrededor de la *Global WordNet Association*. Sin embargo, el ILI para los wordnets del castellano, catalán y euskera era el WordNet 1.5 (Atserias et al., 1997; Benítez et al., 1998), así como la *Top Ontology* de EuroWordNet y los *Base Concepts* asociados. Por tanto, éstos últimos recursos debieron transportarse a la versión WordNet 1.6 (Atserias, Villarejo e Rigau, 2003). Además, la versión final del MCR con ILI basado en WordNet 1.6 terminada al final del proyecto KNOW2 contiene versiones mejoradas de *Base Concepts* (Izquierdo, Suárez e Rigau, 2007), *Top Ontology* (Álvez et al., 2008), *WordNet Domains* (Bentivogli et al., 2004) y *AdimenSUMO* (Álvez, Lucio e Rigau, 2012). Sin embargo, muchos de sus componentes tenían licencias restrictivas y no podían distribuirse de forma libre e integrada con el resto del MCR[38]. Por ello, y para actualizar los recursos existentes decidimos actualizar el ILI a WordNet 3.0.

[30] `http://www.talp.upc.edu/index.php/technology/resources/multilingual-lexicons-and-machine-translation-resources/multilingual-lexicons/98-wordnet-mappings`

[31] `http://ixa2.si.ehu.es/know`

[32] `http://ixa2.si.ehu.es/know2`

[33] `http://www.kyoto-project.eu`

[34] `http://www.paths-project.eu`

[35] `http://www.opener-project.org/`

[36] `http://www.newsreader-project.eu/`

[37] `http://nlp.lsi.upc.edu/skater`

[38] Esta versión puede consultarse en `http://adimen.si.ehu.es/cgi-bin/wei1.6/public/wei.consult.perl`

3 Multilingual Central Repository 3.0

La versión actual del MCR usa un ILI basado en WordNet 3.0 e integra, siguiendo el modelo propuesto por EuroWordNet y MEANING, wordnets de cinco idiomas distintos, incluyendo el inglés, castellano, catalán, euskera y gallego. Como en la versión anterior, los wordnets están conectados a través del *Inter-Lingual-Index* (ILI) permitiendo conectar palabras de un idioma con las palabras equivalentes en los otros idiomas también integrados en el MCR. Como la versión actual del ILI del MCR es la correspondiente a la versión 3.0 del WordNet en inglés, la mayoría del conocimiento ontológico ha sido transportado desde las versiones anteriores al nuevo MCR 3.0. Así, la mayoría de los recursos transportados han tenido que ser alineados a la nueva versión. La descripción completa del proceso empleado para llevar a cabo el transporte y actualización de todos los recursos se puede consultar en (Gonzalez-Agirre, Laparra e Rigau, 2012b; Gonzalez-Agirre, Laparra e Rigau, 2012a).

Además, para poder interactuar con el MCR y actualizar su contenido, hemos actualizado el *Web EuroWordNet Interface* (WEI), una nueva interfaz web para navegar y editar el MCR 3.0.

3.1 Web EuroWordNet Interface

El *Web EuroWordNet Interface* (WEI) (Benítez et al., 1998) permite consultar y editar la información contenida en el MCR. WEI usa technología CGI, lo que significa que todos los datos se procesan sólo en el servidor y los usuarios trabajan con clientes ligeros con capacidad de navegación web HTML. Todos los datos se almacenan en una base de datos MySQL. La interfaz se ha ido actualizando desde su desarrollo inicial en el proyecto EuroWordNet.

WEI permite navegar, consultar y editar la información asociada a un item (que puede ser un *synset*, una palabra, un *variant* o un ILI) de uno de los wordnets integrados en el MCR. La aplicación WEI consulta la información correspondiente a ese item y la información ontológica asociada a los ILIs correspondientes. La aplicación también permite consultar por los *synsets* relacionados del propio wordnet origen de la consulta (usando las relaciones codificadas en el MCR de hiperonimia, hiponimia, meronimia, etc.), o de algún otro wordnet integrado en el MCR (a través del ILI).

