FrameNet Español: un análisis cognitivo del léxico del español[*]

Carlos Subirats Rüggeberg
Universidad Autónoma de Barcelona e International Computer Science Institute
carlos.subirats@gmail.com

1. El objetivo del proyecto FrameNet Español es la creación de una base de datos online de oraciones anotadas semántica y sintácticamente**,** partiendo de la teoría de la semántica de marcos de Fillmore (1982, 1985)**.** Las oraciones anotadas se extraen automáticamente de un corpus español de 350 millones de palabras, en función de las características de las proyecciones sintácticas de los argumentos semánticos de los predicados analizados. La semántica de marcos parte de la base de que el significado de los elementos léxicos se debe analizar en relación con los marcos semánticos que evocan, entendiendo por marco semántico, una representación esquemática de una situación, en la que están implicados varios participantes, utilería y otros roles conceptuales, cada uno de los cuales es un elemento de dicho marco o, simplemente, un argumento semántico. En FrameNet Español se analiza el significado de los predicados partiendo de los marcos semánticos que lo determinan y, a su vez, se estudian las construcciones sintácticas en las que aparecen dichos predicados, tratando de identificar cómo las características semánticas que se definen en un marco semántico adquieren una forma sintáctica en una o más construcciones sintácticas distintas.

2. El proceso de construcción de la base de datos que integra el proyecto FrameNet implica cuatro tareas distintas (Subirats 2005), que consisten básicamente en:

(1) identificar los marcos semánticos a los que pertenecen las unidades léxicas estudiadas y determinar los argumentos semánticos que forman parte de dichos marcos;
(2) determinar, a partir de un corpus, cuáles son las construcciones sintácticas en las que se proyectan los argumentos semánticos de los marcos identificados en (1);
(3) anotar semánticamente las oraciones extraídas automáticamente del corpus y,
(4) verificar los resultados de la anotación vía web con objeto de corregir posibles errores de anotación y, también, para examinar las características semánticas de los predicados estudiados.

2.1. La identificación del marco semántico al que pertenece una unidad léxica requiere, por un lado, la definición del escenario semántico que caracteriza a dicho marco y, por otro, la identificación de sus argumentos o roles semánticos. Así p. ej., si queremos realizar un análisis de *amenaza*, tenemos que dar una caracterización del marco que evoca dicha unidad léxica, así como de sus argumentos semánticos. En este caso, el nombre eventivo *amenaza* evoca un escenario en el que un *emisor* formula un compromiso a un *receptor*, en relación con la ejecución de una cierta acción futura. Esta acción puede ser deseable para el receptor, como en el caso de *prometer, promesa*, etc., pero puede ser también no deseable, como en el caso de *amenazar*, *amenaza*, etc. El escenario que estamos analizando, por lo tanto, tendría tres argumentos

[*] El proyecto de investigación FrameNet Español se está desarrollando en la Universidad Autónoma de Barcelona y en el International Computer Science Institute (ICSI, Berkeley, CA) con financiación del Ministerio de Educación (MEC) de España (TSI2005-01200). Quisiera dar las gracias a Collin Baker, Michael Ellsworth, Charles Fillmore, Covadonga López Alonso, Miriam Petruck y Josef Ruppenhofer, por toda la ayuda que me han ofrecido para el desarrollo de este proyecto. También quisiera dar las gracias al MEC por las becas que me ha concedido para realizar estancias de investigación en el ICSI.

nucleares, que resultan imprescindibles para caracterizarlo, concretamente, un *emisor*, que formula un compromiso ya sea este positivo o negativo, un *receptor* del compromiso formulado por el *emisor*, y, finalmente, un *mensaje* o un *tema* sobre el que se formula el compromiso. Estos tres argumentos semánticos pueden tener una realización nula, pero incluso en estos casos, están presentes en el conocimiento que configura este escenario y que es necesario para comprender el significado del predicado *amenaza*.

2.2. La segunda tarea en el proceso de análisis semántico de las unidades léxicas que integran FrameNet español, como hemos señalado anteriormente en 2., consiste en consultar el corpus para determinar cuáles son las construcciones sintácticas en las que se proyectan los argumentos semánticos de los predicados pertenecientes a los distintos marcos semánticos estudiados (cf. Fig. 1).

