Lingüística de corpus y semántica cognitiva[1]

Carlos Subirats Rüggeberg
Universidad Autónoma de Barcelona
carlos.subirats@gmail.com
Facultad de Letras, Edificio B, Dept. Filología Española, 08193 Bellaterra, España

1. Introducción

Resumen en español

En este artículo, analizamos, en primer lugar, la tradición lingüística que ha buscado tanto fundamentar documentalmente sus propuestas lingüísticas, como contextualizar los datos lingüísticos de dicha documentación. Este primer análisis está motivado por el hecho de que es dicha tradición, así como sus prácticas, las que han precedido el desarrollo de la lingüística de corpus actual. Asimismo, analizamos dos nuevos factores que, en la actualidad, han contribuido al desarrollo de la lingüística de corpus, concretamente, (1) la voluntad explícita de estudiar el lenguaje a partir de enunciados producidos en contextos comunicativos reales, y (2) la posibilidad de procesar semántica y sintácticamente textos de miles de millones de palabras, que actualmente ofrece la lingüística computacional. En segundo lugar, presentamos un sistema de etiquetación léxica, que utiliza un diccionario electrónico del español de 634.000 formas, que se generan a partir de un diccionario de 114.000 lemas, que incluyen tanto palabras formadas por una única unidad ortográfica, como locuciones formadas por más de una palabra. Asimismo, presentamos un sistema de extracción automática de construcciones sintácticas, que utiliza transductores, es decir, autómatas con función de salida. Para llevar a cabo dicha extracción, el sistema realiza una transducción sobre el fichero de salida del sistema de etiquetación léxica. En tercer lugar, mostramos cómo la utilización de la semántica de marcos (Fillmore 1985) nos permite realizar una extracción de construcciones sintácticas, sobre las que podemos realizar un análisis del significado que aportan, a la construcción oracional, los roles semánticos pertenecientes a los marcos semánticos que evocan los predicados analizados. Finalmente, mostramos la vinculación del planteamiento semántico-sintáctico de la semántica de marcos con la tradición lexicográfica iniciada por Cuervo (1896, 1893), quien define el significado de las entradas de su *Diccionario de construcción y régimen* en función de las características semánticas de las construcciones en las que aparecen.

Palabras clave: lingüística de corpus, análisis léxico automático, análisis automático de constituyentes, extracción automática de construcciones, semántica de marcos

Abstract

In this paper, I analyze a linguistic tradition that has tried to document its linguistic theories, and to contextualize the linguistic facts of their proposals. The interest on this tradition is motivated by the fact that it is its theories and methodologies that contributed to establish the foundation of current corpus linguistics. Likewise, I analyze two new aspects that, at present, have contributed to the development of corpus linguistics, specifically, (1) the explicit

intention of cognitive linguistics to study language based on discourses produced in real communicative contexts, and (2) the possibility to process semantically and syntactically texts formed by billions of words, currently offered by computational linguistics. Secondly, in this paper, we present a Spanish POS tagger and lemmatizer, which uses an electronic dictionary of Spanish of 634,000-word forms, which are automatically generated by expanding a dictionary of 114,000 lemmas. The electronic dictionary of word forms includes both single words as well as multi-word idioms. Chunking is carried out by intersecting the output of the tagger and lemmatizer with finite state transducers, which specify the structure of noun phrases, prepositional phrases, etc. The chunking process is also used to create subcorpora of sentences, where predicates appear in specific constructions. The application of frame semantics (Fillmore 1985) allows us to create specific subcorpora, according to the way semantic roles are mapped into sentence constituents. Finally, we show the link between this semantico-syntactic approach and the lexicographic tradition started by Cuervo (1896, 1893), who defines the meaning of lexical entries, based on the semantic contribution they make to the constructions where they appear.

Keywords: Corpus linguistics, POS tagging, lemmatization, chunking, frame semantics

2.1. **Conceptos fundamentales**

Aunque el término *lingüística de corpus* es relativamente reciente, el interés por la documentación contextualizada del léxico tiene una larga historia en nuestra tradición cultural. Así p. ej., las primeras concordancias bíblicas estaban formadas por índices alfabéticos de palabras utilizadas en la biblia, acompañadas por las correspondientes referencias a su localización en dicho texto. Aunque estas concordancias no se crearon con un objetivo propiamente lingüístico, constituyen una aproximación a la contextualización del léxico en un texto, que ha tenido una gran importancia cultural.

Asimismo, desde sus inicios, ha sido la voluntad explícita de proporcionar una documentación empírica de los hechos lingüísticos, lo que ha llevado a los estudiosos del lenguaje a documentar con textos reales, en general pertenecientes al lenguaje escrito, la validación empírica de los fenómenos lingüísticos estudiados independientemente de su naturaleza. A esta voluntad de documentación de los hechos lingüísticos estudiados, se le han sumado en la actualidad otros aspectos, que son (1) los que le han dado una consolidación definitiva a la lingüística que ha tratado de fundamentar sus hallazgos a partir de textos reales y (2) los que han contribuido al desarrollo de lo que hoy denominamos la *lingüística de corpus*. Estos aspectos, a los que hemos hecho referencia anteriormente, son fundamentalmente dos. En primer lugar, las teorías lingüísticas cognitivas, que han surgido como una voluntad de superar el reduccionismo formalista de propuestas teóricas, como p. ej., la gramática generativa, han utilizado la lingüística de corpus no sólo como una forma de documentar sus propuestas semánticas, pragmáticas, etc., sino, fundamentalmente, como una forma de documentación de dichas propuestas, basada en discursos surgidos en contextos reales y como resultado de un propósito comunicativo real ya sea en el lenguaje hablado o escrito. Es decir, el cognitivismo no trata de documentar únicamente sus propuestas, sino que lo que intenta es documentar su investigación a partir de enunciados formulados con un propósito comunicativo real. En segundo lugar, el desarrollo de medios computacionales cada vez más precisos para analizar discursos orales o escritos en soporte electrónico, especialmente, el desarrollo de sistemas de procesamiento léxico, sintáctico y semántico automáticos, que no sólo utilizan procedimientos estadísticos, sino que, además, utilizan

bases de información lingüística, desarrolladas dentro de teorías que integran la sintaxis en la semántica, como p. ej., la sintaxis cognitiva, han producido un cambio cualitativo del nivel de precisión de los sistemas de procesamiento automático de la información textual de las lenguas más estudiadas, como p. ej., el inglés y, otras lenguas, como el español, etc. La consecuencia inmediata de este cambio cualitativo en la calidad del procesamiento automático del lenguaje natural ha sido el aumento del nivel de precisión con el que hoy se pueden realizar (1) búsquedas de construcciones sintácticas específicas dentro de corpus de textos en soporte electrónico o (2) etiquetaciones automáticas de roles semánticos. Finalmente, la creación de grandes corpus textuales, como p. ej., el ESCOW14[2], un corpus español de domino público, que contiene 3.681 millones de palabras o unidades ortográficas, repartidas en 150 millones de oraciones, que están libres de derechos de autor. Iniciativas como la que acabamos de mencionar, han puesto fin a los problemas creados en el mundo hispánico por la Real Academia Española (RAE), que ha recibido decenas de millones de euros de financiación estatal y privada para crear varios corpus del español, que no sólo no se pueden descargar, sino que tampoco se pueden comprar ni total ni parcialmente. La caótica política lingüística oficial española en relación con el desarrollo de los recursos lingüísticos básicos para el desarrollo de la lingüística computacional tanto a nivel científico como empresarial ha sido objeto de múltiples críticas en el ámbito lingüístico, críticas que no han tenido ninguna incidencia para la conversión de la anacrónica red estatal de "reales academias españolas" en verdaderos centros de investigación competitivos.