La aplicación consta de dos marcos. En el marco superior introducimos los parámetros para la búsqueda, y en inferior nos muestra los resultados de la consulta. Los diferentes parámetros de búsqueda son los siguientes:

- **Item:** el item que pretendemos buscar, que puede ser una palabra, un *synset*, un *variant* o un ILI.
- **Tipo de item:** el tipo del item que pretendemos buscar (palabra, *variant*, *synset* o ILI).
- **PoS:** la categoría gramatical del item (nombres, verbos, adjetivos o adverbios).
- **Relación:** que se carga dinámicamente desde la base de datos (sinónimos, hipónimos, hiperónimos, etc.)
- **WordNet origen:** el wordnet desde donde realizamos la consulta.
- **WordNet navegación:** el wordnet al cual seguimos las relaciones.
- **Glossa:** si está seleccionado se muestran las glosas de los *synsets*.
- **Score:** si está seleccionado se muestran los valores de confianza.
- **Rels:** si está seleccionado muestra información acerca de las relaciones que el *synset* tiene en todos los wordnets seleccionados.
- **Full:** si está seleccionado realiza una búsqueda transitiva por todas las relaciones.
- **WordNets mostrados:** wordnets seleccionados.

WEI también permite editar el contenido del MCR. Su funcionamiento es exactamente igual a la consulta, pero en modo edición, tanto los *synsets* como los ILIs pueden seleccionarse y editarse. Al editar un *synset* nos aparece una pantalla de donde podemos añadir, eliminar o modificar los *variants* del *synset*, modificar su glosa y ejemplos, así como las relaciones que tiene con otros *synsets* (en la Figura 1 se puede ver un ejemplo para el sentido 1 de "*party*"). Al editar un ILI nos aparece una pantalla donde podremos añadir, eliminar o modificar la información ontológica asociada al ILI.

3.1.1 Marcas para synsets y variants

En la nueva versión del WEI es posible asignar propiedades especiales o *marcas* a *variants* y a *synsets*. También podemos añadir una pequeña nota o comentario que especifica mejor por qué hemos asignado una marca determinada. Uno de los objetivos de las marcas es permitir una edición más rápida mediante WEI, siendo

Figura 1: Captura de pantalla de la interfaz de edición, mostrando el sentido 1 de "party".

capaces de marcar y anotar un synset, facilitando una posterior revisión y corrección. Otra ventaja es aumentar la cantidad de información almacenada en cada *variant* y *synset*.

Las marcas disponibles son las siguientes:

- Marcas de *variant*:
  - DUBLEX: Para aquellos *variants* con una lexicalización dudosa.
  - INFL: Indica que un *variant* es **flexivo**. Necesario para el wordnet en euskera.
  - RARE: *Variant* muy poco usado u en desuso.
  - SUBCAT: Subcategorización. Se usa para aquellos *variant* que deben tener subcategorización.
  - VULG: Para aquellos *variant* que son vulgares, rudos u ofensivos.
- Marcas de *synset*:
  - GENLEX: Conceptos generales no lexicalizados que son introducidos para organizar mejor la jerarquía.
  - HYPLEX: Indica que el hiperónimo tiene idéntica lexicalización.
  - SPECLEX: Termimos especificos de ciertos dominios, y que deben ser comprobados.

Anotar marcas usando el WEI es muy simple. Tan solo hay que buscar un synset o variant y anotarlo usando la interfaz de edición, seleccionando las marcas deseadas. En una sola ventana se pueden editar las marcas y comentarios de todos los variant contenidos en un synset (incluyendo más de uno, o incluso todos, a la vez), y la marca y el comentario del propio synset.

3.2 Diseño de la Base de Datos para el MCR

Actualmente, el MCR está almacenado en una base de datos relacional consistente de 40 tablas[39]. Este apartado describe cada una de las tablas necesarias para que el MCR funcione correctamente.

La tabla principal del MCR es la tabla que contiene el *Inter-Lingual Index* (ILI):

- **wei_ili_record**: Contiene los identificadores de los ILI, en formato ili-30-xxxxxxx-y (las x indican el offset del *synset*, y las y representan la categoria gramatical o *Part-Of-Speech* (PoS)). Cada registro también almacena el origen del ILI (el WordNet del que proviene), si es un *Base Concept* o no, el fichero lexicográfico, y si es una instancia o no.

Cada uno de los idiomas incluidos en el MCR (incluyendo el inglés) esta ligado al ILI, y compuesto por cinco tablas. Cada idioma tiene su propio código de 3 letras, que está indicado por *xxx* en la lista siguiente:

[39] La distribución actual ha sido probada sobre MySQL y PostgreSQL.