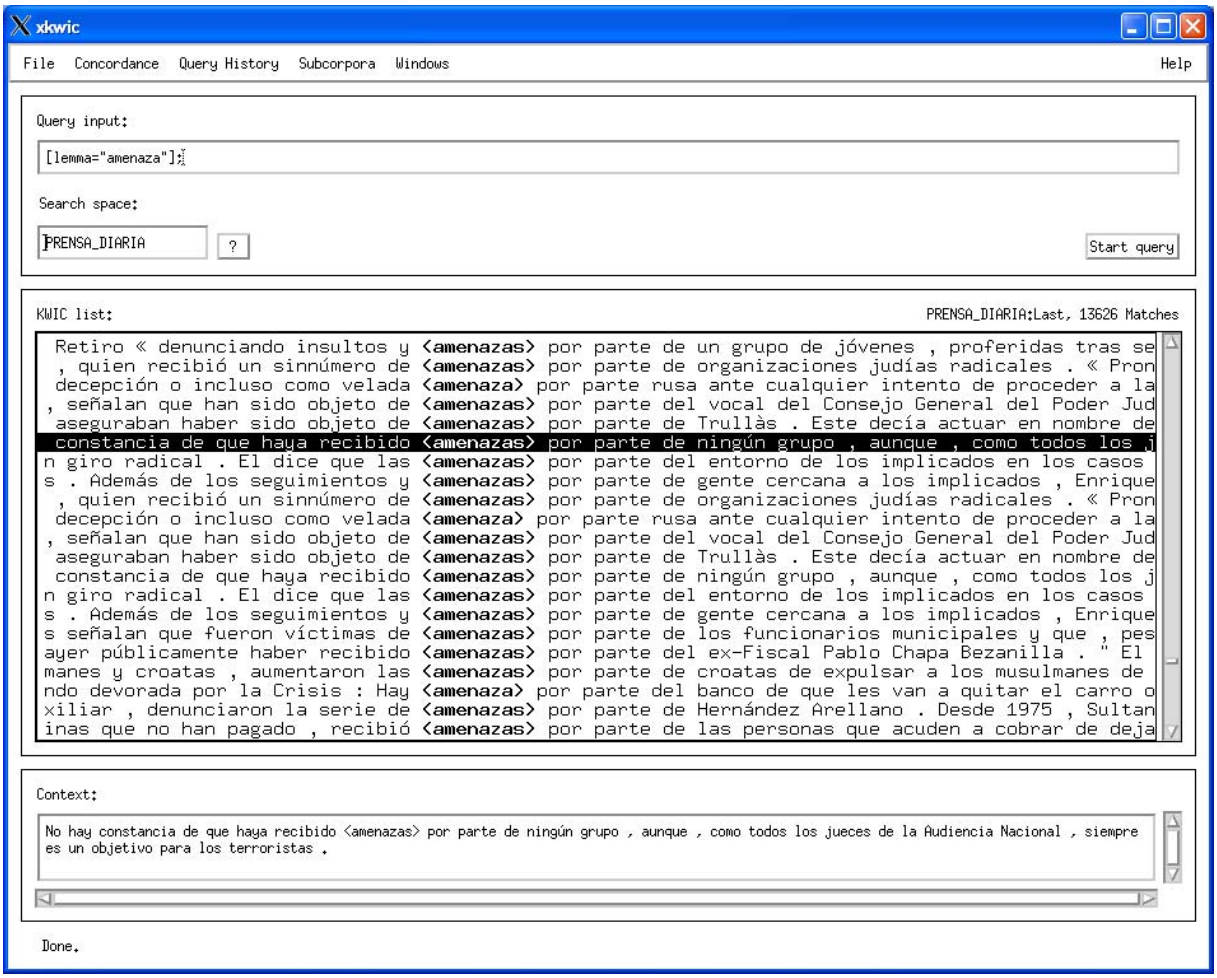

Fig. 1. Búsqueda con la aplicación XKWIC de las construcciones en las que el nombre eventivo *amenaza* va seguido por la locución prepositiva *por parte de*

