La función del corpus en FrameNet Español[1]

Carlos Subirats Rüggeberg
<subirats@icsi.berkeley.edu>
Universidad Autónoma de Barcelona e International Computer Science Institute

Resumen

En este artículo, destacamos la importancia de la utilización de un corpus textual en el desarrollo de FrameNet Español (FNE), un proyecto de análisis semántico del léxico basado en la teoría de la semántica de marcos. En primer lugar, se analizan aplicaciones de tratamiento automático de corpus, que se fundamentan, por un lado, en la utilización de léxicos electrónicos y, por otro lado, en la representación de los resultados del análisis lingüístico en forma de autómatas y en el uso de transductores para sistematizar tanto características formales del léxico –locuciones verbales– como de la sintaxis –gramáticas para la extracción automática de construcciones de un corpus–. Paralelamente, proponemos la utilización de la intersección de los autómatas y transductores como un formalismo de análisis lingüístico para el tratamiento automático de textos. En segundo lugar, planteamos cómo la creación de un corpus de oraciones con anotación semántica ha permitido poner de manifiesto características semánticas del léxico del español. En tercer lugar, destacamos la utilidad del corpus de oraciones anotadas de FNE como corpus de entrenamiento de una aplicación estadística de etiquetación automática de roles semánticos. Finalmente, se proponen nuevas líneas de investigación, concretamente, el desarrollo de sistemas de análisis automático de corpus, que integren el análisis semántico y sintáctico en un proceso único, con objeto de acercar el tratamiento automático de la información textual a los procesos cognitivos de comprensión del lenguaje.

1. Introducción

El objetivo del proyecto de investigación FrameNet Español[2] (FNE) es realizar un análisis semántico del léxico, partiendo de la teoría de la semántica de marcos (Fillmore 1982, 1985) y la gramática de construcciones (Lakoff 1987, Fillmore, Kay y O'Connor

1 FrameNet Español (FNE) se está desarrollando en la Universidad Autónoma de Barcelona con la cooperación del International Computer Science Institute (ICSI). FNE está financiado por el Ministerio de Ciencia e Innovación (MICINN, FFI2008-0875) y la Fundación Comillas. En el marco del acuerdo de colaboración entre el ICSI y el MICINN, una beca para estancias de tecnólogos españoles me ha permitido desarrollar partes del proyecto FNE en el ICSI. Quisiera darles las gracias a Collin Baker y a Charles J. Fillmore por su ayuda en el desarrollo de este proyecto. También deseo expresar mi especial agradecimiento a Michael Ellsworth.

2 http://gemini.uab.es/SFN

1988, Goldberg 1995). Los resultados de este proyecto están organizados en una base de datos integrada por oraciones con anotación semántica y sintáctica. En su estado actual, FNE ha estudiado más de 1.000 unidades léxicas, pertenecientes a 300 marcos semánticos. Los resultados de FNE se pueden consultar libremente en la red[3] mediante las aplicaciones Web Reports y FrameSQL. Los Web Reports permiten acceder a las definiciones de los marcos semánticos, así como a la anotación semántica y sintáctica de cada una de las distintas unidades léxicas que forman parte de dichos marcos semánticos.

La aplicación FrameSQL[4] permite además realizar consultas globales dentro de la base de datos de FNE, lo que posibilita, p. ej., (1) buscar todas las oraciones en las que aparece un determinado rol semántico, como p. ej. SPEAKER, un verbo de soporte, como p. ej. *dar*, *hacer*, etc., o (2) recuperar todas las oraciones de la base de datos, en las que la unidad léxica analizada tiene un uso metafórico. A su vez, el corpus de oraciones anotadas creado en FNE se está utilizando además como un corpus de entrenamiento de una aplicación estadística de etiquetación automática de roles semánticos.

2. El corpus de FNE: análisis léxico y extracción automática de construcciones

1. Las oraciones anotadas semántica y sintácticamente de FNE se extraen a partir de un corpus de 400 millones de palabras. Esta extracción se realiza automáticamente y su objetivo es reunir una selección de las distintas realizaciones sintácticas asociadas a un predicado perteneciente a un marco semántico determinado. Estas oraciones extraídas automáticamente del corpus son las que posteriormente se anotan semántica y sintácticamente mediante procedimientos semiautomáticos.

