Actas del XXXV Simposio Internacional de la Sociedad Española de Lingüística, editadas por Milka Villayandre Llamazares, León, Universidad de León, Dpto. de Filología Hispánica y Clásica, 2006. ISBN: 84-690-3383-2. Publicación electrónica en: http://www3.unileon.es/dp/dfh/SEL/actas.htm

EL PROCESO DE ANOTACIÓN SEMÁNTICA EN *FRAMENET ESPAÑOL*

ROCÍO DONÉS ROJAS
CRISTINA ORTIZ RODRÍGUEZ
Universidad Autónoma de Barcelona

1. ENFOQUE TEÓRICO

La semántica de marcos empezó a desarrollarse en 1968, en pleno auge de la gramática generativa que, como la gramática estructural, parte de la hipótesis de que "existen reglas o estructuras determinadas con independencia del léxico que permiten dar una caracterización general de las secuencias de palabras que constituyen los enunciados lingüísticos de las lenguas naturales" (Subirats 2001:27), y acabó definiéndose, básicamente, a partir del modelo propuesto por Fillmore (1985). Ésta es una de las líneas de estudio más importantes de la lingüística cognitiva, que intenta describir conceptualmente cómo se organiza nuestro conocimiento del mundo, y que es la pieza clave de la semántica cognitiva que se adopta en *FrameNet español*.

Desde el punto de vista de la sintaxis, aspecto gramatical necesario para dar sustento a la línea semántica desarrollada en el proyecto, debe destacarse la teoría de los predicados de Harris (1991), ya que trabaja la relación de dependencia entre predicados y argumentos partiendo de un procedimiento sistemático que nos permite determinar las palabras que integran el léxico de una lengua. De este modo, puede constatarse que "no todas las secuencias finitas

de palabras construidas sobre dicho léxico constituyen enunciados aceptables para los hablantes, aunque sean correctos" (Subirats 2001:27).

A partir del proceso de cambio que sufre la historia de la gramática y de la teoría de los predicados de Harris, se ha ido configurando el marco teórico y metodológico de la sintaxis léxica de Subirats (2001), "cuyo objetivo básico es el estudio de las restricciones léxicamente condicionadas, que estructuran formal y semánticamente los enunciados que determinan la forma de vehicular la información en las lenguas naturales" (Subirats 2001:27).

Este proyecto, por lo tanto, puede enmarcarse dentro del punto de vista semántico de la gramática del español que, por un lado, sigue la línea teórica de la semántica de marcos de Fillmore y, por el otro, la de la sintaxis léxica de Subirats. Principalmente, se utiliza la sintaxis para caracterizar formalmente lo más importante de la anotación: los constituyentes que vehiculan los argumentos semánticos de los predicados. De este modo, podemos afirmar que *FrameNet español* parte de la caracterización conceptual de los predicados del léxico, pero que necesita del estudio de sus proyecciones sintácticas porque son las que reflejan la estructura conceptual de las unidades léxicas. Así pues, se antepone la semántica a la sintaxis.

El objetivo principal de *FrameNet español* es, en definitiva, estudiar la organización conceptual de la red de clases semánticas o escenarios conceptuales que configura el léxico de predicados de la lengua española (Subirats 2004:1).

2. PROPÓSITOS DEL PROYECTO

Existen tres propósitos que sustentan el objetivo principal del proyecto (Subirats 2004:1).

1º) El primer propósito es identificar los marcos semánticos que configuran conceptualmente el léxico de predicados del español. Para esta identificación se procede de lo más amplio a lo más concreto: se delimitan los marcos generales, dentro de ellos se establecen los marcos concretos o escenarios conceptuales, y se definen, por último, los predicados o unidades léxicas (*Lexical Units*) que se incluyen en los anteriores. Por ejemplo, el marco general de *Emoción* (*Emotion*) puede dividirse en varios escenarios

conceptuales, tales como *Cambio del estado emocional* (*Change_emotional_state*) o *Experimentador objetivo* (*Experiencer_object*), y estos, a su vez, incorporan predicados como *sorprenderse* o *aterrar*, respectivamente; el marco general de *Comunicación* (*Communication*) puede dividirse en varios escenarios conceptuales, tales como *Pregunta* (*Questioning*) o *Discusión* (*Discussion*), y estos, a su vez, incorporan predicados como *preguntar* o *discutir*, respectivamente; el marco general de *Movimiento* (*Motion*) puede dividirse en varios escenarios conceptuales, tales como *Cambio de postura* (*Change posture*) o *Llegada* (*Arriving*), y estos, a su vez, incorporan predicados como *agacharse* o *llegar*, respectivamente; o el marco general de *Cognición* (*Cognition*) puede dividirse en varios escenarios conceptuales, tales como *Llegar a saber* (*Coming_to_believe*) o *Memorización* (*Memorization*), y estos, a su vez, incorporan predicados como *concluir* o *memorizar*, respectivamente.

En un principio, se comenzó a analizar predicados verbales, y en el desarrollo del proyecto, se fueron incluyendo también nombres predicativos, adjetivos predicativos, locuciones predicativas y, probablemente, grupos preposicionales predicativos.

2º) El segundo propósito es determinar los argumentos semánticos o conceptuales (*Frame Elements*) de cada marco. Por ejemplo, en un escenario semántico o conceptual como el de *Cambio de postura* (*Change posture*), encontramos, entre otros, los argumentos semánticos siguientes: *Protagonista* (*Protagonist*), *Dirección* (*Direction*) o *Tiempo* (*Time*).

3º) El tercer propósito es anotar semántica y sintácticamente las construcciones en las que aparecen predicados pertenecientes a los diferentes marcos.