Para determinar las distintas construcciones sintácticas en las que se pueden realizar los argumentos semánticos de una unidad léxica, no nos basta con nuestra competencia lingüística, puesto que, a partir de ella, no es posible determinar con precisión cuáles son las construcciones más habituales o las colocaciones más frecuentes en las que puede aparecer una unidad léxica determinada. Pero además de todo ello, en la actualidad disponemos de córpora textuales y, también, de herramientas de tratamiento de dichos córpora, que nos permiten, por un lado, estudiar de forma detallada los contextos sintácticos en los que aparece una unidad léxica y, por otro, determinar las colocaciones más frecuentes –tanto a la derecha

como a la izquierda– en las que puede aparecer. Para ello, es necesario tener herramientas de etiquetación del corpus, que nos indiquen automáticamente para cada forma del corpus (1) cuál es la clase (o clases) de palabras a la que pertenece, (2) cuál es el lema (o lemas) al que está asociado dicha forma y (3) cuáles son las propiedades morfológicas de flexión de los verbos, los nombres, los adjetivos y los participios. Dado que el corpus que utilizamos en el proyecto FrameNet Español posee esta información asociada a las formas que lo integran, el software de consulta del corpus permite realizar búsquedas en función de las mencionadas propiedades léxicas y/o morfológicas en relación con una o más unidades léxicas. Por ello, el corpus de 350 millones de palabras que utilizamos en este proyecto de investigación nos permite verificar con precisión cuáles son las distintas construcciones sintácticas en las que se manifiestan los argumentos semánticos de un predicado como *amenaza*. Así p. ej., podemos constatar que el argumento semántico *mensaje* se puede realizar sintácticamente como una oración o como un complemento infinitivo introducidos por una preposición, tal como podemos observar en las oraciones (1) y (2), respectivamente:

(1) *Ha ignorado la **amenaza** de Estados Unidos de que si no abandona el país deberá enfrentarse a una invasión de la isla* […]
(2) *La Comisión Europea criticó ayer la **amenaza** británica de bloquear la toma de decisiones conjuntas que requieran unanimidad* […]

Asimismo, el corpus pone de manifiesto que el argumento semántico *emisor* no sólo se realiza como un grupo preposicional encabezado por la preposición *de*, como p. ej., *de Estados Unidos* en la anterior oración (1), o como un adjetivo gentilicio, p. ej. *británico* en (2), sino que puede ser también un grupo preposicional encabezado por la locución prepositiva *por parte de*, como podemos observar en (3):

(3) *La presión vendedora de los últimos días se debe a un movimiento especulativo relacionado con la **amenaza** de posibles ventas por parte de los bancos centrales.*

El corpus nos permite documentar los verbos de soporte que acompañan a *amenaza*, como p. ej., *hacer* o *sufrir* en (4) y (5) respectivamente. Estos verbos de soporte introducen dos perspectivas semánticas distintas en relación con el evento que evoca *amenaza*: *hacer* focaliza el *emisor*, mientras que *sufrir* focaliza el *receptor* y, por tanto, consideramos que inducen un cambio de punto de vista semántico en la caracterización del evento que denota *amenaza*.

(4) *A pesar de que no es la primera vez que Karadzic hace **amenazas** de este tipo –y más de una vez las llevó a la práctica–, las que publica hoy Der Spiegel tienen una especial incidencia, dado que se producen en el contexto de una grave crisis humanitaria y militar de la ONU en Bosnia* […]
(5) *El juez del caso, Alberto Costa, quien sufrió **amenazas** anónimas de muerte, presuntamente por simpatizantes del MRTA, informó hoy que el proceso está en su fase de definición.*

Asimismo, el corpus que utilizamos para el desarrollo del proyecto FrameNet nos puede aportar información sobre las colocaciones en las que participan verbos que comparten algún argumento semántico con *amenaza*, como p. ej., *cumplir*, *lanzar*, *proferir*, etc., cuyos *agentes* –que sintácticamente son sus argumentos externos–, concretamente, *el grupo paramilitar priísta* en (6), *El secretario general del Polisario y presidente de la República Árabe Saharaui Democrática (RASD)* en (7), y *Los detenidos* en (8) son, a su vez, el *emisor* del nombre eventivo *amenaza* en (6), (7) y (8), respectivamente:

(6) *Precisó que la emboscada se registró en la zona que mantiene bajo control el grupo paramilitar priísta, que cumplió ayer la* ***amenaza*** *de impedir el transito de los representantes de la Iglesia Católica en las comunidades de Tila.*

(7) *El secretario general del Polisario y presidente de la República Árabe Saharaui Democrática (RASD) lanzó ayer esta* ***amenaza*** *al denunciar la lentitud en la elaboración del censo.*