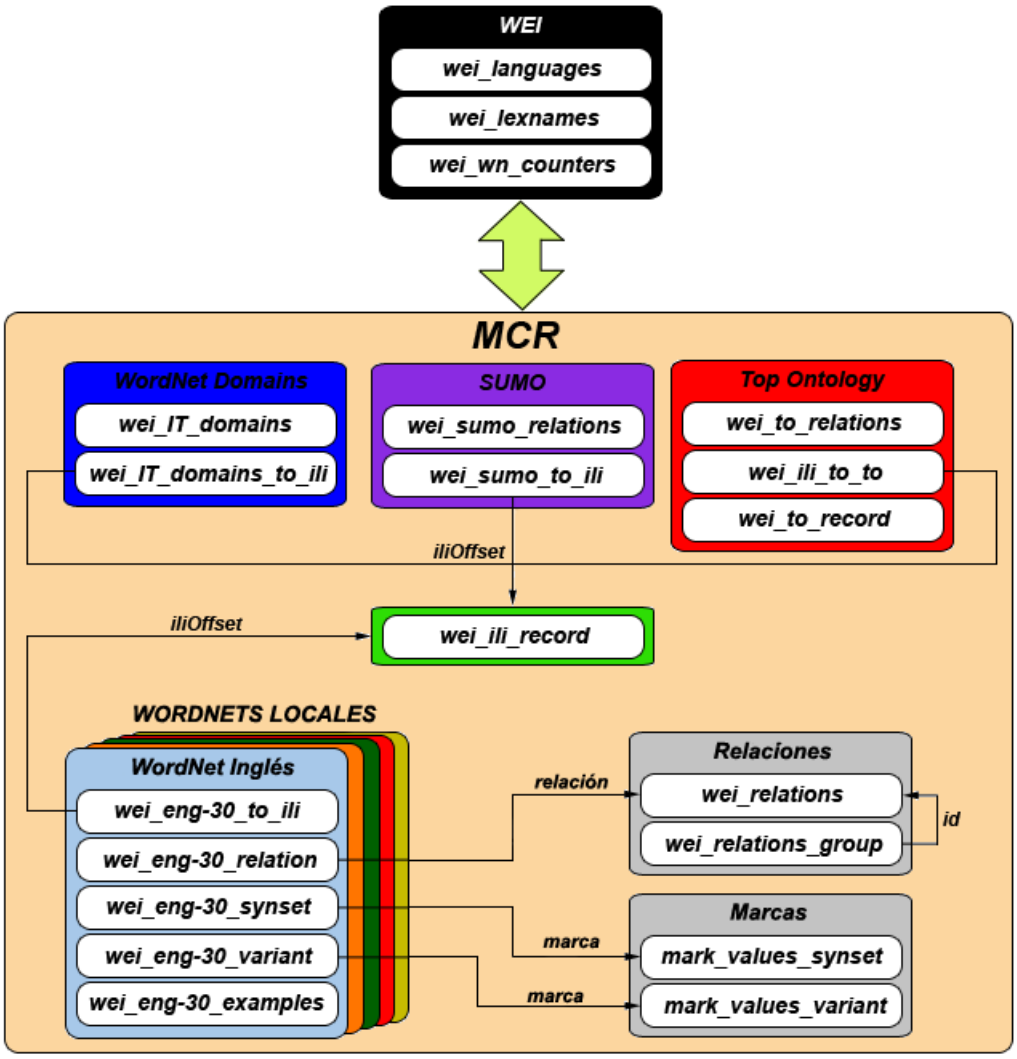

Figura 2: Estructura de la Base de Datos para el MCR y el WEI.

- **wei_xxx-30_to_ili**: Esta tabla conecta el ILI (por ejemplo, ili-30-00001740-a) con el número de *synset* (por ejemplo, eng-30-00001740-a).
- **wei_xxx-30_relation**: Esta tabla contiene todas las relaciones del wordnet. Cada registro (que es la instancia de una relación) guarda un código que indica el tipo de relación (que se almacena de la tabla *wei_relations*), la dirección de la relación (*synset* origen y *synset* destino), el valor de confianza y el wordnet del que proviene.
- **wei_xxx-30_synset**: Aquí se almacena la información acerca del *synset*: el identificador, el número de descendientes, la glosa, el nivel en el que se encuentra (contando desde arriba), y finalmente una marca (opcional) y un comentario del *synset* (opcional).
- **wei_xxx-30_variant**: Aquí se almacenan todos los *variant* del wordnet. Cada registro representa un único *variant* y contiene la siguiente información: el *variant*, el sentido de la palabra, el identificador de *synset*, un valor de confianza, el experimento del que proviene (opcional), y finalmente la marca (opcional) y el comentario del *variant* (opcional).
- **wei_xxx-30_examples**: En esta tabla se listan todos los ejemplos del wordnet. Cada ejemplo está identificado por el número de *synset*, la palabra y el sentido.