Así pues, *agacharse* es un predicado que corresponde al marco general de *Movimiento*, al escenario conceptual de *Cambio de postura* y que se usa en diferentes oraciones que nosotros hemos anotado semántica y sintácticamente.

Hoy en día, es necesario procesar volúmenes cada vez mayores de información textual. Los tres propósitos del proyecto proporcionan los medios para realizar con éxito esta tarea. Para ello, ha sido necesario que usáramos herramientas informáticas de trabajo como *Xkwic*, *Gramcreator*, *FNDesktop* y una base de datos relacional llamada *Sato Tool* que nos ayudan a llevar a cabo nuestros propósitos

y, al mismo tiempo, a simplificar la ardua tarea del procesamiento de textos[1].

Estas herramientas, en primer lugar, permiten que anotemos semántica y sintácticamente oraciones previamente extraídas de forma automática de nuestro corpus textual e importadas en formato *XML* a la base de datos (Subirats 2004:1). El desarrollo de sistemas como los nuestros, principalmente de procesamiento semántico de textos, ha potenciado la creación de corpus lingüísticos con anotación semántica y sintáctica que sirven de entrenamiento para aplicaciones de anotación semántica automática (Gildea y Jurafsky 2002).

En segundo lugar, posibilitan la consulta mediante interfaz web (*Sato Tool*) del resultado de las relaciones de las diferentes intersecciones (algoritmos informáticos), ya que es posible acceder a la organización automática de la información final de la anotación semántica (Subirats 2004:1). La interfaz web de nuestro proyecto está preparada para que se realicen consultas sobre argumentos semánticos y construcciones sintácticas (gracias al *FNDesktop*) y las combinaciones de argumentos asociadas a dicho predicado que pueden darse dentro de un escenario conceptual o de toda la base de datos (gracias a la *Sato Tool*). Estas dos herramientas proporcionarán la creación de un corpus lingüístico en línea de oraciones anotadas semántica y sintácticamente que abrirá nuevas perspectivas para el análisis cognitivo de las características semánticas de los predicados del léxico español: un resultado similar al que persigue *FrameNet inglés*.

La posibilidad que ofrece la *Sato Tool* de realizar consultas cruzadas y simultáneas sobre *FrameNet español* e *inglés* "permitirá que nuestra base de datos se pueda utilizar como un diccionario semántico bilingüe en línea inglés-español y español-inglés, el cual, además de tener aplicaciones para la consulta directa por parte de los usuarios, tendrá sin duda repercusiones en el desarrollo de sistemas de traducción automática basados en el análisis cognitivo del léxico" (Subirats 2004:13).

Así, vemos que hay diferentes proyectos internacionales que persiguen el mismo objetivo de *FrameNet español* y que parten, a su vez, del originario *FrameNet inglés*. Este último es la respuesta innovadora a la necesidad de crear corpus linguísticos con anotación

[1] La explicación de las diferentes herramientas de trabajo se hará en los apartados sucesivos.

semántica. Tiene 130000 oraciones anotadas semántica y sintácticamente; 8800 predicados analizados, repartidos en 610 marcos semánticos distintos y con 4800 argumentos semánticos. Este proyecto ha puesto de manifiesto no sólo la posibilidad y el interés de realizar este tipo de investigaciones, sino su enorme impacto en la mejora de los sistemas de tratamiento automático de la información textual. En Alemania, encontramos el proyecto *Salsa*, que se divide en dos etapas, *Salsa I* y *Salsa II*, y que tiene como objetivo el desarrollo de un corpus con anotación semántica y sintáctica, basado en el análisis semántico propuesto por el proyecto inglés. Y, en Japón, con el mismo objetivo, tenemos el *FrameNet japonés*, que sigue, también, los presupuestos teóricos y metodológicos de *FrameNet inglés*.

De este modo, *FrameNet español* es la respuesta en lengua española al reto que supone en la actualidad el procesamiento automático de la información textual. Su objetivo final es la construcción de un corpus con etiquetación semántica y sintáctica que tenga aplicaciones lingüísticas e informáticas para el tratamiento automático de la información textual.

3. EL PROCESO DE ANOTACIÓN SEMÁNTICA EN *FRAMENET ESPAÑOL*

El proceso de anotación semántica y sintáctica de *FrameNet español* llevado a cabo por las lingüistas se divide básicamente en cuatro tareas fundamentales en la construcción de una red de marcos conceptuales:

- la *identificación de esquemas semánticos*
- la *subcorporación*
- la *etiquetación*
- y la *consulta vía web de los resultados obtenidos*.

3.1. *Identificación de esquemas semánticos*

La primera tarea del proceso de anotación semántica es *la identificación de esquemas semánticos*. Esta primera fase del proceso de anotación consiste en establecer los escenarios conceptuales o

semánticos que configuran la red de predicados del léxico español y fijar cuáles son los argumentos semánticos (*Frame Elements*) que configuran tales esquemas. Una vez determinado esto, se definen cuáles son aquellas unidades léxicas que podrían ser incluidas en cada escenario semántico.

Por ejemplo, si nos encontramos dentro del marco general del *Movimiento* (*Motion*), primero establecemos qué escenarios semánticos de este marco vamos a trabajar. Así, podemos estudiar el escenario conceptual de *Llegada* (*Arriving*) que pertenece al marco general del *Movimiento* y que se caracteriza por el desplazamiento de un tema hacia una meta.

Una vez decidido el escenario semántico que vamos a analizar, fijamos los diferentes argumentos semánticos que configuran conceptualmente dicho escenario. Tenemos tres tipos de argumentos conceptuales:

- Argumentos principales o nucleares (*Cores*): son aquellos argumentos semánticos específicos de un escenario conceptual determinado. Por ejemplo, en el caso del escenario semántico señalado anteriormente, *Llegada*, los argumentos semánticos nucleares son: el *tema* y la *meta*, ya que el escenario conceptual de *Llegada* señala que un tema se desplaza hacia una meta, como podemos ver ejemplificado en la oración (1):

  1) [El atleta]$_{\text{Tema}}$ llegó [a la meta]$_{\text{Meta}}$ muy cansado.