(8) *Los detenidos fueron acusados de promover el bloqueo de la carretera panamericana que une Colombia y Ecuador y de proferir* ***amenazas*** *contra los vehículos que circulaban.*

2.3 La tercera tarea en relación con la construcción de FrameNet Español consiste en la anotación semántica y sintáctica de las oraciones extraídas automáticamente del corpus e importadas en la base de datos de FrameNet. Esta tarea se lleva a cabo con la herramienta FNDesktop (cf. Fig.2). Como podemos observar en la Fig. 2, en la parte izquierda aparece el

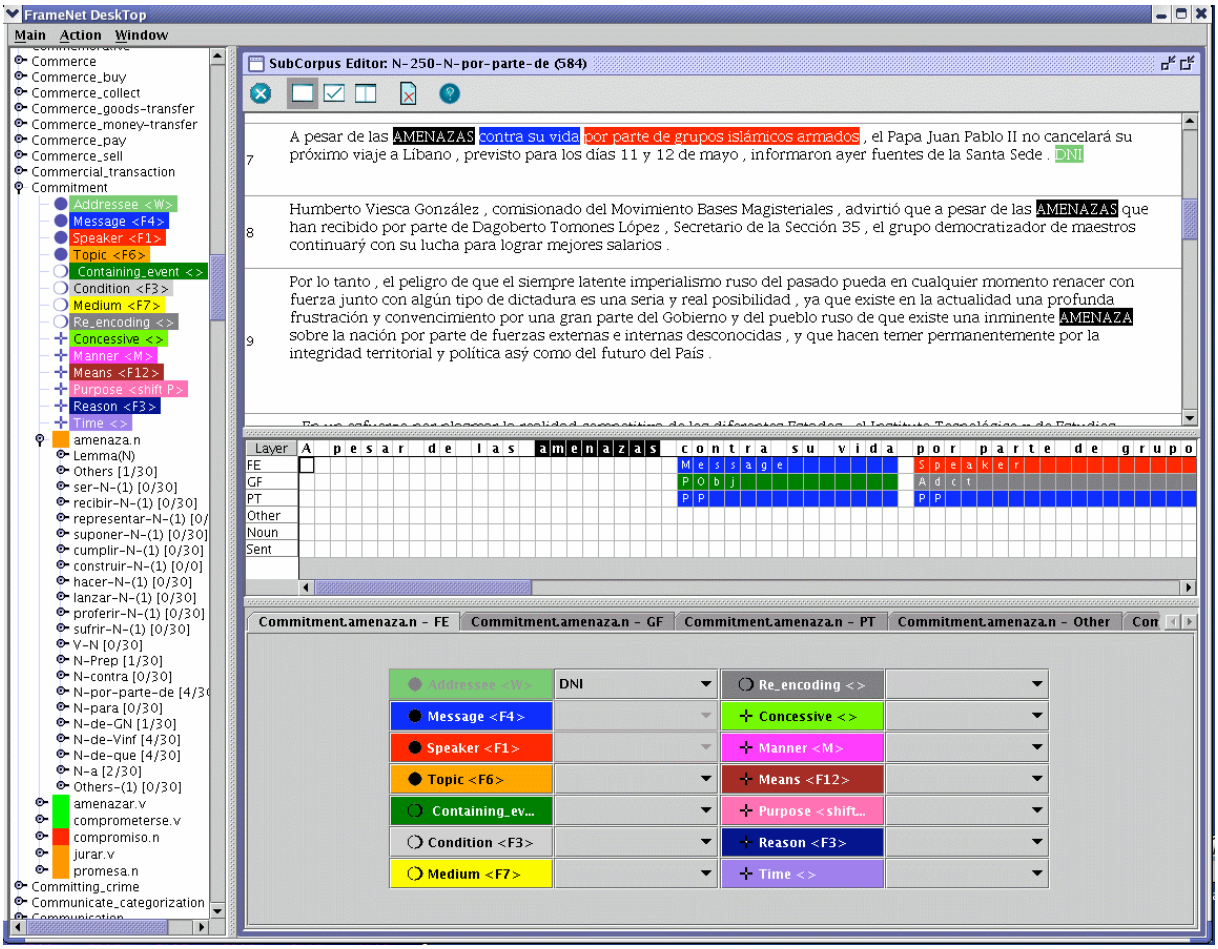

Fig. 2. Anotación semántica y sintáctica del nombre eventivo *amenaza* con la herramienta FNDesktop