Añadir un nuevo idioma es tan fácil como crear las cinco nuevas tablas con el patrón anterior y un código de 3 letras que lo representa.

Además de los wordnets, el MCR integra otros recursos (dominios, ontologías, marcas, etc.). Las tablas que contienen esta información son las siguientes:

Dominios:

- **wei_domains**: Esta tabla representa la jerarquía de *WordNet Domains* usando tuplas origen-destino.
- **wei_ili_to_domains**: Cada registro enlaza un dominio a un ILI. También se indica el wordnet del que proviene. Esta tabla es única, haciendo que la información de dominios esté compartida entre todos los wordnets.

AdimenSUMO:

- **wei_sumo_relations**: Esta tabla representa la jerarquía de *AdimenSUMO* usando tuplas origen-destino. También incluye un campo que indica si se trata de una sub-clase o no.
- **wei_ili_to_sumo**: Cada registro enlaza una etiqueta de *AdimenSUMO* a un ILI. También se indica el wordnet del que proviene. Al igual que la tabla de dominios, esta tabla es única, haciendo que la información de dominios esté compartida entre todos los wordnets.

Top Ontology:

- **wei_to_relations**: Esta tabla representa la jerarquía de *Top Ontology* usando tuplas origen-destino. También incluye un campo que indica el tipo de la relación.
- **wei_ili_to_to**: Cada registro enlaza una etiqueta de *Top Ontology* a un ILI. También se indica el wordnet del que proviene. Al igual que la tabla de dominios y la de *AdimenSUMO*, esta tabla es única, haciendo que la información de dominios esté compartida entre todos los idiomas.
- **wei_to_record**: Esta tabla almacena, para cada etiqueta de *Top Ontology*, la glosa asociada a ella.

Marcas:

- **mark_values_synset**: Valores permitidos para las marcas de *synset*, asi como su descripción.
- **mark_values_variant**: Valores permitidos para las marcas de *variant*, asi como su descripción.

El resto de tablas incluidas en el MCR son las siguientes:

- **wei_relations**: Esta tabla contiene todas las relaciones posibles en el MCR. Cada relación tiene un identificador, un nombre, sus propiedades y una nota (opcional). También se indica si es inversa (en caso de que sea posible) y a que grupo de relaciones pertenece (ver más abajo). El código ID que aparece en esta tabla es el que está reflejado en las tabla *wei_xxx-30_relation*. Esta tabla es la que permite realizar búsquedas mediante el WEI.
- **wei_relations_group**: Aquí se almacenan los super-grupos de relaciones (sinónimos, hiperónimos, merónimos, causa, etc.). El código ID que aparece en esta tabla es el que está reflejado en la tabla *wei_relations*.
- **wei_languages**: Los wordnets disponibles en el MCR. Para cada wordnet se indica el código, el nombre y el color con el que debe de aparecer en el WEI.
- **wei_lexnames**: Aquí se almacenan los ficheros lexicográficos de WordNet. Cada entrada tiene un código (el indicado en la tabla *wei_ili_record*) y un nombre descriptivo.
- **wei_wn_counters**: La interfaz del WEI permite la creación de nuevos *synset*. Para evitar solapamientos y problemas futuros, cada PoS tiene su propio número de *offset*, empezando desde el 800.000. Esta tabla guarda los números que deben adoptar los nuevos *synsets* que se creen en cada categoría gramatical.

La figura 2 muestra la estructura completa del MCR.

3.3 Estado actual del MCR

En este apartado se presenta el estado actual del MCR, incluyendo el progreso respecto al WordNet inglés. La tabla 1 muestra la cantidad actual de *synsets* y *variants*, el número de glosas y el número de ejemplos de cada wordnet, distinguiendo entre las distintas categorías gramaticales o PoS.

Como ejemplo de la información contenida en el MCR podemos analizar el sentido 4 de "*party*", que se muestra en la figura 3. En la columna de la izquierda podemos ver el ILI (*ili-30-07447641-n*), y debajo de él, la información asociada al ILI:

- **WordNet Domains:** *free_time, sociology.*
- **Fichero semántico de WordNet:** *event.*
- **AdimenSUMO:** *Meeting.*
- **Top Ontology:** *Agentive, BoundedEvent, Communication, Purpose, Social.*

En la siguiente columna se muestra los *synsets* asociados de cada wordnet y los *variants* que hay