2.1. Una visión retrospectiva

En la tradición lingüística hispánica, existen obras fundamentales, que constituyen verdaderos ejemplos de lo que hoy denominaríamos lingüística de corpus y que merecen ser estudiadas con mucho detenimiento, con objeto de poder escribir la historia de la documentación de las propuestas lingüísticas dentro de la lingüística hispánica. Así p. ej., una de las grandes obras que fundamenta su investigación en un amplio corpus de textos es el *Diccionario de construcción y régimen* de Rufino J. Cuervo (1886, 1893). Al igual que en la lingüística cognitiva actual, Cuervo estudia en su *Diccionario* las redes de acepciones (Subirats 2018), partiendo de su documentación en construcciones sintácticas extraídas de multitud de textos, tales como obras literarias, ensayos, etc. La novedad de la propuesta de Cuervo no reside únicamente en el hecho de documentar su análisis lexicográfico en un corpus textual, sino en utilizar además una teoría conceptual de la metáfora y la metonimia (Subirats 2010) para establecer una relación lingüísticamente motivada entre las distintas acepciones de las unidades léxicas analizadas en su *Diccionario*, convirtiendo en redes de relaciones semánticas entre distintas acepciones, lo que en la tradición lexicográfica anterior a Cuervo eran meras listas de acepciones. Para llevar a cabo este innovador estudio de la polisemia, Cuervo fundamentó su análisis lexicográfico en ejemplos extraídos de un corpus, poniendo en relación el significado de las diferentes acepciones de las entradas polisémicas con las distintas construcciones sintácticas en las que podían aparecer. Pero Cuervo no se limitó a definir las acepciones de las entradas que aparecían en las oraciones extraídas de su corpus, sino que, además de ello, vinculó los aspectos diferenciales del significado de cada acepción, con características sintácticas específicas de las construcciones extraídas del corpus. Cuervo, por tanto, no estableció una asociación directa entre una definición lexicográfica y una acepción documentada, sino que, a partir de un significado inicial de carácter general, que definía de manera muy amplia el significado de una entrada polisémica, fue añadiendo progresivamente precisiones semánticas en la definición de las distintas acepciones, especificando los constituyentes de cada construcción que determinaban el

significado específico de cada acepción, realizando así una integración de la semántica y la sintaxis en la caracterización lexicográfica de sus entradas, una aproximación que es comparable con los planteamientos de la gramática de construcciones (Fillmore, Lee-Goldman y Rhodes 2012) dentro de la lingüística cognitiva.

La documentación sistemática de la investigación lingüística mediante la utilización de corpus textuales no se pudo establecer como una metodología generalizada hasta que la informática no se consolidó como un instrumento imprescindible en la investigación lingüística. De hecho, la utilización de corpus de textos sin medios informáticos, tal como hizo Cuervo a finales del siglo XIX para realizar su *Diccionario*, resultaría hoy una tarea tan laboriosa, que sería difícilmente realizable. A pesar de lo que hemos señalado, en el año 1942 el Instituto Caro y Cuervo de Bogotá (Colombia) decidió terminar la labor lexicográfica iniciada por Cuervo, quien falleció en 1911. Esta gigantesca tarea terminó en 1994, y en 1998 se publicó la edición completa del *Diccionario* en 8 volúmenes (Cuervo 1998), edición que, en 2002, se publicó además en CD-ROM.

La incorporación de la informática en la lingüística ha sido un proceso paulatino, ya que, en sus inicios, tuvo una aplicación casi exclusiva en el ámbito del cálculo científico. Posteriormente, la informática se extendió a otras áreas de conocimiento y fue entonces cuando empezó a tener una progresiva incidencia en la lingüística. La informática no sólo ha afectado las formas de transmitir y almacenar la información lingüística, sino que ha impulsado el desarrollo de nuevas áreas de investigación como el procesamiento formal del lenguaje y, en la actualidad, el procesamiento semántico automático. No creo que sea exagerado afirmar que los cambios introducidos por la informática en el ámbito del uso, la transmisión y el análisis del lenguaje son comparables o incluso mayores que los que produjo en su momento el descubrimiento de la imprenta.

Una de las primeras aplicaciones de la informática a la lingüística en el ámbito hispánico fue el análisis estadístico del léxico del español que realizó Juilland y Chang (1964), partiendo de un corpus de 500.000 palabras. Este primer análisis automático del léxico desde un punto de vista estadístico puso de manifiesto el potencial de la informática en su aplicación al estudio del léxico, así como la extraordinaria utilidad de la utilización de corpus de textos para la investigación lingüística. Las aplicaciones posteriores de la informática en la lingüística se centraron en el desarrollo de sistemas de etiquetación léxica, es decir, programas que permitían automatizar (1) el análisis de la morfología flexiva de las formas verbales, nominales y adjetivas, así como (2) la lematización automática, es decir, la asociación de un lema a cada una de las formas léxicas, tanto flexivas como no flexivas. Los primeros sistemas de etiquetación léxica constituyeron el punto de partida para el desarrollo de los sistemas extracción automática de información léxica y sintáctica de los corpus de textos. La etiquetación léxica se llevó a cabo inicialmente con sistemas estadísticos, que son los más extendidos en la actualidad. Pero también se desarrollaron etiquetadores léxicos, que utilizaban diccionarios electrónicos, que incluían tanto las formas no flexivas del léxico (adverbios, etc.), como las formas flexivas de los verbos, los nombres y los adjetivos. En dichos sistemas, las formas flexivas se generan automáticamente, a partir de diccionarios de lemas con códigos asociados, que remiten a los programas de generación de formas flexivas a ficheros codificados, que formalizan la flexión morfológica, y es a partir de la información contenida en dichos ficheros cómo se generan automáticamente todas las formas flexivas del léxico.

Uno de los primeros etiquetadores léxicos basado en un diccionario electrónico fue desarrollado por Casajuana, Rodríguez, Sopena y Villar (1987)[3] en el Centro de Investigación de la Universidad Autónoma de Madrid/IBM (Marcos Marín 2009: 393). En dicho sistema a partir de un diccionario de 35.000 lemas se generaba un diccionario de 400.000 formas, que se utilizaba como base de información para la etiquetación léxica. Otro sistema análogo de etiquetación basado en diccionarios se desarrolló en el marco de un proyecto de investigación en la Universidad Autónoma de Barcelona (Subirats 1987, 1989, 1992, 1994a, 1994b). La primera versión de este diccionario electrónico del español contenía 500.000 formas, que se generaban automáticamente a partir de un diccionario que contenía 65.000 lemas. En el siguiente apartado, veremos el desarrollo actual de este diccionario y su integración en un sistema de diccionarios, que incluye tanto formas simples como locutivas. Asimismo, analizaremos los programas de etiquetación léxica automática del español que utilizan dicho sistema. La ventaja de los etiquetadores léxicos que utilizan diccionarios electrónicos es (1) la precisión de sus resultados, (2) la posibilidad de controlar y corregir los errores de la etiquetación, introduciendo cambios en el diccionario de formas de base o, incluso, añadiendo nuevas entradas en dicho diccionario, cuando ello es necesario, y finalmente, (3) la posibilidad de combinar dichos sistemas con etiquetadores estadísticos.

En los próximos apartados, queremos analizar el proceso de etiquetación del corpus utilizado en el proyecto FrameNet español, así como la utilización de dicho corpus etiquetado para la extracción automática de construcciones sintácticas para estudiar la proyección de la semántica en la sintaxis.

3. Estado de la cuestión

El corpus del proyecto de investigación FrameNet Español[4], que contiene 1.000 millones de palabras, se ha etiquetado léxicamente mediante la utilización de un diccionario electrónico del español[5], que contiene 634.503 formas[6], las cuales se generan automáticamente a partir de un diccionario de 113.825 lemas, que está integrado por los siguientes componentes:

1) un diccionario de 86.104 lemas simples, como p. ej., *unir*, *inmoralidad*, *allí,* etc., (Subirats 1989, 1992, 1994a) y
2) un diccionario de 25.721 lemas locutivos, que se expanden automáticamente en un diccionario de locuciones flexivas de 55.000 formas; dicho diccionario incluye las siguientes clases de palabras:
   a) locuciones nominales, p. ej., *muerte cerebral*, *carga de profundidad*, *ácido graso no saturado*, etc., (Subirats 1994b);
   b) locuciones adjetivas, *azul marino, sano y salvo,* y locuciones adjetivas comparativas con verbo de soporte, *(estar) fresca como una rosa* (Garrido 1999);
   c) locuciones adverbiales, *poco a poco*, *de una vez por todas,* etc.; y
   d) grupos preposicionales predicativos que admiten verbos de soporte, *(estar) en peligro, (estar) de moda, (ser) de armas tomar* (Mogorrón 1994);

Además de los diccionarios electrónicos de formas simples y locutivas, el sistema de etiquetación léxica incluye 3.009 transductores (Ortega 2010), como los de la Fig. 2, que formalizan las características lexicosintácticas de 2.043 locuciones verbales, como p. ej., *dar por sentado, tener en cuenta*, *tener lugar* (Bobes 2000), y locuciones verbales con estructuras comparativas, *aburrirse como una ostra* (Garrido 1999).