- Argumentos secundarios o periféricos (*Peripherals*): son aquellos argumentos semánticos característicos del marco general en el que trabajamos, pero que no son específicos del escenario semántico en el que se inserta un predicado. Veamos los ejemplos (2) y (3) pertenecientes al escenario conceptual de *Llegada*.

  2) [Los ponentes]$_{\text{Tema}}$ llegaron [desde Barcelona] $_{\text{Origen}}$
  3) [Los escaladores]$_{\text{Tema}}$ llegaron [rápidamente]$_{\text{Velocidad}}$ [a la cima de la montaña]$_{\text{Meta}}$ [por un camino secreto] $_{\text{Trayectoria}}$

Al pertenecer *Llegada* al marco general de *Movimiento*, ésta puede incluir argumentos semánticos periféricos como el *origen* (*desde Barcelona*), la *trayectoria* (*por un camino secreto*) o la

velocidad (*rápidamente*). Estos argumentos son característicos del marco general de *Movimiento*, pero no son específicos del escenario conceptual de *Llegada*.

- Argumentos generales (*Extra-thematics*): son aquellos elementos semánticos que pueden formar parte de cualquier escenario semántico, sea cual sea su marco general, como pueden ser la especificación del tiempo, del lugar, de la finalidad, de la manera, etc. Veamos las oraciones (4) y (5) con dos predicados pertenecientes a dos marcos semánticos generales diferentes:

    4) Los escaladores llegaron [ayer]$_{\text{Tiempo}}$ a la cima de la montaña.
    5) [Ayer]$_{\text{Tiempo}}$ hablé con la profesora de mi hijo.

Los predicados *llegar* y *hablar* que aparecen en las oraciones (4) y (5) pertenecen a marcos generales distintos: *Movimiento* y *Comunicación*; sin embargo, en ambos casos podemos encontrar una especificación del tiempo.

Asimismo, sería posible encontrar dos oraciones como (6) y (7):

6) María hablaba [muy deprisa]$_{\text{Manera}}$
7) María llegó [rápidamente]$_{\text{Velocidad}}$ al punto de encuentro.

El constituyente *muy deprisa* se podría pensar en anotarlo, al igual que en (7), con la etiqueta que especifica la velocidad, pero teniendo en cuenta que este predicado no pertenece al *Movimiento*, sino a la *Comunicación*, lo anotaríamos como una especificación de *Manera* (argumento semántico general) ya que en el escenario semántico de la *Comunicación* no resulta relevante la especificación de la velocidad.

Una vez fijados los argumentos semánticos que a priori podemos pensar que pertenecen a un determinado escenario semántico, seleccionamos las unidades léxicas que inicialmente pueden formar parte de cada escenario conceptual definido. Por ejemplo, el escenario conceptual de *Llegada* incluye inicialmente unidades léxicas como *llegar*, *llegada*, *entrar*, *avanzar*, *irrumpir*, etc.

Determinar los escenarios conceptuales, sus unidades léxicas y los posibles argumentos semánticos es la base de la identificación de los esquemas semánticos, porque los escenarios conceptuales deben ser un reflejo de las características semánticas del léxico.

Por ello, *FrameNet español* no propone esquemas conceptuales abstractos desvinculados del léxico, puesto que éstos no permiten construir un análisis conceptual del léxico ni de su organización en redes semánticas.

3.2. *Subcorporación*

La segunda tarea del proceso de anotación semántica consiste en seleccionar del corpus aquellas oraciones que contienen distintas proyecciones sintácticas que ejemplifican o vehiculan los argumentos semánticos de los predicados fijados y que integran un escenario semántico determinado. A continuación, se crean semiautomáticamente transductores de estas proyecciones en forma de expresiones regulares simplificadas (o autómatas) para llevar a cabo la extracción automática de las oraciones que las poseen y que en una tercera tarea serán anotadas semántica y sintácticamente.

El corpus de *FrameNet español* o *SFN Corpus* se ha ido construyendo, desarrollando y manteniendo desde 1989 hasta la actualidad. Se compone de 350 millones de palabras o, lo que es lo mismo, de 18 millones de oraciones pertenecientes en un 60 % a textos del español de América y en un 40% a textos del español europeo. El corpus contiene géneros y estilos (textos periodísticos, jurídicos, literarios, transcripciones del lenguaje hablado, ensayos sobre humanidades, noticias de agencia, prensa cultural...) y están divididos en cuatro subcorpus según el género que predomina en ellos: prensa española, noticias de agencia, literario y transcripciones de sesiones del Parlamento Europeo. Estos textos electrónicos están integrados en nuestra base de datos por un fichero con marcas de *XML* que especifica la procedencia, el nombre del fichero, el género textual, el título y el número de párrafo con objeto de facilitar tanto las búsquedas interactivas en las oraciones del corpus como la extracción automática de construcciones sintácticas para su posterior anotación semántica y sintáctica.

Desde el punto de vista del *lema*, podemos decir que *FrameNet español* dispone de 93000 lemas, diferenciados en 68000 lemas simples, como por ejemplo, *salpicar*, *desplome*, etc., y 25000 lemas locutivos, como por ejemplo, *bombas atómicas*, *maestros de escuela*, etc.