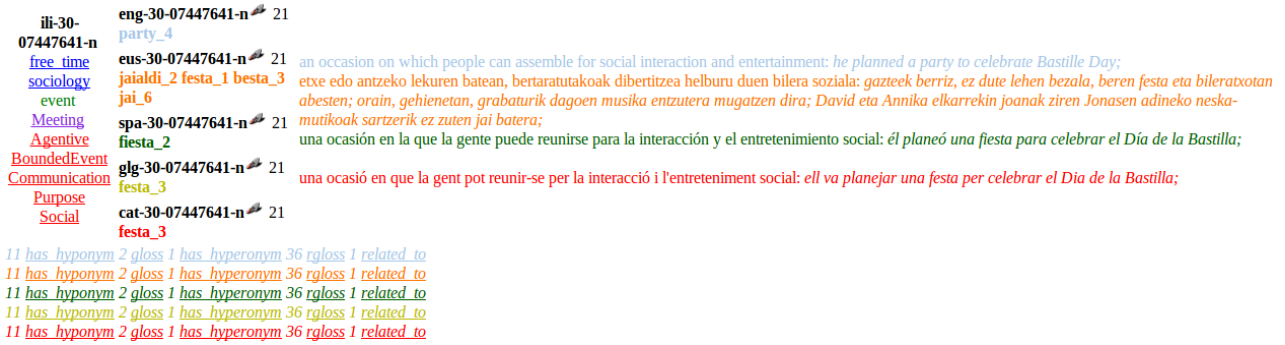

Figura 3: Información almacenada en el MCR para el sentido 4 de "party".

en cada uno de ellos, indicando el **variant** y el número de **sentido** de dicha palabra (el número final después del simbolo "_"):

- **Inglés:** party (4).
- **Euskera:** jaialdi (2), festa (1), besta (3), jai (6).
- **Castellano:** fiesta (2).
- **Gallego:** festa (3).
- **Catalán:** festa (3).

En la columna de la derecha están las glosas o definiciones para cada uno de los wordnets (en este caso, el gallego no tiene esta información), así como ejemplos de uso de las palabras (que se muestran en cursiva).

Por último, en la parte inferior se enumeran las relaciones de estos synset (que en este caso coinciden):

- **has_hyponym:** Indica que el synset tiene 11 hipónimos.
- **gloss:** Indica que 2 de las palabras que aparecen en la glossa estan desambiguadas, y ligadas al synset correspondiente. En este caso son *entertainment* e *interaction*, que podemos ver en la glossa en inglés.
- **has_hyperonym:** Indica que el synset tiene un hiperónimo.
- **rgloss:** *Reverse gloss*. Indica que este *synset* aparece desambiguado en 36 glosas de otros *synets*.
- **related_to:** Indica qué otros *synset* está relacionados con este. En este caso es el sentido 1 de "*party*" en inglés, definido como *have or participate in a party*, o el sentido 4 de "festejar" en castellano.

4 Conclusiones y Trabajo Futuro

Como hemos visto, la construcción de bases de conocimiento con cobertura suficiente pare procesar textos de dominio general requiere de un esfuerzo enorme. Este esfuerzo sólo pueden realizarlo grandes grupos de investigación durante largos periodos de desarrollo. Afortunadamente, en los ultimos años, la comunidad investigadora ha desarrollado un amplio conjunto de recursos semánticos de amplia cobertura vinculados a distantas versiones de WordNet. A lo largo de los últimos años, muchos de estos recursos han sido integrados en el *Multilingual Central Repository* (MCR). Primero usando un ILI basado en WordNet 1.6. Esta versión, aunque demostró el potencial de la propuesta contenía recursos con licencias restrictivas que impedían su distribución integrada. La versión que usa el ILI basado en WordNet 3.0 no contiene recursos que tengan licencias restrictivas y puede distribuirse de forma integrada[40].

Como vemos, integrar todos estos recursos en una única infraestructura también es una tarea compleja que requiere de un esfuerzo continuado. Por un lado, continuamente aparecen nuevos recursos potencialmente interesantes, y por otro, los antiguos recursos se siguen actualizando.