Cada una de las formas simples y locutivas del diccionario electrónico de formas expandidas van acompañadas:

1) del lema al que pertenecen,
2) de etiquetas[7] que proporcionan información sobre su pertenencia a una –o más– clases de palabras[8] (nombre, verbo, adjetivo, etc.), y
3) de las siguientes propiedades morfológicas flexivas[9]:
   a. el tiempo (presente, pretérito, futuro, etc.), el modo (indicativo, subjuntivo), la persona (1ª, 2ª, 3ª) y el número (singular, plural) de los verbos, y el género (masculino, femenino) y el número de los participios verbales;
   b. el número de los nombres y adjetivos, y el género, en el caso de los nombres o adjetivos que tienen flexión de género.

El educto o salida del etiquetador léxico es un conjunto de autómatas –un autómata por cada oración (Fig. 1)–, cuyas transiciones están etiquetadas con la información que posee el diccionario electrónico, de cada una de las formas léxicas simples y locutivas que contiene dicha oración (Ortega 2010).

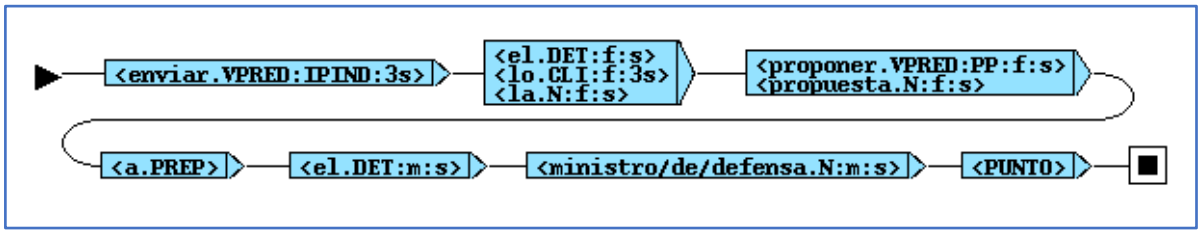

Fig. 1. Representación gráfica del autómata resultante de la etiquetación léxica automática de la oración *Envió la propuesta al ministro de defensa*.

Las ambigüedades que genera el proceso de etiquetación léxica se representan gráficamente en el autómata como diferentes transiciones entre dos estados. Así p. ej., en la Fig.1, partiendo del estado inicial, que se representa mediante un triángulo, el rectángulo correspondiente a la segunda transición del autómata, muestra las ambigüedades de la forma *la*, que está etiquetada como (1) un determinante femenino singular, asociado al determinante *el*, (2) un pronombre clítico femenino en 3ª persona del singular, asociado al clítico *lo*, o (3) un nombre femenino singular, asociado al lema nominal *la*, que denota una nota dentro de la escala musical. Análogamente, la tercera transición del autómata de la Fig. 1 muestra las ambigüedades de la palabra *propuesta*, que está etiquetada como (1) un participio verbal femenino singular del verbo *proponer* o (2) el nombre femenino singular *propuesta*. Ambigüedades como las que acabamos de analizar se pueden resolver intersectando los autómatas resultantes de la etiquetación léxica, como p. ej., en la Fig. 1, con transductores que especifiquen condiciones locales de desambiguación, no sólo condiciones distribucionales positivas que especifiquen de combinaciones posibles de clases de palabras con determinadas características flexivas, etc., sino también con condiciones negativas, es decir, teniendo en cuenta aspectos distribucionales locales, que no se pueden dar en un determinado contexto.

Como hemos señalado anteriormente, las locuciones verbales se identifican y se etiquetan intersectando los autómatas[10] que constituyen la salida del etiquetador léxico (Fig. 1), con 3.009 transductores, que especifican las propiedades léxicas y sintácticas relevantes desde el punto de vista de su identificación y de su etiquetación como unidades léxicas. Estos transductores recogen propiedades sintácticas de las locuciones verbales, que son imprescindibles para su reconocimiento, el cual, como veremos a continuación, puede ser una

forma indirecta de eliminar ambigüedades de la etiquetación léxica. P. ej., el reconocimiento de una locución verbal, en los casos en los que dicha locución tiene pronombres clíticos, adverbios u otros elementos léxicos entre el núcleo verbal y su parte fija o conexa, como en la Fig. 3, se puede llevar a cabo intersectando el autómata resultante de la etiquetación léxica (Fig. 3) con el transductor de la Fig. 2, porque dicho transductor especifica la posición y el tipo de material léxico que puede aparecer entre el núcleo verbal, *dar*, y la parte conexa, *por sentado*, de la locución *dar por sentado*. Por ello, el transductor de la Fig. 2 puede reconocer cada una de las formas que integran *dar por sentado* y puede convertir dichas formas en un único lema verbal locutivo, representado gráficamente por una transición única en el autómata resultante de la transducción de la Fig. 4. Asimismo, dentro del mismo proceso de transducción, el adverbio *siempre* se desplaza a la derecha de la locución verbal. Además, en todo este proceso, la transducción elimina la ambigüedad de las formas que integran la locución, concretamente, la ambigüedad que afecta al verbo *dar*, que puede ser (1) un verbo de soporte o (2) un verbo predicativo y, asimismo, la ambigüedad que afecta a *sentado*, que puede ser (1) un adjetivo predicativo, (2) un participio vinculado al verbo *sentarse*, o (3) un participio asociado al lema verbal *sentar*. Tras la transducción léxica (Fig. 4), *dar* y *sentado* dejan de ser formas ambiguas como en la Fig. 3, porque la secuencia formada por las palabras *dar*, *por* y *sentado* se han convertido en componentes de una única unidad léxica locutiva *dar por sentado*, y, por tanto, las palabras –o formas– que integran dicha locución están integradas en una única transición del autómata de la Fig. 4. Por tanto, después de la transducción se eliminan todas las ambigüedades relacionadas con las formas que integran la locución verbal *dar por sentado*. Esta forma indirecta de eliminar ambigüedades se da en todos los casos en los que se detectan automáticamente elementos léxicos locutivos.

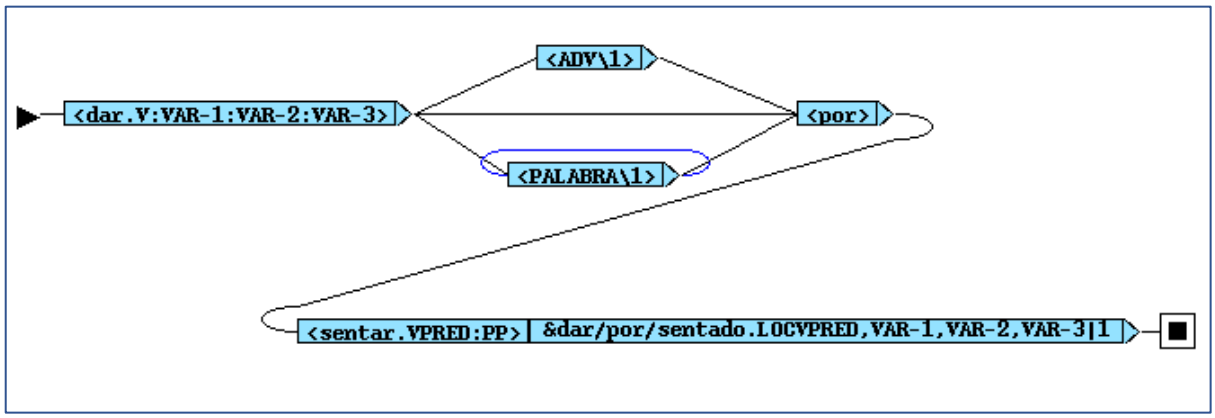

Fig. 2. Representación gráfica del transductor que permite identificar la locución verbal *dar por sentado*.