La consulta del corpus se consigue mediante la herramienta *Corpus Workbench* (CWB) desarrollada por el *Institut fur Maschinelle Sprachverarbeitung* de la Universidad de Stuttgart en Alemania. Uno de los módulos de los que se compone CWB es *XKWIC* (*Key Word In Context Xwindows*), la interfaz gráfica del procesador de consultas de CWB, una herramienta que está especialmente adaptada para llevar a cabo consultas interactivas en el corpus. Por ello, las consultas que realizamos habitualmente las lingüistas se llevan a cabo con este módulo, ya que permite realizar búsquedas, ordenaciones de resultados y selecciones de ejemplos de forma sencilla e intuitiva para *observar y analizar aquellas oraciones que contienen distintas proyecciones sintácticas*.

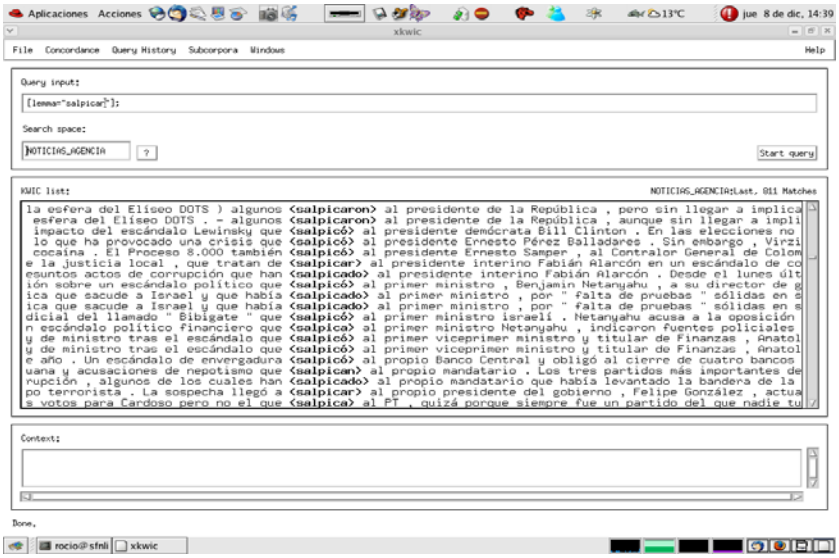

Figura 1. *Consulta con XKWIC de las oraciones del "Corpus de FrameNet español" en la que aparece el verbo "salpicar" seguido de un grupo preposicional encabezado por "a". Desde la ventana central se puede navegar entre los ejemplos y, en la parte inferior, se puede visualizar el contexto de la oración seleccionada.*

Una vez observadas las construcciones que vehiculan los significados de los predicados, se *crean transductores* en forma de expresiones regulares simplificadas con la herramienta *GramCreator*, desarrollada por los informáticos del proyecto, que permite extraer automáticamente las construcciones seleccionadas anteriormente. Por ejemplo, como vemos en la Figura 1, tenemos la expresión regular: V + *a* GN del predicado *salpicar*.

A partir del *SFN Corpus* y usando el procesador de consultas CQP (*Corpus Query Processor*), otro módulo de CWB, *se crea el subcorpus* de todas las oraciones en las que aparece el predicado (*target*).

De dicho subcorpus *se detectan*, automáticamente, las construcciones que formalizan los transductores, gracias a la aplicación *Query System Xwindows* (XQS), desarrollada también por nuestro grupo de investigación.

Finalmente, *se extraen*, automáticamente, las construcciones sintácticas del corpus que ejemplifican los argumentos semánticos y que responden al contexto sintáctico fijado en los transductores, gracias a un *software* de intersección de autómatas llamado *ALIA* y que está desarrollado por los informáticos del proyecto. Este software selecciona aleatoriamente treinta ejemplos del corpus y permite realizar una selección mucho más variada que la que se podría conseguir mediante otros procedimientos (por ejemplo, los manuales). A partir de la selección aleatoria de las oraciones extraídas automáticamente, *se recuperan las oraciones en formato XML* del CWB, el mismo de la base de datos de *FrameNet español*, a partir de las referencias que se almacenan en el autómata. Posteriormente, este archivo con las treinta oraciones seleccionadas *se importa automáticamente a nuestra base de datos*.

Por todo ello, el objetivo de esta segunda tarea es automatizar el proceso de extracción de oraciones con determinadas proyecciones sintácticas de los argumentos conceptuales de un predicado para organizar y facilitar su posterior anotación semántica (Subirats 2004:9-10).

3.3. *Etiquetación*

La tercera tarea del proceso de anotación semántica es la etiquetación o anotación semántica y sintáctica semiautomática, mediante la herramienta *FNDesktop*, de las oraciones extraídas del corpus en las que aparece un predicado en el contexto de las construcciones sintácticas que previamente se han establecido (Subirats 2004:10). Esta herramienta fue creada en el *International Computer Science Institute* (*ICSI*) en Berkeley, California y cedida a nuestro proyecto, donde se ha adaptado a las necesidades del español.

Esta tarea constituye el núcleo central del desarrollo del proyecto *FrameNet español*.

FNDesktop está dividido en dos partes (véase Figura 2) (Subirats 2004:10): la primera, un marco de navegación, donde observamos los elementos que integran la base de datos de *FrameNet español*, es decir, los escenarios semánticos, sus argumentos conceptuales y las unidades léxicas; la segunda, un marco central donde se encuentran los objetos necesarios para llevar a cabo la etiquetación, es decir, las oraciones extraídas a partir de la subcorporación y las etiquetas semánticas que responden a los diferentes argumentos semánticos.