Para poder extraer automáticamente las construcciones sintácticas relacionadas con un predicado, el corpus de FNE ha sido previamente sometido a un análisis léxico automático, que asigna la clase de palabra, el lema y las propiedades de flexión a los nombres, los verbos y los adjetivos (Subirats y Ortega 2000). El sistema de análisis léxico permite reconocer y etiquetar tanto las formas simples como las locutivas y, dentro de estas últimas, el sistema detecta tanto las locuciones que se pueden reconocer a partir de un diccionario, como p. ej., *carga de profundidad*, como las locuciones que requieren tanto información léxica como sintáctica, como p. ej. *tener en cuenta* en *Lo hizo* ***teniéndo****lo siempre* ***en cuenta****,* donde el reconocimiento de la locución verbal no se puede llevar a cabo únicamente con un diccionario o con procedimientos estadísticos. Asimismo, un analizador sintáctico que utiliza transductores permite detectar las construcciones sintácticas que se quieren extraer para su posterior anotación semántica.

2. Para poder realizar una extracción automática precisa de construcciones a partir del corpus de FNE, es necesario etiquetarlo. El proceso de etiquetación se realiza mediante un analizador léxico que utiliza un diccionario electrónico expandido del español integrado por 610.000 formas, que incluyen tanto formas simples, es decir, palabras que constituyen una unidad ortográfica, como locuciones (Subirats y Ortega 2000). El diccionario expandido se genera automáticamente a partir de un diccionario que contiene 103.000 lemas, concretamente, 78.000 formas simples, como p. ej., *unir*, *inmoralidad*,

3 http://gemini.uab.es:9080/SFNsite/sfn-data

4 FrameSQL ha sido desarrollado por el Prof. Hiroaki Sato en la Universidad de Senshu (Japón).

anticonstitucional, etc. y 25.000 locuciones –a excepción de locuciones verbales–, como p. ej., *muerte cerebral*, *carga de profundidad*, etc. Cada una de las formas del diccionario expandido está asociada a un conjunto de etiquetas, que indican el lema, la clase de palabra y las propiedades de flexión de los verbos (tiempo, modo, persona y número), y los nombres y adjetivos (género y/o número).

El output del etiquetador léxico es un conjunto de autómatas –un autómata por cada oración–, cuyas transiciones están etiquetadas con la información que el diccionario expandido posee sobre cada una de las formas simples o locutivas del texto analizado (Fig. 1).

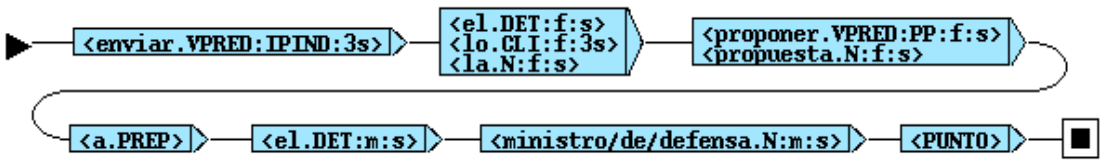

Fig. 1. Etiquetación léxica automática de la oración *Envió la propuesta al ministro de defensa*

Las ambigüedades que genera el proceso de etiquetación léxica se formalizan como distintas transiciones entre dos estados consecutivos del autómata y se esquematizan en su representación gráfica como conjuntos de transiciones dentro de un mismo bloque orientado. Estas ambigüedades se pueden suprimir mediante la intersección del autómata que formaliza el análisis léxico de la oración con un transductor que especifique condiciones locales de desambiguación, basadas en regularidades distribucionales observadas en el corpus. Así p. ej., la forma *la* puede ser un determinante, un pronombre clítico o un nombre, que se refiere a una nota de la escala diatónica. Sin embargo, la regularidad observada en el corpus es que cuando *la* se encuentra delante de un verbo en forma personal, *la* es un pronombre clítico y, por tanto, en este caso se pueden descartar las etiquetas de determinante y nombre.

3. El reconocimiento de locuciones verbales se realiza mediante la intersección de los autómatas resultantes del análisis léxico con 2.000 transductores que formalizan las características sintácticas más relevantes de las locuciones verbales en español (Fig. 2). Estos transductores permiten el reconocimiento de una locución verbal, aunque tenga pronombres clíticos, adverbios u otro material léxico entre su núcleo verbal y su parte constante, ya que el transductor especifica la posición y el tipo de material léxico que puede aparecer en esta posición. Obsérvese que la identificación de una locución verbal puede implicar además un proceso indirecto de desambiguación. Así p. ej., el transductor de la Fig. 2 permite identificar la locución verbal *dar por sentado* en la oración de la Fig. 3, como podemos observar en la Fig. 4. Pero obsérvese que el propio proceso de identificación de la locución *dar por sentado*, elimina también las ambigüedades asociadas a su núcleo verbal *dar* y a su parte constante *por sentado*, ambigüedades que podemos apreciar en la Fig. 3.