menú de navegación de la herramienta de anotación. En este menú, aparece la lista de marcos semánticos; al pinchar sobre cada uno de ellos, se despliega un nuevo menú, que incluye todos los argumentos semánticos y la lista de todas las unidades léxicas pertenecientes al marco semántico correspondiente. A su vez, al pinchar sobre una unidad léxica, se despliega la lista de subcórpora que se han extraído automáticamente del corpus y cuyas oraciones han sido cargadazas previamente en la base de datos de FrameNet. Asimismo, pinchando sobre

cada uno de los subcórpora, aparece en el marco superior derecho la lista de oraciones que contiene. Para anotar una oración, se pincha sobre ella en el marco superior y esta se despliega horizontalmente en el marco central derecho, tal como podemos observar en la Fig. 1, y, en la parte inferior del marco superior, es decir, en el marco central derecho–, aparece la lista de todos los argumentos semánticos relacionados con el marco semántico en el que se está efectuando la anotación semántica. Para realizar la anotación, se selecciona con el ratón un constituyente y se le asigna el argumento semántico elegido dentro de la lista de argumentos que se encuentra en la parte inferior del marco central donde aparece la oración. Una vez seleccionado el argumento, la aplicación FNDesktop le asigna automáticamente al constituyente correspondiente su función sintáctica, p. ej., sujeto, objeto directo, etc., y la clase de constituyente a la que pertenece, p. ej., grupo nominal, grupo preposicional, etc.