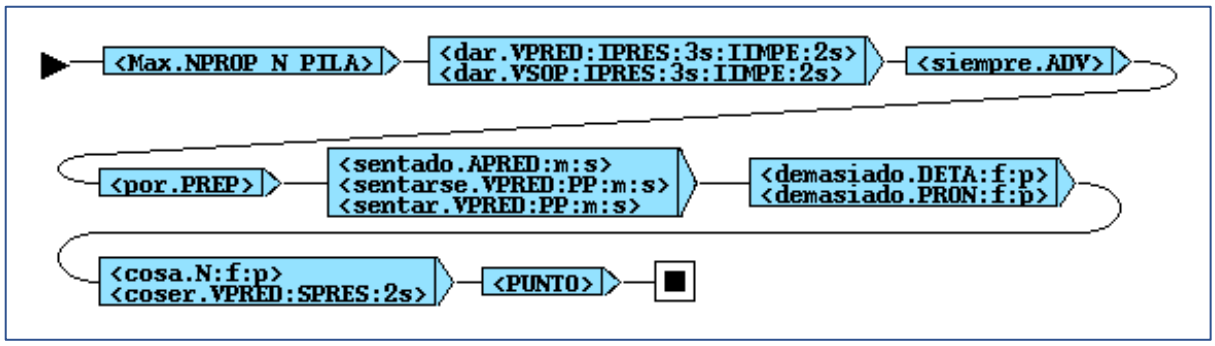

Fig. 3. Autómata resultante de la etiquetación léxica de la oración *Max da siempre por sentado demasiadas cosas* antes de su intersección con el transductor de la Fig. 2.

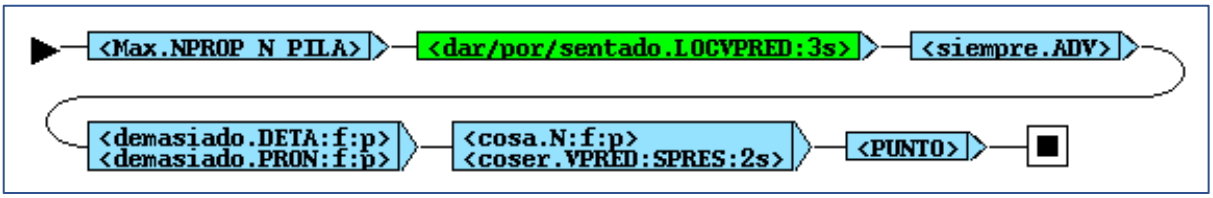

Fig. 4. Autómata resultante de la intersección del autómata de la Fig. 3 con el transductor de la Fig. 2. La locución verbal reconocida aparece en el recuadro de color verde.

La cobertura léxica de los diccionarios electrónicos de formas simples y locutivas ha sido verificada sobre el corpus utilizado en el proyecto FrameNet Español, corpus que contiene 1.000 millones de palabras. Este proceso nos ha permitido incluir entradas que no se encuentran en los diccionarios tradicionales del español, como p. ej., el *Diccionario de la lengua española* de la Real Academia Española, que, a pesar de sus enormes deficiencias (Subirats 2007), ha constituido tradicionalmente el punto de referencia de toda la lexicografía española. Asimismo, la utilización del corpus como forma de verificación del diccionario electrónico nos ha permitido también eliminar entradas que aparecen a menudo en los diccionarios tradicionales, pero que, en realidad, no se usan en la lengua española actual (ya sea oral o escrita). En suma, la inclusión de formas léxicas utilizadas en la lengua actual, que no figuran en los diccionarios tradicionales, junto con la eliminación de palabras que no se usan en la actualidad, aunque aparezcan en las obras de la lexicografía tradicional, nos ha permitido adaptar el diccionario electrónico a las características reales de la lengua española actual.

4.1. Consideraciones metodológicas: semántica y sintaxis

El proceso de identificación de las locuciones verbales (LocV[11]), como hemos visto en el aparado anterior en relación con la LocV *dar por sentado*, tiene muchos aspectos en común con el reconocimiento de los grupos preposicionales predicativos (GPpred) con verbo de soporte, p. ej., *estar de moda*, y también de los nombres y adjetivos predicativos (Npred y Apred) con verbo de soporte (Vsop), p. ej., *hacer una solicitud*, *estar cansado*, puesto que tanto (1) las LocV, como (2) los GPpred, los Npred y los Apred, requieren una identificación de las formas que los integran, en función de sus propiedades combinatorias con otros constituyentes. En efecto, los GPpred, los Npred y los Apred con Vsop –de forma análoga a lo que ocurre con las LocV– admiten material léxico entre el GPpred, el Npred y el Apred, y su Vsop correspondiente, p. ej., *estar ahora de moda*, *hacer siempre peticiones inútiles, estar en este momento ocupado*, donde los adverbios *ahora*, *siempre*, y el grupo preposicional con valor adverbial *en este momento*, se encuentran, respectivamente, entre el GPpred, el Npred y el Apred, y su Vsop correspondiente. Sin embargo, a pesar de sus similitudes, los GPpred, los Npred y los Apred reciben un tratamiento distinto al que reciben las LocV, lo cual está motivado por el hecho de que los GPpred, los Npred y los Apred con Vsop requieren un tratamiento menos particularizado desde el punto de vista léxico que las LocV, puesto que sus propiedades combinatorias con otros elementos léxicos, fundamentalmente, pronombres clíticos, adverbios y sintagmas con valor adverbial, son muy similares, puesto que dichos elementos léxicos aparecen siempre en el mismo lugar, concretamente, entre el Vsop y su núcleo predicativo. Precisamente por ello, es posible construir gramáticas locales en forma de transductores de aplicación general para el reconocimiento de los GPpred, los Npred y los Apred con Vsop, que posibiliten el reconocimiento de dichas unidades léxicas y sus respectivos verbos de soporte, sin tener que hacer transductores individualizados como en el caso de las LocV. Es cierto, no obstante, que la creación de transductores que tengan en

cuenta el Vsop de las clases de predicados no verbales mencionados, es decir, GPpred, Npred y Apred, potenciaría aún más la precisión de su reconocimiento automático.

A partir de las cuestiones que hemos abordado hasta ahora en relación con el reconocimiento automático del léxico locutivo con propiedades léxico-semánticas específicas, se pone de manifiesto además que no existe una frontera precisa entre el léxico y la sintaxis, sino que lo que realmente existe es una gradación –tal como ha puesto de manifiesto la gramática de construcciones–, gradación que empezaría con las formas léxicas simples, continuaría con las unidades léxicas predicativas –simples o locutivas– con Vsop y las LocV, y terminaría con las construcciones sintácticas propiamente dichas. Precisamente por el hecho de haber una gradación de estas características, no existe una diferencia en los algoritmos que utilizan los programas de etiquetación léxica y los analizadores sintácticos que se utilizan para reconocimiento de constituyentes sintácticos suboracionales, como p. ej., grupos nominales y preposicionales, núcleos verbales con argumentos nominales, etc. Por ello, los analizadores sintácticos utilizan los mismos algoritmos de intersección de autómatas con transductores que los programas de etiquetación léxica. La diferencia entre los etiquetadores léxicos y los analizadores sintácticos reside en el hecho de que, en general, estos últimos utilizan transductores, que recogen las características de construcciones sintácticas específicas, formuladas como combinaciones de clases de palabras que no están necesariamente ligadas a propiedades léxico-sintácticas de unidades léxicas específicas. Así p. ej., en la Fig. 5, podemos observar un autómata, que recoge las características distribucionales de los grupos nominales del español y que permite, por tanto, reconocer grupos nominales de un determinado nivel de complejidad.

