Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 111-114 recibido 29-11-12 revisado 08-02-13 aceptado 19-02-13

Nominalizaciones deverbales: denotación y estructura argumental

Deverbal nominalizations: denotation and argument structure

Aina Peris
Universitat de Barcelona
Gran Via de les Corts Catalanes, 585
aina.peris@ub.edu

Resumen: Tesis doctoral en Lingüística Computacional realizada por Aina Peris en la Universitat de Barcelona (UB) bajo la dirección de la Dra. Mariona Taulé (UB) y el Dr. Horacio Rodríguez (Universitat Politècnica de Catalunya). El acto de defensa de la tesis tuvo lugar el viernes 11 de mayo de 2012 ante el tribunal formado por los doctores Piek Vossen (Vrije Universiteit of Amsterdam), Lidia Moreno (Universitat Politècnica de Valencia) y Mª Antònia Martí (UB). La calificación obtenida fue Sobresaliente *Cum Laude* por unanimidad con mención europea.
Palabras clave: Nominalizaciones deverbales, desambigüación automática, etiquetador de roles semánticos

Abstract: Ph.D. Thesis in Computational Linguistics, written by Aina Peris at the University of Barcelona (UB), under the supervision of Dr. Mariona Taulé (UB) and Dr. Horacio Rodríguez (Technical University of Catalonia). The author was examined on friday, 11$^{\text{th}}$ of May 2011, by a commitee formed by the doctors Piek Vossen (Vrije Universiteit of Amsterdam), Lidia Moreno (Technical University of Valencia) and Mª Antònia Martí (UB). The grade obtained was Excellent *Cum Laude* unanimously (with European mention).
Keywords: Deverbal nominalizations, automatic disambiguation, semantic role labeling

1 Introducción

Las nominalizaciones deverbales del español son construcciones lingüísticas que se caracterizan por presentar propiedades propias de los sustantivos pero al mismo tiempo por heredar la estructura argumental de los verbos de los que derivan. Esta dualidad les confiere un notable interés lingüístico porque pueden denotar tanto un estado o el resultado de la acción denotada por el verbo base correspondiente, y también pueden denotar la misma acción o evento que expresa el verbo base, y por tanto, ser paráfrasis de cláusulas oracionales. Por otra parte, son sustantivos que tienen capacidad argumental, es decir, seleccionan argumentos y, en este sentido, es relevante observar los patrones de realización sintáctico-semántica de los argumentos de las nominalizaciones, ya que suponen una manera alternativa de expresar el significado contenido en una oración.

Por lo tanto, dado que las nominalizaciones deverbales pueden expresar el mismo contenido semántico que los predicados verbales y que son construcciones bastante frecuentes en el lenguaje escrito, nos parecía necesario estudiarlas desde el punto de vista de la Lingüística Computacional, contribuyendo, así, a los trabajos que hasta ahora han ido un paso más allá de los verbos en la representación semántica de los textos. Sin embargo, estos trabajos se centran básicamente en las nominalizaciones deverbales del inglés, por lo que también creímos necesario emprender este estudio en español. Veamos ejemplos del tipo de fenómeno con el que tratamos:

(1) [La **construcción** hotelera] ha sido derribada tras la sentencia judicial que así lo ordenaba.

(2) La reflexión fue necesaria para [la posterior **construcción** de la democracia].

En el ejemplo 1 la nominalización *construcción* hace referencia al edificio resultado de las acción del verbo mientras que en el ejemplo 2 se refiere a la acción o evento de

ISSN 1135-5948

construir. En ambos ejemplos, además, las nominalizaciones tienen complementos del nombre (CN) que indican el objeto construido. Por lo tanto, ambos CNs pueden ser asociados a la posición argumental de paciente (arg1-pat).

Además del intrínseco valor lingüístico que tiene el estudio de estas construcciones, también desde un punto de vista del Procesamiento del Lenguaje Natural (PLN) resulta interesante disponer de herramientas y recursos que traten y representen las nominalizaciones deverbales del español, tanto en lo que se refiere a la denotación como a la estructura argumental. Tareas como la resolución de la correferencia o la detección de paráfrasis pueden beneficiarse de una herramienta o un recurso que trate el tipo denotativo de las nominalizaciones, y aplicaciones de extracción de información o sistemas de etiquetado semántico, pueden aprovechar herramientas y recursos que representen la estructura argumental de las nominalizaciones.

2 Organización de la tesis

Esta tesis se estructura en cuatro partes: los antecedentes en el estudio de las nominalizaciones deverbales, la estructura argumental, la denotación y los recursos derivados que las representan. La primera parte introduce el concepto de nominalización deverbal, la importancia de su estudio (Capítulo 1) y ofrece una panorámica de los trabajos realizados, tanto desde el punto de vista lingüístico como computacional (Capítulo 2). La segunda parte centra su atención en la estructura argumental de las nominalizaciones deverbales, tanto el estudio empírico realizado sobre este aspecto (Capítulo 3) como el sistema automático desarrollado (RHN) para la anotación de dicha información en el corpus (Capítulo 4). La tercera parte trata la distinción denotativa entre evento y resultado, tanto el estudio empírico realizado sobre este aspecto (Capítulo 5), como el sistema de clasificación automático desarrollado (ADN) para la anotación de dicha información en el corpus (Capítulo 6) y los experimentos desarrollados con este clasificador (Capítulo 7). En la cuarta parte se describen los recursos lingüísticos derivados de esta investigación, el corpus AnCora-Es enriquecido con la anotación de las nominalizaciones deverbales (Capítulo 8) y el léxico derivado AnCora-Nom (Capítulo 9). Finalmente, en el Capítulo 10 se recogen las conclusiones globales de este trabajo, las aportaciones del mismo y las líneas de trabajo futuro.