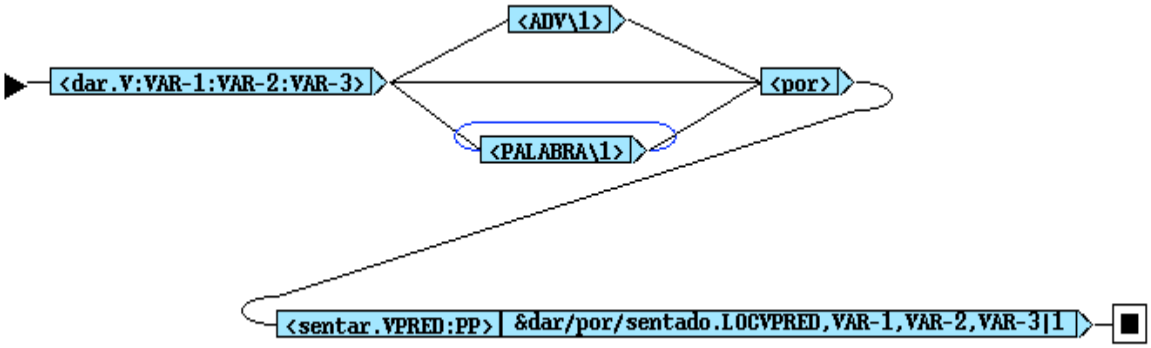

Fig. 2. Transductor con variables que permite reconocer la locución verbal *dar por sentado.*

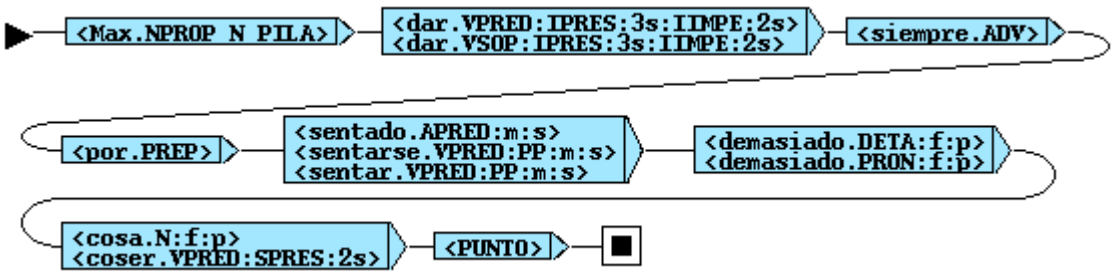

Fig. 3. Autómata resultante del análisis léxico de la oración *Max da siempre por sentado demasiadas cosas.*

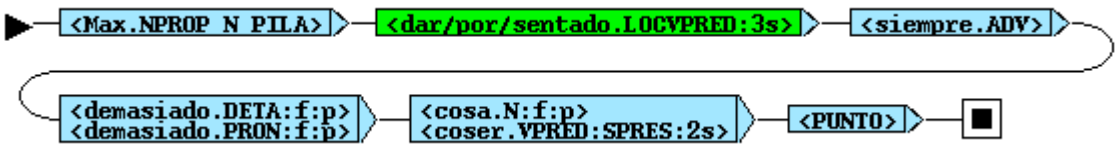

Fig. 4. Autómata resultante intersección del autómata de la Fig. 3 con el transductor de la Fig. 2.

Tras la realización del proceso de análisis léxico del corpus, se procede a identificar las formas verbales compuestas, como p. ej., *había comido*, *hubieran tenido en cuenta*, etc., mediante transductores, que formalizan las características sintácticas de los tiempos verbales compuestos del español. El proceso de extracción de construcciones sintácticas asociadas a un predicado para su posterior anotación semántica se realiza también con transductores, que especifican las características formales de las construcciones que se quieren extraer del corpus.