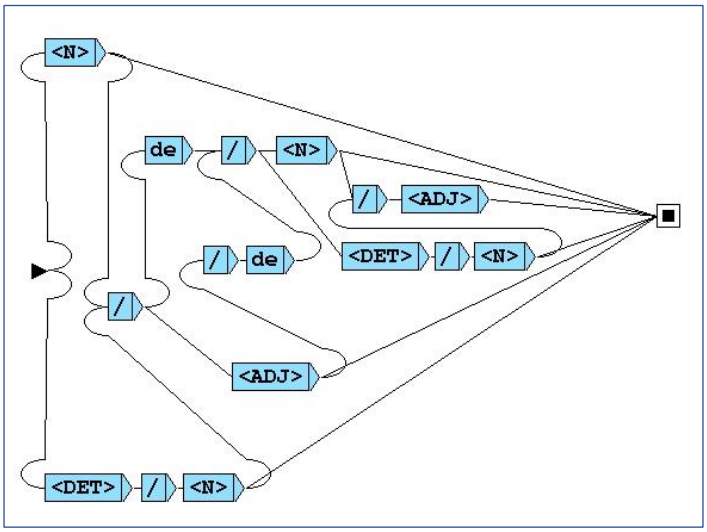

Fig. 5. Representación gráfica en forma de autómata de la expresión regular (<E> + <DET>/) <N> (<E> + /<ADJ>) (<E> + /de/ (<E> + <DET>/) <N> (<E> + /<ADJ>)

4.2. Consideraciones teóricas y metodológicas

En este apartado, queremos presentar algunos conceptos fundamentales de la semántica de marcos, conceptos que nos permitirán comprender la integración de la semántica en la sintaxis, que es una de las características teóricas fundamentales de la lingüística cognitiva y, también, una característica de algunos aspectos del sistema de procesamiento que presentamos en este artículo. En primer lugar, hay que partir de la base de que el objetivo

fundamental de la semántica y la sintaxis cognitivas es realizar un análisis semántico-sintáctico del léxico, partiendo de la teoría de la semántica de marcos (Fillmore 1976, 1982, 1985) y de la teoría de la gramática de construcciones (Fillmore et al. 2012). La teoría de la semántica de marcos, al igual que todas las propuestas semánticas dentro de la lingüística cognitiva, rechaza frontalmente la concepción referencialista del significado, es decir, rechaza la reducción del significado a una mera cuestión de referencia del lenguaje a las entidades de una supuesta realidad "objetiva", externa e independiente del hablante. En contraposición con la semántica referencialista, la semántica de marcos parte de la base de que el significado es una *construcción*, que se crea a partir de los marcos semánticos que evocan las unidades léxicas. Un marco semántico es una esquematización de un situación o escenario, en el que hay un conjunto de participantes, que constituyen los roles semánticos de dicho marco. Los marcos semánticos suelen incluir un conjunto de predicados, verbales, nominales, adjetivos, etc., que evocan el mismo marco, aunque focalicen aspectos distintos dentro de él. Analicemos algunos ejemplos. El marco semántico `Request`[12] (petición) incluye predicados que evocan una situación, en la que un SPEAKER (locutor) se dirige a un ADDRESSEE (destinatario) para transmitirle un MESSAGE (mensaje), en el que el SPEAKER le pide al ADDRESSEE que realice una acción. Así p. ej., los verbos *pedir*, *ordenar*, *suplicar* y *reclamar* entre otros, evocan todos ellos el marco semántico `Request`:

(1) *Le **pedimos** que se fuera para siempre.*
(2) *El capitán nos **ordenó** abandonar el barco.*
(3) *Les **suplicó** en vano que no le mataran.*
(4) *Les **reclamé** el envío de la factura.*

Sin embargo, cada uno de los cuatro predicados señalados, aunque evocan todos ellos la situación que activa el marco `Request`, destacan aspectos distintos dentro de dicho marco. Así p. ej., en (1), un SPEAKER, que tiene una realización nula, formula una petición a un ADDRESSEE, *le*, para comunicarle que un MESSAGE, en el que le solicita al ADDRESSEE que realice una acción determinada, *que se fuera para siempre*. Pero *pedir*, a diferencia de *ordenar*, como veremos a continuación, no denota si existe o no una relación jerárquica entre el SPEAKER y el ADDRESSEE. Por el contrario, en el caso de *ordenar* en (2), el verbo denota una relación de superioridad jerárquica del SPEAKER, *el capitán*, sobre los ADDRESSEES, *nos*, relación que le permite al SPEAKER conminar u obligar a los ADDRESSEES a realizar su petición, *abandonar el barco*. Asimismo, mientras *suplicar* en (3) resalta la humildad con la que se realiza una petición, *pedir* no lexicaliza la manera humilde o no de realizar dicha petición. Análogamente, *reclamar* en (4) denota el derecho que el SPEAKER posee –o considera que posee– para poder realizar una petición, mientras que *pedir* en (1) es completamente neutro en relación con la existencia o no de un derecho por parte del locutor para realizar una petición determinada.

El proyecto FrameNet español se centra en el estudio de los marcos semánticos que permiten analizar cognitivamente el significado que evocan las unidades léxicas que conforman el léxico. Y, además de ello, analiza también la proyección sintáctica de los roles semánticos del marco que evocan los predicados estudiados en construcciones sintácticas específicas. En la práctica, estudiar la proyección sintáctica de los roles semánticos implica dos tareas: en primer lugar, buscar previamente en un corpus las ocurrencias de las oraciones en las que aparece un determinado predicado y, en segundo lugar, analizar la proyección de los roles semánticos en la sintaxis. En el caso de los predicados verbales, se trata de identificar en el corpus los distintos tipos de argumentos verbales en los que se proyectan los roles, así como las distintas preposiciones que pueden introducir dichos argumentos verbales. En el caso de los predicados no verbales, como p. ej., nombres o adjetivos predicativos, etc., se trata de

determinar (1) los distintos tipos de verbos de soporte que seleccionan, (2) las características de los argumentos preposicionales dependientes de dichos predicados no verbales, y (3) el tipo de dependencia sintáctica que mantienen dichos argumentos preposicionales con su predicado, p. ej., objeto preposicional, complemento adjunto, etc. Asimismo, también hay que precisar la posibilidad de que un predicado, independientemente de su naturaleza verbal o no verbal, pueda tener argumentos oracionales o la posibilidad de que dichos argumentos oracionales puedan ser o no complementos infinitivos, que pueden, a su vez, compartir roles semánticos con el predicado del que dependen.

Los roles semánticos del marco que evoca un predicado se proyectan sintácticamente de muy diversas formas. Así p. ej., los tres roles semánticos nucleares, SPEAKER, ADDRESSEE y MESSAGE del marco semántico `Request`, que evocan determinados predicados de comunicación, se realizan sintácticamente de formas muy diversas en el caso de las distintas unidades léxicas que evocan dicho marco semántico y estas diferencias de realización de los roles semánticos se mantienen incluso dentro de predicados pertenecientes a la misma clase de palabras, p. ej., verbos, nombres o adjetivos predicativos. El objetivo del proyecto FrameNet es tratar de determinar las características sintáctico-semánticas de esta proyección, de forma que se pueda dar una caracterización de esta proyección dentro de la gramática de construcciones, evitando, por supuesto, las simplificaciones de los modelos formalistas. El propósito es analizar las características semántico-sintácticas de la proyección de la semántica en la sintaxis, tratando de determinar qué partes de la construcción contribuyen a expresar un significado determinado, tal como propone la gramática de construcciones y tal como hizo Cuervo (1886, 1993) en su *Diccionario*, con objeto de diferenciar las distintas acepciones de las entradas polisémicas (cf. el apartado 1. en este artículo).

Analicemos algunos ejemplos. El rol semántico SPEAKER se puede proyectar sintácticamente de muy distintas formas, poniendo así de manifiesto que no existe una forma univoca de realizar esta proyección, como hemos señalado anteriormente. Así p. ej., en el siguiente ejemplo documentado en el corpus de oraciones anotadas semántica y sintácticamente de FrameNet Español, el rol semántico SPEAKER, *el primero que*, perteneciente al marco semántico `Request`, marco que evoca el nombre eventivo *solicitud*, se realiza como el sujeto del verbo de soporte de *solicitud*, es decir, *hizo*:

> *"Diversas autoridades colombianas han reconocido que Perafán probablemente será extraditado a Estados Unidos, debido a que allí tiene los cargos más graves y fue ese país [el primero que]SPEAKER.EXT hizo la* ***solicitud*** *respectiva a las autoridades venezolanas."*

Sin embargo, en construcciones sin verbo de soporte, como en el siguiente ejemplo documentado, el rol SPEAKER se proyecta en un grupo preposicional dependiente de *solicitud*, *de los acusados*, que es, sintácticamente, un complemento adjunto introducido por la preposición *de*:

> "[…] *denegó la* ***solicitud*** *[de los acusados]SPEAKER/COMP.ADJ de desestimar varios de los cargos en su contra* […]"

En cambio, en el siguiente ejemplo, el rol SPEAKER se realiza en el grupo preposicional *por el gobierno de Estados Unidos*, un complemento agente de la construcción pasiva, cuya forma verbal pasiva, *hecha*, es un participio del verbo de soporte *hacer*:

*"La declaración del presidente responde a la **solicitud** hecha la semana pasada [por el gobierno de Estados Unidos]SPEAKER/COMP.AGENTE ."*

A su vez, en el siguiente ejemplo, el rol SPEAKER, *por parte de los jugadores cubanos*, se realiza como un complemento adjunto dependiente de *solicitud*, introducido por la preposición *por*:

"*Nosotros no hemos recibido hasta el momento ninguna **solicitud** [por parte de los jugadores cubanos] SPEAKER/COMP.ADJ* […]"

Por el contrario, en el siguiente caso, el rol SPEAKER se realiza como un adjetivo posesivo, *su*, que es un modificador de *solicitud* y su argumento externo:

"*Irak dijo que ha cumplido sus obligaciones de revelar sus secretos nucleares y pidió a los delegados asistentes a la conferencia de la AIEA que respalden [su]SPEAKER/EXT **solicitud** de poner fin a las sanciones económicas de la ONU*."