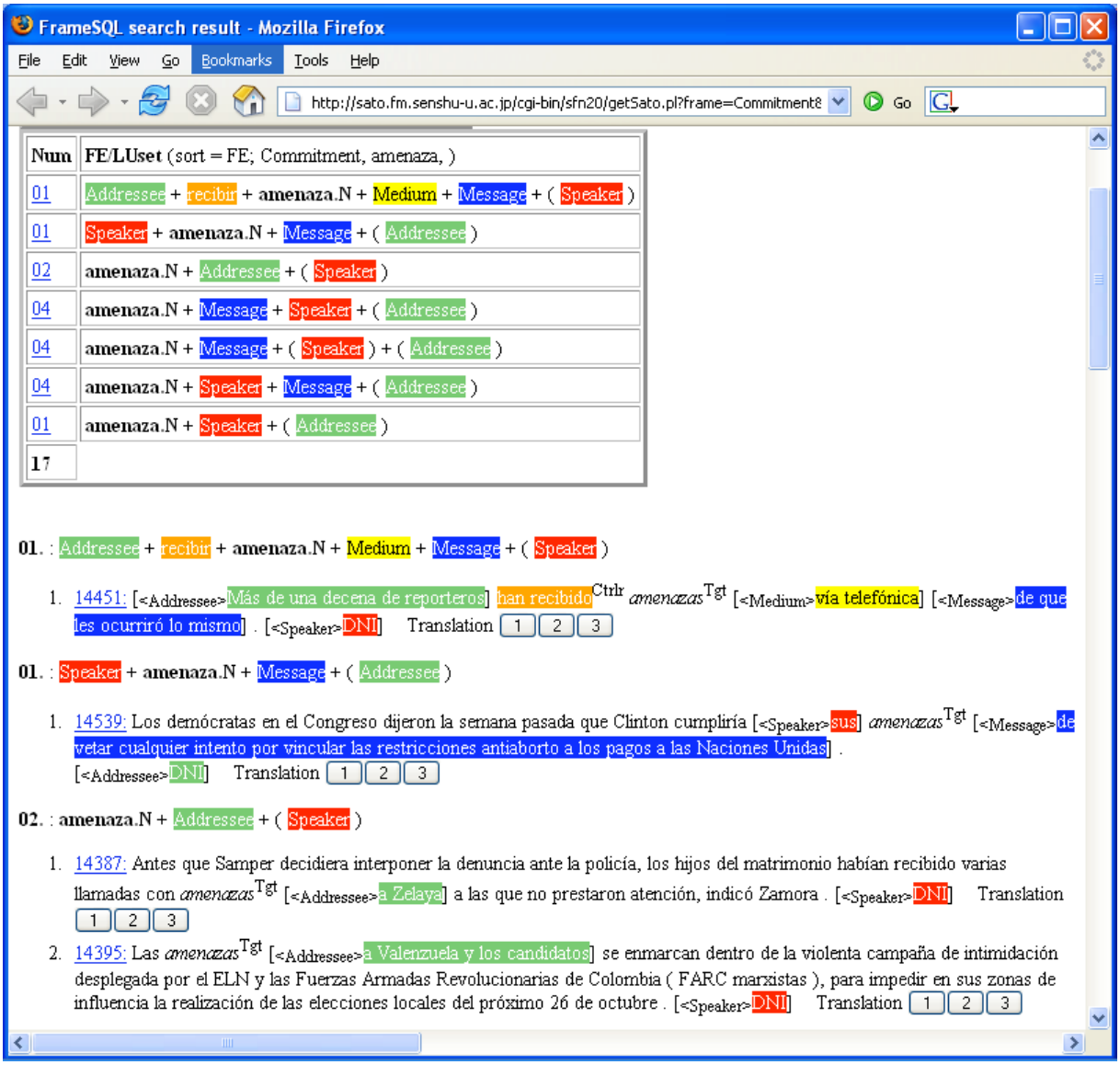

Fig. 3. Visualización parcial de la anotación semántica y sintáctica de amenaza en el marco de *Compromiso* (Commitment)

2.4. La última tarea en la construcción de FrameNet Español consiste en el análisis de los resultados de la anotación con las distintas herramientas de las que disponemos en estos momentos en el proyecto. Así p. ej., en la Fig. 3, podemos observar los resultados parciales de la anotación de *amenaza* en el marco semántico de *Compromiso* (Commitment) con la

herramienta Spanish FrameSQL, que ha sido desarrollado por el Prof. Hiroaki Sato en la Universidad de Senshu (Japón) y, posteriormente, adaptada al español.

Como podemos observar en la Fig. 3, una de las modalidades de visualización que ofrece la herramienta Spanish FrameSQL nos permite ver las combinaciones de argumentos semánticos en el mismo orden en el que aparecen en la oración anotada, en relación con la posición que ocupa en dicha oración el nombre eventivo *amenaza*. Los argumentos se encuentran entre paréntesis cuando tienen una realización nula en la oración.

3. La base de datos del proyecto FrameNet Español será de dominio público a partir de julio de 2007 y, por tanto, los resultados de este proyecto de investigación se podrán consultar libremente en la red mediante las aplicaciones FNDesktop y Spanish FrameSQL (cf. Fig. 3). Posteriormente, tanto el contenido de la base de datos, es decir, el corpus de oraciones anotadas semántica y sintácticamente, como el software para su gestión y consulta, se podrá descargar libremente desde la web del proyecto **(**http://gemini.uab.es/SFN**)** tras solicitar una licencia gratuita. La base de datos de FrameNet Español permitirá desarrollar nuevas aplicaciones en el ámbito del tratamiento automático de la información textual en español, que posibilitarán el desarrollo de nuevas tecnologías para el procesamiento semántico automático y las nuevas formas de tratamiento de la información textual que va a requerir el futuro desarrollo de la web semántica en español. La visualización vía web de la reorganización automática de la información de la base de datos de FrameNet español mediante FNDesktop y FrameSQL, en función de las clases semánticas y sus argumentos, la combinatoria de argumentos, etc., proporcionará un diccionario semántico *online***,** que abrirá nuevas perspectivas para el análisis cognitivo de las características semánticas de los predicados del léxico español. Finalmente, la posibilidad que ofrece Spanish FrameSQL 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 *online* inglés-español y español-inglés, el cual, además de tener aplicaciones para la consulta humana, tendrá sin duda repercusiones en el desarrollo de sistemas de traducción automática basados en el análisis cognitivo del léxico.

Referencias

Baker, Collin F.; Fillmore, Charles J.; Cronin, Beau. 2003. The Structure of the Framenet Database. *International Journal of Lexicography* 16.3:281-296.

Boas, Hans C. 2006. A frame-semantic approach to identifying syntactically relevant elements of meaning". In P. Steiner, H. C. Boas y S. Schierholz, eds. *Contrastive Studies and Valency.Studies in Honor of Hans Ulrich Boas.* Frankfurt / New York: Peter Lang, págs. 119-149.