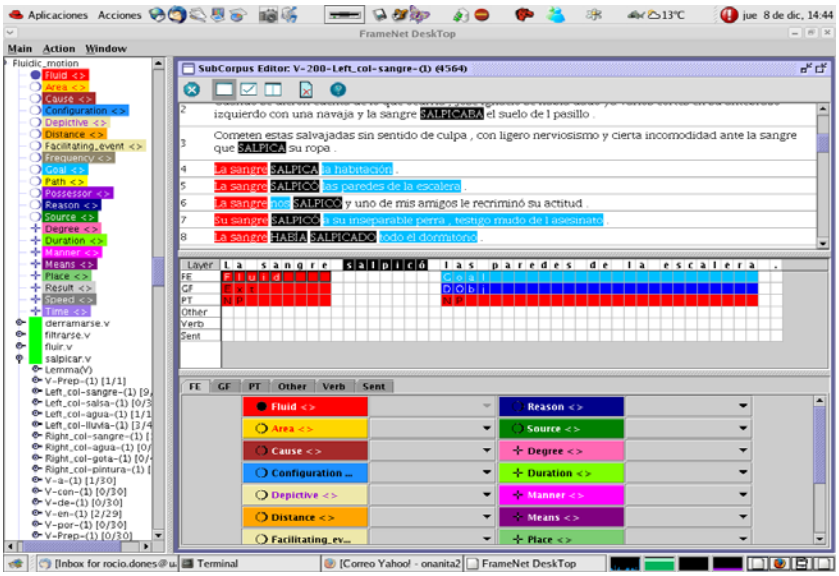

Figura 2. *Anotación del verbo predicativo "salpicar" perteneciente al escenario conceptual de predicados de "Movimiento de líquido" (Fluidic_Motion) con la herramienta FNDesktop*.

Funciones del marco de navegación (Subirats 2004:10-11):

- Se pincha con el ratón sobre un escenario semántico, como por ejemplo, *Movimiento de un líquido* (*Fluidic_motion*) y se despliegan sus argumentos semánticos junto con la lista de unidades léxicas de dicho escenario.
- Se pincha sobre un predicado, como por ejemplo *salpicar*, y "se despliega la lista de subcorpus asociados a dicho predicado y extraídos automáticamente del corpus. El nombre de los subcorpus está en relación con las características de las construcciones sintácticas que los integran" (Subirats 2004:11).

- Se pincha sobre los subcorpus y aparecen, en el marco central, las treinta oraciones de entre las que las lingüistas escogerán las que prefieran anotar.

Funciones del marco central:

- En la sección superior tenemos las oraciones. Se pincha una oración para que se despliegue en la sección intermedia del marco central.
- Empieza la fase de anotación: "se selecciona en la sección intermedia del marco el constituyente que se quiere etiquetar y, en la sección inferior, se pincha el argumento semántico con el que se quiere anotar el constituyente seleccionado" (Subirats 2004:11).
- Existen tres niveles de anotación: la especificación del argumento semántico (*Frame Element*), su función sintáctica (*Grammatical Function*) y el tipo de constituyente (*Phrase Type*). En la práctica, no es necesario especificar manualmente estos dos últimos niveles de anotación, puesto que, al seleccionar la etiqueta correspondiente a un argumento semántico, como por ejemplo, *Fluid*, *Area*, *Cause*, etc., la herramienta *FNDesktop* asigna automáticamente –con un margen de error muy bajo– su función gramatical, como por ejemplo, *External*, *Direct Object*, etc., y el tipo de constituyente, por ejemplo, *NP*, *PP*, etc. Por ello, la anotación semántica y sintáctica constituye un proceso semiautomático, ya que, al asignarle una etiqueta semántica a un constituyente, *FNDesktop* añade automáticamente la especificación de su función sintáctica y el tipo de constituyente (Subirats 2004:11).

Así pues, la sistematización práctica de las posibilidades que propone *FNDesktop* y su aplicación a todos los marcos semánticos en los que hasta ahora se está haciendo una etiquetación semiautomática, permitirá que el proceso de anotación se automatice completamente. Esto simplificará enormemente el proceso de anotación semántica y, asimismo, posibilitará que todos los ejemplos de cada subcorpus estén anotados semántica y sintácticamente. En efecto, la base de datos de oraciones de *FrameNet español* contendrá oraciones anotadas de forma totalmente automática y oraciones cuya anotación automática habrá sido supervisada por nuestro equipo de lingüistas.

Este nuevo procedimiento incrementará notablemente el número de oraciones que se podrán utilizar para los programas de entrenamiento para la etiquetación semántica automática de textos.

3.4. *Consulta vía web de los resultados de la anotación semántica*

La última tarea del proceso de anotación semántica se caracteriza por la revisión de los resultados obtenidos tras la etiquetación. Este proceso lo llevamos a cabo mediante unas aplicaciones en las que se organizan automáticamente los resultados. Estas aplicaciones son:

- *FNDesktop*
- *Web Reports*[2]
- *Sato Tool*[3]

La consulta mediante la herramienta *FNDesktop* permite acceder a los resultados de la anotación de una unidad léxica para verificar las oraciones anotadas y las proyecciones sintácticas de los argumentos semánticos que caracterizan a dicha unidad léxica sin tener que salir de la herramienta con la que anotamos las oraciones de nuestro corpus. Además, gracias a esta herramienta, se pueden verificar los datos obtenidos y también existe la posibilidad de editar las oraciones anotadas en el caso de que lo creamos pertinente. Todo ello, facilita nuestro trabajo (véase Figura 3).