3. La anotación semántica y las características semánticas del léxico español

Una vez seleccionadas con los procedimientos señalados anteriormente en 2. las construcciones sintácticas asociadas a un predicado, se procede a su anotación semántica y sintáctica mediante la aplicación FNDesktop[5]. La anotación semántica consiste en identificar los roles semánticos asociados a un predicado en función del marco semántico al que pertenece. Una vez seleccionado un constituyente y tras asignarle su rol semántico, el sistema especifica automáticamente el tipo de constituyente, p. ej., grupo nominal, oración subordinada, etc., y, asimismo, en la mayoría de los casos, le asigna también automáticamente su función sintáctica (Fig. 5)

El análisis semántico llevado a cabo en FNE ha permitido poner de manifiesto algunas características semánticas del léxico del español. Así p. ej., el análisis de los verbos de emoción nos ha permitido observar que en español la conceptualización de cambios de estado emocional se lexicaliza con verbos incoativos, como p. ej. *sorprenderse*, mientras que en lenguas como el inglés, se expresan en general mediante estados, p. ej., *be surprised* (Subirats y Petruck 2003). A su vez, el análisis de los marcos semánticos de movimiento, ha puesto de manifiesto algunas de las diferencias en la conceptualización de los eventos de movimiento en español y en inglés. Así, mientras que el inglés prefiere conceptualizar los eventos de movimiento en su totalidad, como p. ej. en *He ran out of the house into the garden*, oración que no tiene una equivalencia directa en español, como podemos observar en ?*Salió corriendo de la casa al jardín*, en español preferimos centrarnos en el estado inicial y final del evento de movimiento (Ellsworth et al. 2006). Por el contrario, en español es posible incluir complementos de intención asociados a los eventos de movimiento, p. ej., *Jorge fue a Madrid* ***a ver un amigo*** *para pedirle dinero*, mientras que en inglés se tiene que separar la intencionalidad, del evento de movimiento, p. ej., *John went to Madrid to visit a friend and* ***ask him for money*** (Subirats y Sato 2004).

5 La aplicación FNDesktop ha sido desarrollado por el proyecto FrameNet (http://framenet.icsi.berkeley.edu) y ha sido adaptada al español en el marco de nuestro proyecto de investigación.

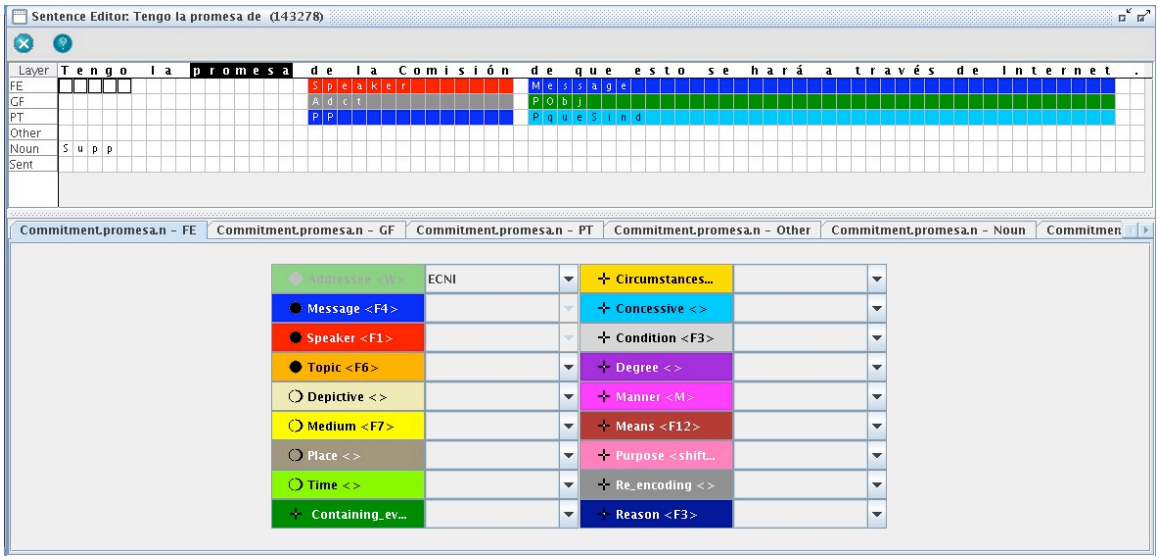

Fig. 5. Anotación semántica y sintáctica de la oración *Tengo la promesa de la comisión de que esto se hará a través de Internet* mediante la aplicación FNDesktop.

4. Aplicaciones de FNE: etiquetación semántica automática en español

El corpus de oraciones anotadas semánticamente de FNE se utiliza como corpus de entrenamiento de Shalmaneser (Erk y Padó 2006), una aplicación estadística de etiquetación automática de roles semánticos (Fig. 6). Tras su fase de entrenamiento, Shalmaneser puede llevar a cabo una anotación automática de roles semánticas de un texto en español, siempre que las palabras que integran dicho texto hayan sido analizadas en FNE y, por tanto, hayan formado parte de su corpus de entrenamiento. El input de Shalmaneser tiene que ser un texto español que haya sido sometido a un análisis léxico y sintáctico, lo que en el marco de nuestro proyectos se consigue con un sistema de transducciones en cascada (Subirats y Ortega 2000).