En el marco del proyecto SKaTer planeamos seguir enriqueciendo el MCR con nuevos recursos. Entre otros, los ya integrados en la versión del MCR con el ILI basado en WordNet 1.6. Por ejemplo, el resto de versiones de WordNet inglés (1.5, 1.6, 1.7, 1.7.1, 2.0 y 2.1, ya que resulta muy práctico poder consultar todas las versiones de WordNet simultáneamente). También pensamos recuperar y integrar al nuevo ILI las preferencias de selección adquiridas de SemCor, así como in-

[40] La mayor parte de los recursos integrados en esta versión tiene licencia Creative Commons Attribution 3.0 Unported (CC BY 3.0) `http://creativecommons.org/licenses/by/3.0`

Variants	**Nombres**	**Verbos**	**Adjetivos**	**Adverbios**	**Synsets**	**%WN**
EngWN3.0	147.360	25.051	30.004	5.580	118.431	100 %
SpaWN3.0	39.142	10.824	6.967	1.051	38.702	33 %
CatWN3.0	51.605	11.577	7.679	2	46.033	39 %
EusWN3.0	40.939	9.470	148	0	30.615	26 %
GalWN3.0	18.949	1.416	6.773	0	19.312	16 %
Glosas						
EngWN3.0	82.379	13.767	18.156	3.621	117.923	100 %
SpaWN3.0	12.533	3.325	1.917	670	18.445	16 %
CatWN3.0	6.294	44	840	1	7.179	6 %
EusWN3.0	2.690	2	0	0	2.692	2 %
GalWN3.0	4.997	2	3.111	0	8.111	7 %
Ejemplos						
EngWN3.0	10.433	11.583	15.615	3.674	41.305	100 %
SpaWN3.0	465	30	195	193	606	2 %
CatWN3.0	2.105	46	368	0	2.201	5 %
EusWN3.0	2.376	0	0	0	2.075	5 %
GalWN3.0	270	2	4.291	0	2.416	6 %

Cuadro 1: Número actual de *variants*, *synsets*, definiciones y ejemplos de cada wordnet.

formación sobre predicados verbales y nominales, tanto de VerbNet (Kipper et al., 2006), como de FrameNet (Baker, Fillmore e Lowe, 1998; Laparra, Rigau e Cuadros, 2010). Entre nuestras prioridades también está la integración y explotación de recursos para el análisis de sentimiento. Por ejemplo, Q-WordNet[41] (Agerri e García-Serrano, 2010).

Agradecimentos

Este trabajo ha sido posible gracias al apoyo de los proyectos europeos MEANING (IST-2001-34460), KYOTO (ICT-2007-211423), OpeNER (ICT-2011-296451) y NewsReader (ICT-2011-316404), así como a los nacionales KNOW (TIN2006-15049-C03), KNOW2 (TIN2009-14715-C04-04), SKaTer (TIN2012-38584-C06-01), y varias acciones complementarias asociadas al proyecto KNOW2. Aitor GonzalezAgirre también recibe el apoyo del Ministerio Español de Educación, Cultura y Deporte a través de una beca pre-doctoral FPU (FPU12/06243).

Bibliografía

Agerri, Rodrigo e Ana García-Serrano. 2010. Q-wordnet: Extracting polarity from wordnet senses. Em *Seventh Conference on International Language Resources and Evaluation, Malta (retrieved May 2010)*.

Agirre, Eneko, Olatz Ansa, Xabier Arregi, José Mª Arriola, Arantza Diaz de Ilarraza, E. Pociello, e L. Uria. 2002. Methodological issues in the building of the basque wordnet: quantitative and qualitative analysis. Em *Proceedings of the first International Conference of Global WordNet Association*, Mysore, India.

Agirre, Eneko e David Martínez. 2001. Learning class-to-class selectional preferences. Em *Proceedings of CoNLL01*, Toulouse, France.

Agirre, Eneko e David Martínez. 2002. Integrating selectional preferences in wordnet. Em *Proceedings of the 1st International Conference of Global WordNet Association*, Mysore, India.

Alfonseca, Enrique e S. Manandhar. 2002. Distinguishing concepts and instances in wordnet. Em *Proceedings of the first International Conference of Global WordNet Association*, Mysore, India, 21-25 January, 2002.

Atserias, Jordi, Salvador Climent, Javier Farreres, German Rigau, e Horacio Rodríguez. 1997. Combining multiple methods for the automatic construction of multilingual wordnets. Em *Proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP'97)*, Tzigov Chark, Bulgaria.

Atserias, Jordi, Luís Villarejo, e German Rigau. 2003. Integrating and porting knowledge across languages. Em *Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'03)*,

[41] http://www.rodrigoagerri.net/sentiment-analysis

pp. 31–37, Borovets, Bulgaria, September, 2003.

Atserias, Jordi, Luís Villarejo, German Rigau, Eneko Agirre, John Carroll, Piek Vossen, e Bernardo Magnini. 2004. The meaning multilingual central respository. Em *Proceedings of the Second International Global WordNet Conference (GWC'04)*.

Baker, Collin F., Charles J. Fillmore, e John B. Lowe. 1998. The berkeley framenet project. Em *Proceedings of the COLING-ACL*, pp. 86–90.