Finamente, el rol semántico SPEAKER se proyecta en un adjetivo, *española*, que es un modificador de *solicitud:*

"[...] *se espera que la Comisión Europea atienda la **solicitud** [española]SPEAKER/MOD de recalificar la zona del Bajo Llobregat como de "objetivo 2" o zona en declive."*

En contraposición con la diversidad de construcciones en las que se realiza la proyección sintáctica del rol semántico SPEAKER, el rol ADDRESSEE del marco `Request` presenta una variación mucho menor: se proyecta (a) en *a las autoridades venezolanas*, un objeto indirecto del verbo de soporte *hacer* en la oración (5), y (b) en *al Gobernador de Virginia, George Allen*, un objeto preposicional dependiente de *solicitud* en (6), donde *solicitud* aparece sin verbo de soporte y como un argumento de *dirigió*:

(5) "[…] *fue ese país, el primero que hizo la **solicitud** respectiva [a las autoridades venezolanas]*ADDRESSEE/OI"
(6) "[…] *dirigió una **solicitud** [al Gobernador de Virginia, George Allen]* ADDRESSEE/ OPREP*, para que conmutara la sentencia del reo o detuviera la ejecución* […]"

También el rol semántico MESSAGE presenta una proyección sintáctica más restringida que el rol SPEAKER, ya que sólo se realiza como un objeto preposicional, introducido por la preposición *de*, que tiene las siguientes características: (a) en (7), es una locución, *de asilo político*; (b) en (8), es una oración subordinada sustantiva con el verbo en subjuntivo, *de que sea suspendido el proceso de deportación*; y (c) en (9), es un complemento infinitivo, *de aumentar la partida de financiación al Fondo*:

(7) *"[…] se estudia las **solicitudes** [de asilo político]MESSAGE/O.PREP de ambos […]"*

(8) *"Un juez federal estudiará mañana la **solicitud** de Ruiz Massieu [de que sea suspendido el proceso de deportación, cuya audiencia tendrá lugar el viernes por parte de un juez de inmigración]MESSAGE/OPREP"*

(9) *"[…] el Congreso aprobará la **solicitud** de la Administración [de aumentar la partida de financiación al Fondo]MESSAGE/OPREP […]"*

El objetivo de este análisis semántico-sintáctico en el léxico permitirá la determinación precisa de cuáles son las partes de las distintas construcciones en las que se proyectan los roles semánticos pertenecientes a marcos semánticos específicos y permitirá establecer así una relación entre el significado que aporta cada parte de una construcción sintáctica, en relación con la proyección sintáctica de los roles semánticos de los marcos que evocan las unidades léxicas predicativas del léxico.

4.3. Consideraciones metodológicas: extracción automática de construcciones

La extracción automática de oraciones del corpus de FrameNet español para su posterior anotación semántica y sintáctica requiere la realización de dos procesos consecutivos. En primer lugar, es necesario identificar en un corpus las construcciones en las que se proyectan los distintos roles semánticos pertenecientes al marco semántico evocado por la unidad léxica analizada. En segundo lugar, se debe extraer automáticamente del corpus, mediante gramáticas sintagmáticas en forma de transductores, las construcciones identificadas y seleccionadas previamente a partir del análisis de los ejemplos del corpus. Dada la diversidad de construcciones en las que se proyectan los roles semánticos pertenecientes a los marcos que evocan las unidades léxicas analizadas (cf. 5. en este artículo), la creación de transductores para realizar la extracción automática de construcciones sintácticas del corpus se tiene que realizar de manera particularizada para cada unidad léxica estudiada. La extracción de oraciones del corpus se realiza de forma automática, pero para poderla automatizar se requiere la construcción de transductores, lo cual resulta un proceso muy laborioso, si se efectúa de forma enteramente manual con un editor. La laboriosidad de la construcción manual de transductores para la realización de transducciones sintácticas se debe fundamentalmente al hecho de que los transductores son formalismos más complejos que los autómatas, dado que tienen una parte de "lectura" –al igual que los autómatas–, pero, además, tienen una parte de "escritura", que especifica los cambios que se tienen que realizar en la cadena reconocida, es decir, los cambios que tiene que realizar la transducción para generar el autómata de salida. Por ello, con objeto de facilitar la tarea de creación de transductores, el proyecto FrameNet Español creó la aplicación Gramcreator, que permite automatizar una parte del proceso de creación de transductores, que están destinados a la extracción automática de oraciones del corpus (Fig. 6).

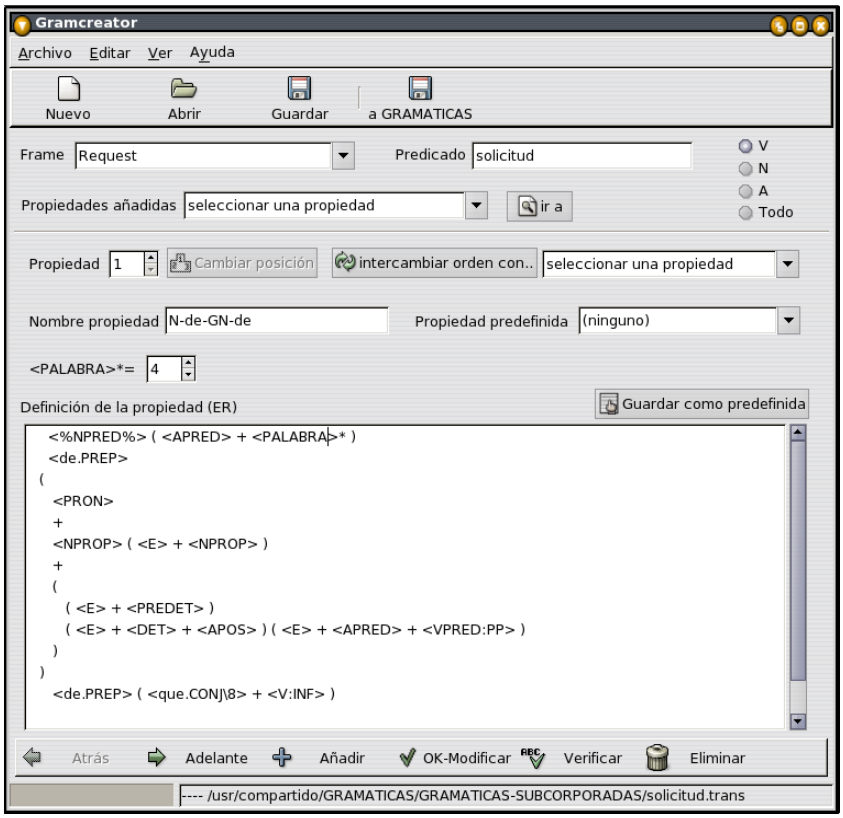

Fig. 6. Gramcreator, aplicación para crear transductores semiautomáticamente para realizar la extracción automática de construcciones del corpus, partiendo de criterios semántico-sintácticos**.**