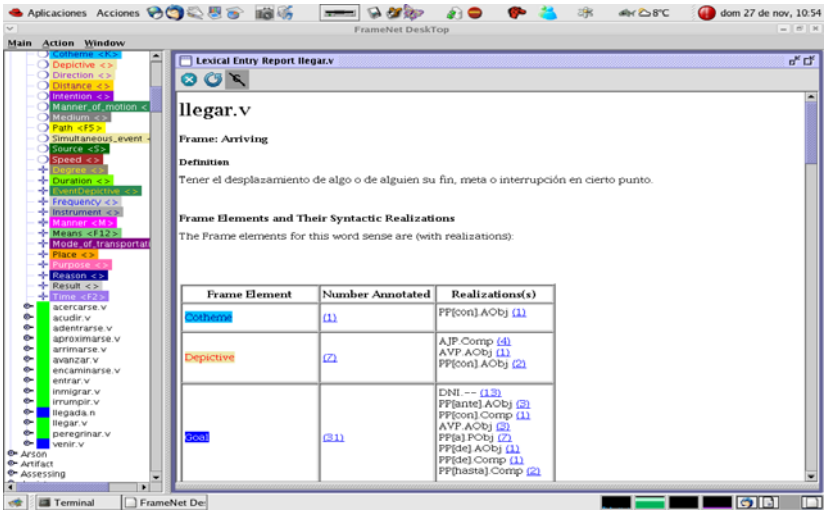

Figura 3. *Consulta de resultados mediante la herramienta FNDesktop*.

[2] URL: http://gemini.uab.es/SFN-internal/
[3] URL: http://sato.fm.senshu-u.ac.jp/sfn20/notes/index2.html

Los resultados de la anotación se organizan automáticamente en una interfaz web llamada *Web Reports*. En cuanto a esta aplicación, son varias las funciones que podemos llevar a cabo:

- Consulta de las oraciones anotadas a partir del enlace *Annotation by Lexical Unit*: en el marco de navegación de la izquierda escogemos el escenario semántico que nos interesa, por ejemplo, *Llegada*. Seleccionamos uno de los predicados que se encuentran en este escenario conceptual, por ejemplo, *llegar*, y accedemos a toda la anotación de este predicado (véase Figura 4).

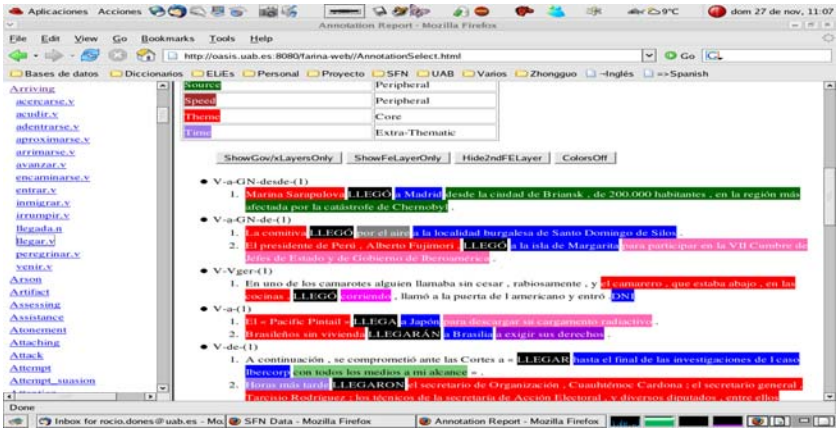

Figura 4. *Consulta de resultados mediante la aplicación Web Reports (Annotation by Lexical Unit).*

- Consulta de las proyecciones sintácticas que vehiculan los argumentos semánticos que caracterizan un predicado, a partir del enlace *Lexical Entry Report*: como en el caso anterior, escogemos el escenario semántico que nos interesa del marco de navegación de la izquierda, seleccionamos un predicado y la aplicación web nos las muestra. Esta aplicación nos ofrece un cuadro en el que podemos observar, a la izquierda, aquellos argumentos semánticos utilizados en la anotación de este predicado y, a la derecha, el tipo de constituyente y su función sintáctica. Este cuadro ofrece la posibilidad de acceder, también, a la oración u oraciones en las que podemos encontrar el argumento conceptual seleccionado (véase Figura 5).

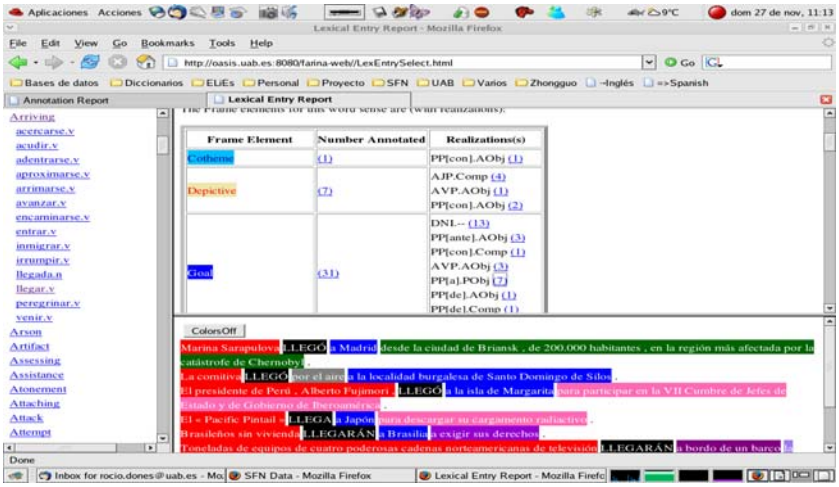

Figura 5. *Consulta de resultados de la anotación mediante la aplicación Web Reports (Lexical Entry Report).*

- Consulta de las oraciones anotadas y de las proyecciones sintácticas de los argumentos semánticos que caracterizan a cada predicado a partir del enlace *Lexical Unit Index*: esta aplicación web nos ofrece los mismos resultados que hemos presentado en los dos apartados anteriores, aunque en este caso accedemos a ellos, no a partir del escenario conceptual, sino directamente a partir de los predicados, ya que estos se encuentran ordenados alfabéticamente en el marco de navegación superior (véase Figura 6).