Bentivogli, Luisa, Pamela Forner, Bernardo Magnini, e Emanuele Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. Em *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pp. 101–108. Association for Computational Linguistics.

Bentivogli, Luisa, E. Pianta, e C. Girardi. 2002. Multiwordnet: developing an aligned multilingual database. Em *Proceedings of the First International Conference on Global WordNet*, Mysore, India.

Benítez, Laura, Sergi Cervell, Gerard Escudero, Mónica López, German Rigau, e Mariona Taulé. 1998. Methods and tools for building the catalan wordnet. Em *Proceedings of ELRA Workshop on Language Resources for European Minority Languages*, Granada, Spain.

Castillo, Mauro, Francis Real, e German Rigau. 2004. Automatic assignment of domain labels to wordnet. Em *Proceeding of the 2nd International WordNet Conference*, pp. 75–82.

Chaves, R. P. 2001. Wordnet and automated text summarization. Em *Proceedings of 6th Natural Language Processing Pacific Rim Symposium NLPRS 2001*, pp. 109–116, Tokyo, Japan, Jan, 2001.

Daudé, Jordi. 2005. *Enlace de Jerarquías Usando el Etiquetado por Relajación*. Tese de doutoramento, Universitat Politècnica de Catalunya, July, 2005.

Fellbaum, Christiane. 1998. *WordNet. An Electronic Lexical Database*. Language, Speech, and Communication. The MIT Press.

González, Aitor, German Rigau, e Mauro Castillo. 2012. A graph-based method to improve wordnet domains. Em *Proceedings of the Computational Linguistics and Intelligent Text Processing (CICLING'12)*, pp. 17–28. Springer.

Gonzalez-Agirre, A., E. Laparra, e G. Rigau. 2012a. Multilingual central repository version 3.0. Em *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pp. 2525–2529.

Gonzalez-Agirre, A., E. Laparra, e G. Rigau. 2012b. Multilingual central repository version 3.0: upgrading a very large lexical knowledge base. Em *Proceedings of the 6th Global WordNet Conference (GWC'12)*.

Gonzalez-Agirre, Aitor, Mauro Castillo, e German Rigau. 2012. A proposal for improving wordnet domains. Em Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, e Stelios Piperidis, editores, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may, 2012. European Language Resources Association (ELRA).

Izquierdo, R., A. Suárez, e G. Rigau. 2007. Exploring the automatic selection of basic level concepts. Em *Proceedings of RANLP*.

Izquierdo, Rubén, Armando Suárez, e German Rigau. 2009. An empirical study on class-based word sense disambiguation. Em *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 389–397. Association for Computational Linguistics.

Izquierdo, Rubén, Armando Suárez, e German Rigau. 2010. Gplsi-ixa: Using semantic classes to acquire monosemous training examples from domain texts. Em *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 402–406, Uppsala, Sweden, July, 2010. Association for Computational Linguistics.

Kipper, Karin, Anna Korhonen, Neville Ryant, e Martha Palmer. 2006. Extending verbnet with novel verb classes. Em *Proceedings of LREC*, volume 2006, pp. 1.

Laparra, Egoitz, German Rigau, e Montse Cuadros. 2010. Exploring the integration of wordnet and framenet. Em *Proceedings of the 5th Global WordNet Conference (GWC'10)*, Mumbai (India), January, 2010.

Laparra, Egoitz, German Rigau, e Piek Vossen. 2012. Mapping wordnet to the kyoto ontology. Em Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard,

Joseph Mariani, Jan Odijk, e Stelios Piperidis, editores, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may, 2012. European Language Resources Association (ELRA).

Lenat, Douglas B. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Magnini, Bernardo e G. Cavaglià. 2000. Integrating subject field codes into wordnet. Em *Proceedings of the Second Internatgional Conference on Language Resources and Evaluation (LREC)*, Athens. Greece.

Magnini, Bernardo, C. Satrapparava, G. Pezzulo, e A. Gliozzo. 2002. The role of domains informations. Em *In Word Sense Disambiguation*, Treto, Cambridge, July, 2002.

McCarthy, Diana. 2001. *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Aternations, Subcategorization Frames and Selectional Preferences*. Tese de doutoramento, University of Sussex.

Mihalcea, Rada e Dan Moldovan. 2001. extended wordnet: Progress report. Em *Proceedings of NAACL Workshop* WordNet and Other Lexical Resources: Applications, Extensions and Customizations, pp. 95–100, Pittsburg, PA, USA.