A pesar de la complejidad de la proyección de la semántica en la sintaxis, las distintas clases de palabras predicativas, como nombres, verbos, adjetivos, etc., presentan algunas características sintácticas en común en relación con las construcciones en las que pueden aparecer, es decir, la clase de palabras a la que pertenece un predicado determina en parte los tipos de construcciones sintácticas en los que puede aparecer. Por ello, para facilitar el proceso de extracción de oraciones del corpus, la biblioteca de transductores que posee Gramcreator está organizada en función de la clase de palabras a la que pertenece el predicado analizado, es decir, verbo, nombre, adjetivo, etc. Así, cuando el usuario de la aplicación necesita un transductor para realizar un determinado tipo de extracción sintáctica del corpus, no tiene que crear el transductor correspondiente, sino que puede usar un transductor de la biblioteca y, en el caso de que el transductor seleccionado no tenga el nivel de especificidad deseado, el usuario sólo tiene que introducir los cambios necesarios para adaptarlo a las características específicas de su extracción y, si lo considera útil, puede incluso guardar la versión modificada del transductor en la biblioteca. Asimismo, Gramcreator tiene también tiene un modo de funcionamiento manual, que permite crear y archivar nuevos transductores. Para facilitar más aún la creación de un nuevo transductor, Gramcreator no requiere que el usuario especifique la parte de escritura del transductor, sino únicamente, la parte de lectura. En la práctica, eso implica que el usuario crea un autómata con el editor y Gramcreator lo convierte automáticamente en un transductor. Por todo ello, Gramcreator reduce sensiblemente la laboriosidad de creación de transductores para la extracción de oraciones asociada a las unidades léxicas previamente seleccionadas para su anotación semántica y sintáctica.

El interés del procedimiento de extracción de construcciones que describimos no reside únicamente en las ventajas prácticas de su funcionamiento, sino en el hecho de que nos permite unificar el proceso de extracción de construcciones sintácticas con el significado que aportan dentro de su contexto oracional, de manera que lo que aparentemente podría parecer una extracción automática de estructuras formales es, en realidad, una extracción de construcciones en función de sus características semánticas, que están ligadas al rol semántico que se proyecta en dichas construcciones. Por todo ello, el proceso de extracción de oraciones del corpus constituye una forma de unificar la semántica con la sintaxis y crear así proceso semántico-sintáctico de extracción automática de construcciones.

5. **Conclusiones y desarrollos futuros**

En este artículo, hemos analizado dos aspectos, que hemos considerado cruciales para la consolidación de la lingüística de corpus: por un lado, la voluntad explícita de la lingüística cognitiva de estudiar el lenguaje a partir de enunciados producidos en contextos comunicativos reales y, por otro lado, los avances de la lingüística computacional en el desarrollo de herramientas para el procesamiento automático de textos y la existencia de corpus de dominio público de miles de millones de palabras. Por otro lado, hemos analizado un sistema de etiquetación léxica, basado en la utilización de diccionarios electrónicos de formas simples y locutivas del léxico general, y hemos explicado cómo la salida de este etiquetador sirve de base para la realización de una extracción automática de construcciones sintácticas mediante un proceso de intersección de un autómata, que es el resultado de la etiquetación léxica, con un transductor, que determina las características de una construcción sintagmática, como p. ej., un determinado tipo de sintagma nominal, verbal, etc. Finalmente, hemos mostrado de qué manera la semántica de marcos nos permite realizar extracciones automáticas de construcciones sintácticas, que pueden ser utilizadas posteriormente para estudiar la proyección sintáctica de los roles semánticos pertenecientes al marco semántico que evoca el predicado analizado. Finalmente, hemos analizado la vinculación del planteamiento semántico-sintáctico de la sintaxis cognitiva con la labor lexicográfica realizada por Cuervo (1896, 1893), quien caracterizó el significado de las distintas acepciones de las entradas de su *Diccionario de construcción y régimen* en relación con las características semánticas de las construcciones en las que aparecían, apoyándose para ello en la documentación de su análisis a partir de ejemplos extraído de un corpus de textos.

La investigación futura en el ámbito de las áreas que hemos abordado en este artículo podría extenderse en seis direcciones distintas e igualmente importantes, ya que todas ellas están relacionadas entre sí y van a determinar el desarrollo orgánico de una nueva concepción lingüística, que estudiará la sintaxis como una proyección de la semántica, y contribuya al desarrollo de sistemas de procesamiento semántico automático de las lenguas naturales.

En primer lugar, sería necesario ampliar la cobertura léxica de las bases de datos de oraciones y textos etiquetados semántica y sintácticamente, siguiendo los planteamientos de la semántica de marcos, lo que, en la práctica, implicaría ampliar el número de unidades léxicas analizadas en los proyectos integrados en Global FrameNet[13]. Asimismo, sería necesario aumentar el número de lenguas integradas dentro del mencionado proyecto global, lo cual permitiría crear una base de conocimiento que posibilitaría realizar una comparación semántica de las distintas lenguas estudiadas, lo que a su vez daría lugar a un análisis detallado de sus características semánticas diferenciales. En segundo lugar, sería importante desarrollar la labor iniciada con el *constructicon*[14], desarrollado en el marco del proyecto FrameNet de Berkeley, que consistió en un estudio piloto de cien construcciones sintácticas,

analizadas desde el punto de vista de la gramática de construcciones. Sólo a partir de estudios sintácticos multilingües basados en la gramática de construcciones se podrá llegar a una comprensión mas detallada de la contribución de la sintaxis en el proceso de la construcción del significado. En tercer lugar, es necesario integrar la semántica de marcos y la gramática de construcciones en los nuevos programas de análisis semántico de textos, como p. ej., el que propone Bryant (2008), quien presenta un sistema de análisis textual, que combina e integra el análisis semántico basado en esquemas de imagen –que se pueden convertir fácilmente en marcos semánticos–, con el análisis de construcciones gramaticales. En cuarto lugar, sería muy interesante, tal como propone Matos (2014) en su modelo de análisis semántico, integrar la jerarquía de marcos semánticos, es decir, la red de relaciones semánticas establecidas entre marcos semánticos, dentro de la información utilizada por los sistemas de procesamiento semántico, ya que ello permitiría aumentar la precisión de los análisis semánticos automáticos. En quinto lugar, se deberían desarrollar aplicaciones lexicográficas de la semántica de marcos (Ostermann 2015), ya que ello posibilitaría una transformación radical de los diccionarios actuales, transformación que abriría nuevas vías de modernización de la lexicografía. Asimismo, también sería importante reflexionar sobre las aplicaciones de la semántica de marcos y la gramática de construcciones para la enseñanza del español como lengua extranjera, ya que un planteamiento cognitivo de la enseñanza de lenguas extranjeras podría mejorar considerablemente la eficacia de su metodología. Y, finalmente, en sexto lugar, sería importante ahondar en las raíces cognitivas de la lingüística hispánica del s. XIX y principios del s. XX, para poder relacionar la lingüística cognitiva y la lingüística de corpus, con nuestra tradición anterior, a la que podemos dar una nueva interpretación en la actualidad, a la luz de los conocimientos que ha aportado la lingüística cognitiva y la lingüística de corpus.

Referencias citadas

Bobes, E. de. 2000. *Gramática electrónica de las locuciones verbales.* Laboratorio de Lingüística Informática, Universidad Autónoma de Barcelona.

Bryant, John E. 2008. *Best-Fit Constructional Analysis*. Tesis doctoral, University of California Berkeley (California, EE.UU), http://www2.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-100.pdf

Casajuana, R. y C. Rodríguez. 1985. "Verificación ortográfica en castellano: la realización de un diccionario en ordenador". *Español Actual* 44: 5-76.

Casajuana, R. y C. Rodríguez. 1986. "Clasificación de los verbos castellanos para un diccionario en ordenador", *Actas del 1er Congreso de Lenguajes Naturales y Lenguajes Formales*. Barcelona, 221-44.

Casajuana, R., C. Rodríguez, L. Sopena y C. Villar. 1987. "Towards an Integrated Environment for Spanish Document Verification and Composition". *Proceedings of the Third Conference on European Chapter of the Association for Computational Linguistics (EACL '87)*, 52-5: https://dl.acm.org/citation.cfm?id=976867

Cuervo, R. J. 1886-1893. *Diccionario de construcción y régimen de la lengua castellana*, 2 vols. París: A. Roger y F. Chernoviz:

https://bibliotecanacional.gov.co/content/conservacion?idFichero=86745

Cuervo, R. J. 1998. *Diccionario de construcción y régimen de la lengua castellana*, 8 vols. Barcelona: Herder.