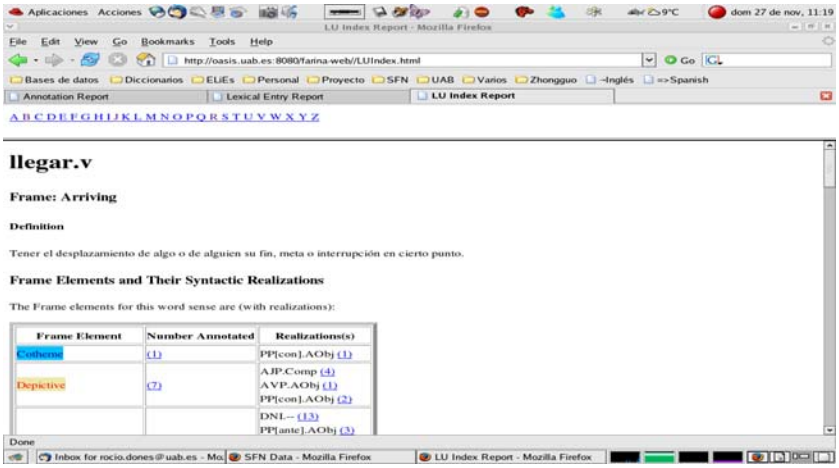

Figura 6. *Consulta de resultados de la anotación mediante la aplicación Web Reports (Lexical Entry Index).*

Otra forma que tenemos de consultar los resultados del proceso de anotación es mediante la aplicación *Sato Tool*, desarrollada por el prof. Hiroaki Sato de la Universidad de Senshu, Tokio (Japón). Esta aplicación nos permite hacer consultas más complejas que las descritas más arriba. Con esta aplicación podemos realizar búsquedas sobre argumentos semánticos y sus combinaciones y/o construcciones sintácticas, pero no sólo dentro de un solo escenario semántico como en las aplicaciones anteriores sino también en toda la base de datos. Asimismo, esta aplicación es útil, también, para las búsquedas transversales de los resultados de *FrameNet español* y *FrameNet inglés* (véase Figura 7 *infra*).

En resumen, el proyecto *FrameNet español* pretende ofrecer un mapa de la organización semántica del léxico de la lengua española. Este mapa de la red conceptual del léxico español estará acompañado, además, de un corpus de oraciones anotadas semántica y sintácticamente para ejemplificar cada uno de los predicados que conformen dicho mapa.

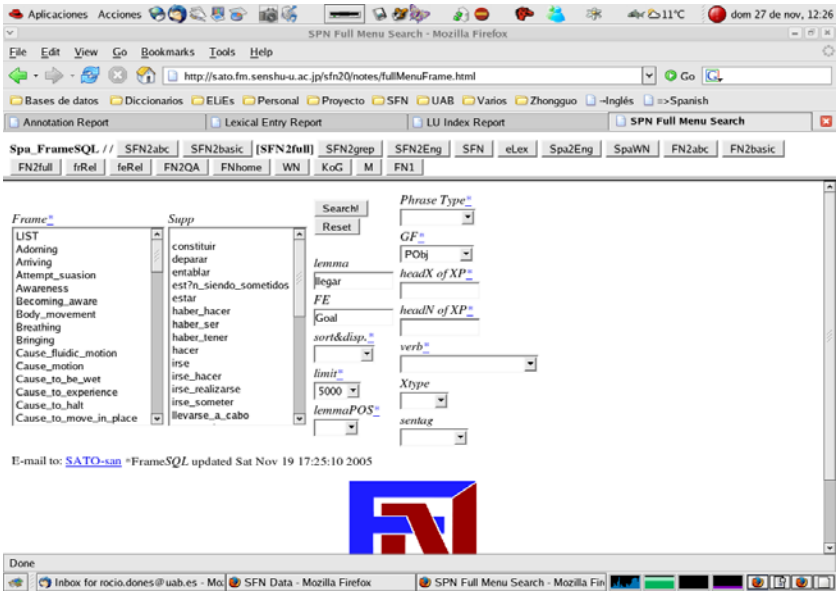

Figura 7. *Consulta de resultados de la anotación mediante la aplicación Sato Tool.*

4. DIFUSIÓN DE LOS RESULTADOS DEL PROYECTO

Los resultados de *FrameNet español* podrán visualizarse vía web de forma gratuita mediante los *Web Reports* y la *Sato Tool*, a mediados del 2006 y, más adelante, sobre el 2008, se pretende que tanto el contenido de la base de datos, es decir, de las oraciones anotadas semántica y sintácticamente, como el *software* para su gestión y consulta (*FNDesktop*), se puedan descargar libremente desde la web del proyecto tras solicitar una licencia gratuita. Los datos del proyecto se distribuirán en formato *XML*, *HTML* y *OWL*.

4.1. *Aplicaciones lingüísticas*

Gracias a la visualización de los resultados de la anotación mediante la aplicación *Web Reports* y *Sato Tool*, se dispondrá de un diccionario semántico en línea que proporcionará no sólo definiciones de cada predicado, sino también un ejemplario de oraciones anotadas que reflejarán los usos de cada predicado dentro de su escenario conceptual y unas tablas organizadas automáticamente que pondrán de manifiesto las realizaciones sintácticas de los argumentos semánticos asociados a cada predicado. Todo ello, permitirá llevar a cabo investigaciones y estudios sobre el léxico de la lengua española.

Por otra parte, gracias a la posibilidad que nos ofrece la aplicación del prof. Sato de realizar búsquedas transversales entre las base de datos de *FrameNet español* y *FrameNet inglés*, nuestro proyecto proporcionará un diccionario semántico bilingüe en línea español-inglés inglés-español que no sólo será interesante para la consulta de los usuarios sino que también puede ser de importancia significativa en el campo de la traducción automática basada en el análisis cognitivo.