Miller, G. A., R. Beckwith, Christiane Fellbaum, D. Gross, K. Miller, e R. Tengi. 1991. Five papers on wordnet. *Special Issue of the International Journal of Lexicography*, 3(4):235–312.

Moldovan, Dan e Vasile Rus. 2001. Logic form transformation of wordnet and its applicability to question answering. Em *In Proceedings of ACL 2001*, pp. 394–401.

Niles, I. e Adam Pease. 2001. Towards a standard upper ontology. Em Chris Welty e Barry Smith, editores, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, Maine.

Niles, I. e Adam Pease. 2003. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. Em *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, Las Vegas, Nevada.

Parapar, David, Alvaro Barreiro, e David E. Losada. 2005. Query expansion using wordnet with a logical model of information retrieval. Em *Proceedings of IADIS AC*, pp. 487–494.

Pianta, Emanuele, Luisa Bentivogli, e Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. Em *Proceedings of the First International Conference on Global WordNet*, January, 2002.

Rigau, German, Bernardo Magnini, Eneko Agirre, Piek Vossen, e John Carroll. 2002. Meaning: A roadmap to knowledge technologies. Em *Proceedings of COLING Workshop A Roadmap for Computational Linguistics*, Taipei, Taiwan.

Rodríguez, Horacio, Salvador Climent, Piek Vossen, Laura Bloksma, Wim Peters, Antonietta Alonge, Francesca Bertagna, e Adriana Roventini. 1998. The top-down strategy for building eurowordnet: Vocabulary coverage, base concepts and top ontology. *Computers and the Humanities*, 32(2-3):117–152.

Rodríguez, Horacio, David Farwell, Javier Farreres, Manuel Bertran, Musa Alkhalifa, Mª Antònia Martí, William J. Black, Sabri Elkateb, James Kirk, Adam Pease, Piek Vossen, e Christiane Fellbaum. 2008. Arabic wordnet: Current state and future extensions. Em *Proceedings of the Fourth International Global WordNet Conference - GWC 2008, Szeged, Hungary*, January, 2008.

Rosch, Eleanor e B. Lloyd. 1978. *Cognition and Categorization*. Lawrence Erlbaum Associates, Hillsdale NJ, USA.

Sagri, Maria Teresa, Daniela Tiscornia, e Francesca Bertagna. 2004. Jur-wordnet. Em *Proceedings of the Second International Global WordNet Conference (GWC'04). Panel on figurative language*, January, 2004.

Sornlertlamvanich, Virach, Thatsanee Charoenporn, e Hitoshi Isahara. 2010. Language resource management system for asian wordnet collaboration and its web service application. Em *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may, 2010. European Language Resources Association (ELRA).

Stamou, Sofia, Alexandros Ntoulas, Jeroen Hoppenbrouwers, Maximiliano Saiz-Noeda, e Dimitris Christoudoulakis. 2002a. Euroterm: Extending the eurowordnet with domain-specific terminology using an expand model approach. Em *Proceedings of the 1st Global WordNet Association conference*, Mysore, India.

Stamou, Sofia, Kemal Oflazer, Karel Pala, Dimitris Christoudoulakis, Dan Cristea, Dan Tufis,

Svetla Koeva, George Totkov, Dominique Dutoit, e Maria Grigoriadou. 2002b. Balkanet: A multilingual semantic network for the balkan languages. Em *Proceedings of the 1st Global WordNet Association conference*.

Stevenson, Mark e Mark A. Greenwood. 2006. Learning Information Extraction Patterns Using WordNet. Em *Proceedings of the 5th Intl. Conf. on Language Resources and Evaluations, LREC 2006 22 - 28 May 2006*, volume 2006, pp. 95–102.

Vossen, Piek. 1998. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.

Vossen, Piek, L. Bloksma, Horacio Rodríguez, Salvador Climent, A. Roventini, F. Bertagna, e A. Alonge. 1997. The eurowordnet base concepts and top-ontology. Relatório técnico, Deliverable D017D034D036 EuroWordNet LE2-4003.

Álvez, Javier, Jordi Atserias, Jordi Carrera, Salvador Climent, Egoitz Laparra, Antoni Oliver, e German Rigau. 2008. Complete and consistent annotation of wordnet using the top concept ontology. Em *Proceedings of the the 6th Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech (Morocco), May, 2008.

Álvez, Javier, Paqui Lucio, e G. Rigau. 2012. Adimen-sumo: Reengineering an ontology for first-order reasoning. *International Journal on Semantic Web and Information Systems*, 8(4):80–116.