Fillmore, C. J. 1976. "Frame Semantics and the Nature of Language". En *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, vol. 280: 20-32.

Fillmore, C. J. 1982. "Frame Semantics". En *Linguistics in the Morning Calm*, ed. The Linguistic Society of Korea, 111-37. Seúl: Hanshin.

Fillmore, C. J. 1985. "Frames and the Semantics of Understanding". *Quaderni di Semantica* 6.2: 222-54.

Fillmore, C. J., R. R. Lee-Goldman y R. Rhodes. 2012. "The FrameNet Constructicon". En *Sign-Based Construction Grammar,* eds H. C. Boas y I. A. Sag. Stanford: CSLI Publications: https://pdfs.semanticscholar.org/7678/b53e512d492256c4810f6337957743362750.pdf?_ga=2.84839641.1584376741.1588345206-1335818265.1579895470

Garrido, P. 1999. *Estudio sintáctico del adverbio fijo en predicados comparativos. Estudios de Lingüística del Español* 7: http://elies.rediris.es/elies7/

Juilland, A. G. y E. Chang-Rodríguez. 1964. *Frequency Dictionary of Spanish Words*. La Haya: Mouton.

Marcos-Marín, Franciso A. 2009. "Historia humana de la lengua española y su computación". *Studies in Hispanic and Lusophone Linguistics* 2.2: 387-415: http://www.lllf.uam.es/~fmarcos/articulo/09SHLLcompu.pdf

Matos, Ely. 2014. *LUDI: Um framework para desambiguação lexical com base no enriquecimento da Semântica de Frames*. Tesis doctoral, Universidade Federal de Juiz de Fora (Minas Gerais, Brasil): https://www.researchgate.net/publication/308794122_LUDI_Um_framework_para_desambiguacao_lexical_com_base_no_enriquecimento_da_Semantica_de_Frames

Mogorrón, P. 1994. *Estudio contrastivo de las frases 'ser/estar + Prep X' en español y 'être + Prep X' en francés.* Tesis doctoral, Universidad de Valencia.

Ortega, M. 2010. "Etiquetación automática de roles semánticos en FrameNet Español". En *Analizar datos, describir variación*. *XXVIII Congreso Internacional de la Asociación Española de Lingüística Aplicada (AESLA 2010)*, eds. J. L. Bueno et al. Vigo: Universidade de Vigo, Servizo de Publicacións: http://papers.spanishfn.org/public/Ortega2010.pdf

Ostermann, Carolin. 2015. *Cognitive Lexicography: A New Approach to Lexicography Making Use of Cognitive Semantics*. Berlín: De Gruyter.

Subirats, C. 1987. “El Diccionario Electrónico del Español”. *Procesamiento del Lenguaje Natural. Número monográfico sobre las III Jornadas SEPLN*, Boletín 5: 63-72: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/download/4300/2591

Subirats, C. 1989. “Verbal Morphology in the Electronic Dictionary of Spanish”. *Lingvisticae Investigationes* 13.1: 179-201: http://papers.spanishfn.org/public/subirats89.pdf

Subirats, C. 1992. “Verbal, Nominal and Adjectival Inflection in the Electronic Dictionary of Simple Forms of Spanish”. *Lingvisticae Investigationes* 16.2: 345-71: http://papers.spanishfn.org/public/subirats92.pdf

Subirats, C. 1994a. “Sistema de Diccionarios Electrónicos del Español”. *Actas del Congreso de la Lengua Española, Sevilla, 1992*, 316-30. Madrid: Instituto Cervantes: https://cvc.cervantes.es/obref/congresos/sevilla/tecnologias/ponenc_subirats.htm

Subirats, C. 1994b. “La flexión nominal en el Diccionario Electrónico de Formas Compuestas del Español”. *Lingua Franca* 1: 63-9. http://papers.spanishfn.org/public/subirats94.pdf

Subirats, C. 1998. “Automatic Extraction of Textual Information in Spanish”. *Language Design: Journal of Theoretical and Experimental Linguistics* 1: 1-13: http://papers.spanishfn.org/public/subirats94.pdf

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

Subirats, C. y M. Ortega. 2001. “Extracción automática de información de grandes corpus”. En *Lingüística con corpus: catorce aplicaciones sobre el español*, ed. Josse De Kock, 155-75. Salamanca: Ediciones Universidad de Salamanca: http://papers.spanishfn.org/public/subirats_ortega2001.pdf

Subirats, C. 2007. “La lingüística en España”. *Hispanic Issues On Line*, vol. 2, *Estudios Hispánicos. Perspectivas Internacionales*, 169-79: http://hdl.handle.net/11299/182522

Subirats, C. 2010. La teoría conceptual de la metáfora de Gómez Hermosilla. In C. Assunção, G. Fernandes, and M. Loureiro, eds. *Ideias Linguísticas na Península Ibérica.* Münster: Nodus Publikationen, pp. 826–834: http://papers.spanishfn.org/public/La_teoria_de_la_metafora_de_Gomez_Hermosilla.pdf

Subirats, C. 2018. La metáfora y el análisis de la polisemia en la tradición lingüística hispánica. En *Lenguaje figurado y competencia interlingüística (I)*. *Aspectos teóricos*, eds. A. Pamies, M.ª I. Balsas y A. Magdalena, 51-64. Granada: Comares.

Notas

[1] Este trabajo se ha desarrollado en el marco del proyecto de investigación "FrameNet Español (FNE): aplicación al procesamiento semántico automático" (FFI2017-84460-P), financiado por el Ministerio de Economía y Competitividad de España. El proyecto FNE se está desarrollando en colaboración con el proyecto FrameNet de Berkeley, CA (EE.UU), y con la ayuda del Grupo de Investigación de Lexicología y Lingüística de Corpus (GreLIC) de la Universidad de Barcelona.

[2] Disponible en https://corporafromtheweb.org/escow14/ .

[3] Cf. también Casajuana y Rodríguez (1985, 1986).

[4] http://spanishfn.org/corpus

[5] El desarrollo de los diccionarios electrónicos de unidades léxicas simples y locutivas, así como el desarrollo de los programas de generación y consulta de los diccionarios, han sido financiados por el Ministerio de Educación (CAICYT PB85-371, CICYT PB87-780 y PB92- 0635) y por el Ministerio de Obras Públicas y Transportes (TIC90-403).

[6] http://spanishfn.org/dictionary/dictionary-information-lemmas-and-expanded-forms

[7] http://spanishfn.org/dictionary/dictionary-information-set-of-tags

[8] http://spanishfn.org/dictionary/dictionary-information-independent-tags

[9] http://spanishfn.org/dictionary/dictionary-information-inflectional-properties

[10] El desarrollo de los transductores léxicos de locuciones verbales, así como los programas de intersección de autómatas, que forma parte el núcleo del etiquetador léxico, han sido financiados por el Ministerio de Educación (CICYT TIC96-0804 y TIC1999-0753). Queremos mencionar nuestro agradecimiento a †Amelia de Irazazábal por habernos permitido la utilización de una versión electrónica de las Tesauro SPINES, publicado por el desaparecido Centro de Información y Documentación Científica (CINDOC). También queremos agradecer al Prof. Octavio Santana, investigador principal del Grupo de Estructura de Datos y Lingüística Computacional de la Universidad de las Palmas de Gran Canaria por habernos proporcionarnos un diccionario electrónico de nombres propios, que hemos integrado también en nuestro sistema.

[11] Utilizamos abreviaturas en este apartado para facilitar su lectura.

[12] Los nombres de los marcos semánticos, así como los nombres de sus roles semánticos, están en inglés en la web del proyecto FrameNet Español (FNE) http://spanishfn.org/data. Mantenemos los nombres en inglés de los marcos y sus roles en este artículo para facilitar la consulta y la identificación de los datos referentes a la definición de marcos semánticos y la etiquetación de roles semánticos en la web del proyecto FNE.

[13] https://www.globalframenet.org/

[14] http://sato.fm.senshu-u.ac.jp/frameSQL/cxn/CxNeng/cxn00/21colorTag/index.html