4.2. *Aplicaciones informáticas*

En el campo del tratamiento informático de la lengua, *FrameNet español*, gracias al gran número de oraciones anotadas semántica y sintácticamente, servirá como corpus de entrenamiento para las futuras aplicaciones especializadas en el desarrollo de la web

semántica y en la búsqueda en bases de datos y la extracción de información textual mediante herramientas de anotación semántica automática (Gildea y Jurafsky 2002).

Por otra parte, utilizamos una aplicación en el proceso de anotación que se ha ido desarrollando en proyectos anteriores a *FrameNet español* y que en la actualidad es de dominio público. Esta aplicación es el *e-Lexicon*.

La aplicación *e-Lexicon*[4] es un diccionario electrónico en línea de formas a disposición del público de manera gratuita vía web. Esta herramienta desarrollada por nuestro equipo de investigación está integrada por una base de datos *MySQL* (el mismo formato de la *Sato Tool*) que organiza y gestiona toda la información léxica de los diccionarios y de los transductores léxicos. Esta aplicación proporciona información léxica, categorial y flexiva de todas las palabras consultadas. Por ejemplo, al consultar una forma flexiva como *hablaban*, *e-Lexicon* ofrece la siguiente información:

- el lema al que está asociado dicha forma flexiva, es decir, *hablar*;
- su categoría, es decir, un verbo predicativo (VPRED);
- y sus propiedades morfológicas flexivas, en este caso, tercera persona del plural (3P) del pretérito imperfecto del indicativo (IPIMP).

La aplicación de esta herramienta en *FrameNet español* permite la etiquetación de todos nuestros textos electrónicos y el análisis y la extracción automática de la información del corpus. Todo ello facilita el proceso de anotación de las oraciones seleccionadas.

Asimismo, nuestro grupo de investigación está desarrollando otra herramienta, el *e-Parser*, también a disposición del público de forma gratuita[5]. El *e-Parser* es un anotador sintáctico en línea que reconoce y etiqueta enunciados u oraciones independientemente de que en ellas aparezcan formas simples o formas locutivas.

En relación con *FrameNet español*, como señalamos en el apartado 3.3., esta herramienta posibilita la anotación sintáctica de

---

[4] A la que se puede acceder desde la URL: http://gemini.uab.es/elexicon/busca_dico.html

[5] A la que se puede acceder desde la URL: http://gemini.uab.es/SFN/etiquetador/index.html

forma automática de aquellos constituyentes seleccionados por el lingüista: ahora mismo esta aplicación es capaz de anotar sintácticamente constituyentes con un margen de error muy bajo.

En definitiva, mediante el estudio del mapa de la organización semántica del léxico de la lengua española, vemos que los resultados del proyecto *FrameNet español* serán de importancia significativa en el campo del tratamiento automático de la información textual de la lengua española gracias a las aplicaciones lingüísticas e informáticas comentadas.

REFERENCIAS BIBLIOGRÁFICAS

BAKER, C. F.; FILLMORE, C. y CRONIN, B. (2003): "The Structure of the FrameNet Database", *International Journal of Lexicography*, 16, 3, 281-296.

FILLMORE, C. J.; BAKER, C. F. y SATO, H. (2002): "Seeing Arguments through Transparent Structures", *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, 787-791. Publicación electrónica en: http://framenet.icsi.berkeley.edu/~framenet/papers/LREC12.pdf

FILLMORE, C. J.; BAKER, C. F. y SATO, H. (2002): "The FrameNet Database and Software Tools", *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, 1157-1160. Publicación electrónica en: http://framenet.icsi.berkeley.edu/~framenet/papers/demo4.pdf

FILLMORE, C. J. (1985): "Frames and the semantics of understanding", *Quaderni di Semántica*, 6, 2, 222-254.

GILDEA, D. y JURAFSKY, D. (2002): "Automatic Labeling of Semantic Roles", *Computational Linguistics*, Vol. 28, 3, 245-288.

HARRIS, Z. (1991): *A Theory of Language and Information. A Mathematical Approach*, Oxford: Clarendon Press.

ORTEGA, M. (2002): *Transductores en el análisis léxico y sintáctico de un texto*, tesis de licenciatura, Universidad Politécnica de Cataluña.

RUPPENHOFER, J. *et al*. (2005): *FrameNet II: Extended Theory and Practice*, Publicación electrónica en: http://www.icsi.berkeley.edu/~josef/book.html

SUBIRATS, C. (2001): *Introducción a la sintaxis léxica del español*, Madrid/Frankfurt: Iberoamericana/Vervuert.

SUBIRATS, C. (2004): "FrameNet Español. Una red semántica de marcos conceptuales", *VI International Congress of Hispanic Linguistics (October 2003)*, Germany, Leipzig University. Publicación electrónica en: http://gemini.uab.es:9080/SFNsite/papers/sfn-papers

SUBIRATS, C. y PETRUCK, M. R. L. (2003): "Surprise: Spanish FrameNet!", *Workshop on Frame Semantics, International Congress of Linguists (July 29, 2003)*, Praga, Publicación electrónica en: http://framenet.icsi.berkeley.edu/~framenet/papers/SFNsurprise.pdf

SUBIRATS, C. y SATO, H. (2004): "Spanish FrameNet and FrameSQL", *4th International Conference on Language Resources and Evaluation (LREC), Workshop on Building Lexical Resources from Semantically Annotated Corpora (May 30, 2004)*, Lisboa, Publicación electrónica en: http://seneca.uab.es/csubirats/Subirats-Sato_LREC-2004.doc

Enlaces de interés

FrameNet español. URL: http://gemini.uab.es/SFN/index.html

FrameNet inglés. URL: http://framenet.icsi.berkeley.edu

FrameNet alemán (SALSA). URL: http://www.coli.uni-saarland.de/projects/salsa/

FrameNet japonés. URL: http://jfn.st.hc.keio.ac.jp