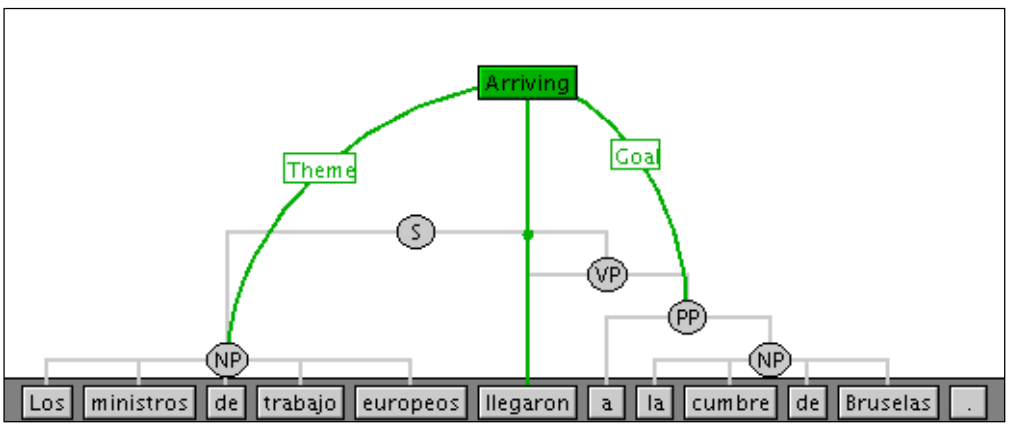

Fig. 6. Asignación automática de roles semánticos con Shalmaneser en la oración *Los ministros de trabajo europeos llegaron a la cumbre de Bruselas*.

5. Conclusiones

FNE ha puesto de manifiesto la importante función de la utilización de un gran corpus textual del español y, asimismo, el interés de utilizar la semántica de marcos para:

(1) mostrar las características semánticas del léxico español e
(2) impulsar aplicaciones de tratamiento automático de la información textual, como la etiquetación automática de roles semánticos.

Uno de los objetivos inmediatos fundamentales de FNE es, en primer lugar, ampliar su cobertura léxica y, en segundo lugar, desarrollar un procedimiento de tratamiento automático de la información textual en español que permita integrar el análisis léxico, sintáctico y semántico en un módulo único (Bryant 2008), acercando así los procedimientos de tratamiento automático a los procesos cognitivos de comprensión del lenguaje.

Referencias

Bryant, John. 2008. *Best-Fit Constructional Analysis*. Ph. D. diss., University of California Berkeley: http://www.icsi.berkeley.edu/~jbryant/bryantdissertation.pdf

Ellsworth, Michael; Ohara, Kyoko; Subirats, Carlos; Schmidt, Thomas. 2006. Frame-semantic analysis of motion scenarios in English, German, Spanish, and Japanese. ICCG4, *Fourth International Conference on Construction Grammar, Tokyo (Japan)*

Erk, Katrin; Padó, Sebastian. 2006. Shalmaneser. A flexible toolbox for semantic role assignment. *Proceedings of LREC* 2006: http://www.coli.uni-saarland.de/~pado/pub/papers/lrec06_erk.pdf

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

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

Fillmore, Charles J.; Kay, Paul; O'Connor. Catherine. 1988. Regularity and idiomaticity in grammatical constructions: The case of *Let alone*. *Language* 64/3: 501-538.

Goldberg, Adele. 1995. *Constructions. A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.

Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: The University of Chicago Press.

Subirats, Carlos. (en prensa). Spanish FrameNet: A Frame Semantic analysis of the Spanish lexicon. En Hans Boas, ed. *Multilingual FrameNets in Computational Lexicography*. New York/Berlin: Mouton de Gruyter.

Subirats, Carlos; Ortega, Marc. 2000. Tratamiento automático de la información textual en español mediante bases de información lingüística y transductores. *Estudios de Lingüistica del Español* 10: http://elies.rediris.es/elies10/

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

Subirats, Carlos; Sato, Hiroaki. 2004. Spanish FrameNet and FrameSQL. *4th International Conference on Language Resources and Evaluation*. Workshop on Building Lexical Resources from Semantically Annotated Corpora, Lisbon (Portugal), mayo de 2004: http://gemini.uab.es/SFNpub/papers/subirats-sato.pdf