Burchardt, Aljoscha; Erk, Katrin; Frank, Anette; Kowalski, Andrea; Padó, Sebastian; Pinkal, Manfred. 2006. The SALSA Corpus: a German Corpus Resource for Lexical Semantics. *Proceedings of LREC 2006*, Génova: http://www.coli.uni-saarland.de/%7Epado/pub/papers/lrec06_burchardt1.pdf

Castellón, Irene; Fernández, Ana.; Vázquez, Gloria; Alonso, Laura; Capilla, Joan A. 2006. The Sensem Corpus: a Corpus Annotated at the Syntactic and Semantic Level. *Fifth International Conference on Language Resources and Evaluation (LREC)*: http://grial.uab.es/archivos/LREC2006def.pdf

Erk, Katrin; Padó, Sebastián. 2006. Shalmaneser. A flexible toolbox for Semantic Role Assignment. *Proceedings of Language Resources and Evaluation (*FREC) 2006: http://www.coli.uni-saarland.de/~pado/pub/papers/lrec06_erk.pdf

Fillmore, Charles J. 1985. Frames and the semantics of understanding. *Quaderni di Semántica* 6.2:222-254.

Fillmore, Charles J. 1982. Frame semantics. En *Linguistics in the Morning Calm*, Seúl: Hanshin Publishing Co., págs.111-137.

Fillmore, Charles J.; Jonson, Christopher R; Petruck, Miriam R.L. 2003. Background to FrameNet. *International Journal of Lexicography* 16.3:235-250.

García-Miguel, J. M.; Albertuz, Francisco J. 2005. Verbs, Semantic Classes and Semantic Roles in the ADESSE project. *Interdisciplinary Workshop on Verb Features and Verb Classes*. Saarbrücken: http://webs.uvigo.es/adesse/textos/saarb05.pdf

Gildea, Daniel; Jurafsky, Daniel. 2002. Automatic Labeling of Semantic roles. *Computational Linguistics* 28.3:245-288.

Ohara, Kyoko H.; Fujii, Seiko; Saito, Hiroaki; Ishizaki, Shun; Ohori, Toshio; Suzuki, Ryoko. 2003. The Japanese FrameNet Project: A Preliminary Report. *Proceedings of Pacific Association for Computational Linguistics (PACLING'03)*, págs. 249-254. http://jfn.st.hc.keio.ac.jp/publications/PACLING03.pdf

Petruck, Miriam R. L. 1996. Frame Semantics. In J. Verschueren, J.-O. Östman, J. Blommaert y C. Bulcaen, eds. *Handbook of Pragmatics*. Ámsterdam / Philadelphia: John Benjamins. http://framenet.icsi.berkeley.edu/papers/miriamp.FS2.pdf

Pradhan, Sameer.; Ward, Wayne.; Hacioglu, Kadri.; Martin, James.; Jurafsky, Dan. 2004. Shallow Semantic Parsing using Support Vector Machines. *Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL-2004), Boston, MA, May 2-7, 2004*: http://oak.colorado.edu/~spradhan/publications/pradhan-hlt-2004-a.pdf

Ruppenhofer, Josef; Ellsworth, Michael; Petruck, Miriam R. L.; Johnson, Christopher R. 2005. *FrameNet: Theory and Practice*: http://framenet.icsi.berkeley.edu/book/book.html

Scheffczyk, Jan; Baker, Collin F.; Narayanan, Srini. 2006. Ontology-based Reasoning about Lexical Resources. *OntoLex 2006: Interfacing Ontologies and Lexical Resources for Semantic Web Technologies.*

Subirats Rüggeberg, C. 2005. FrameNet español. Una red semántica de marcos conceptuales. En E. Serra y G. Wotjak, eds. *Cognición y percepción lingüísticas.* Valencia:

Universidad de Valencia y Universidad de Leipzig, págs. 182-196. http://gemini.uab.es/SFN/papers/Leipzig_Paper.pdf

Subirats Rüggeberg, C.; Petruck, Miriam R. L. 2003. Surprise: Spanish FrameNet! In E. Hajicova, A. Kotesovcova y J. Mirovsky eds. *Proceedings of CIL 17*. (CD-ROM). Prague: Matfyzpress: http://www.icsi.berkeley.edu/%7Eframenet/papers/SFNsurprise.pdf