3 Contribuciones

Las contribuciones de esta tesis se resumen a continuación:

- Conjunto de criterios lingüísticos que permiten establecer una distinción entre nominalizaciones eventivas y nominalizaciones resultativas del español. Estos criterios se han obtenido a partir del estudio empírico sobre un subconjunto de 100.000 palabras del corpus AnCora-Es, que nos permitió establecer qué criterios de la bibliografía eran válidos para el español y detectar también una serie de criterios nuevos que ayudan a distinguir entre estas dos lecturas denotativas.
- Estudio lingüístico de la estructura argumental de las nominalizaciones deverbales, es decir, de los distintos patrones de realización sintáctica de los argumentos de estos predicados. A partir de las observaciones iniciales del estudio empírico y su implementación en las reglas de proyección de RHN, hemos obtenido nuevas e interesantes observaciones lingüísticas.
- Construcción del ADN-Classifier, un sistema de clasificación automática de nominalizaciones deverbales según su denotación.
- Implementación de RHN, conjunto de reglas heurísticas que tienen en cuenta la información del léxico AnCora-Verb y a partir de las cuales se ha anotado automáticamente la estructura argumental de las nominalizaciones deverbales del corpus AnCora-Es.
- Enriquecimiento del corpus AnCora-Es con la validación manual de los procesos automáticos de anotación (denotación y estructura argumental) de las nominalizaciones deverbales.
- Creación de AnCora-Nom, un léxico de 1.655 nominalizaciones deverbales en español.

Estas contribuciones se clasifican en tres grandes grupos que detallamos en las siguientes subsecciones: 1) caracterización lingüística de las nominalizaciones deverbales (denotación y estructura argumental); 2) herra-

mientas computacionales para tratar estos dos aspectos de las nominalizaciones deverbales automáticamente, y 3) creación de recursos lingüísticos que representan estas construcciones lingüísticas.

3.1 Caracterización Lingüística

En relación a la distinción denotativa entre evento y resultado de las nominalizaciones deverbales, se han definido una serie de criterios que permiten identificar una de las dos lecturas (Peris y Taulé, 2009). Se analizó si los criterios establecidos en la bibliografía para el inglés eran válidos para el español. Entre los criterios evaluados, los más relevantes para el español son: 1) la clase semántica del verbo del que deriva la nominalización; 2) su capacidad de pluralización; 3) los tipos de determinantes; 4) la preposición que introduce al complemento agentivo; y 5) la presencia obligatoria de un argumento interno (arg1). Estos rasgos se han representado como atributos en las entradas léxicas nominales del léxico AnCora-Nom. Además, el estudio lingüístico llevado a cabo nos permitió encontrar criterios nuevos para la identificación, especialmente, de las nominalizaciones eventivas (puesto que con los criterios de la bibliografía no eran todas identificables): los selectores y el criterio de la paráfrasis. Los selectores pueden ser de dos tipos: (i) selectores externos, elementos que desde fuera del SN indican la denotación de la nominalización (la preposición *durante* por ejemplo); y (ii) selectores internos, prefijos de la nominalización que indican un tipo concreto de denotación (el prefijo *re-* reiterativo se aplica a acciones, por lo tanto las nominalizaciones que lo emplean son eventivas). En cuanto al criterio de la paráfrasis, si un SN cuyo núcleo es una nominalización y puede parafrasearse por una oración con el verbo base, se considera que es una nominalización eventiva.

Respecto a la estructura argumental de las nominalizaciones deverbales, se realizó un estudio lingüístico basado en corpus que permitió definir una serie de patrones de realización sintáctico-semántica que luego se implementaron en la herramienta de etiquetado semántico RHN. A partir del análisis de errores de esta herramienta, hemos podido establecer algunas características de la estructura argumental de las nominalizaciones deverbales. En primer lugar, la hipótesis de trabajo inicial de que las nominalizaciones deverbales heredan la estructura argumental del verbo base correspondiente se confirma ya que RHN consigue un F1 del 77 % y se basa principalmente en la información contenida en el léxico AnCora-Verb. En segundo lugar, se muestra que el orden de los constituyentes de los SNs de núcleo deverbal es más libre que el de los complementos verbales, y que hasta cierto punto depende del contexto. En tercer lugar, cabe destacar que los argumentos de las nominalizaciones están marcados por un alto grado de opcionalidad. Esto afecta especialmente al arg0, que no aparece realizado en numerosas ocasiones. Finalmente, detallamos las características argumentales de los constituyentes que pueden ser complementos de las nominalizaciones deverbales: los SAs no relacionales, los Sadv y las oraciones subordinadas no son argumentos en un SN de núcleo deverbal. Respecto a los SNs complementos de nominalizaciones deverbales, se puede establecer que aquellos anotados como una entidad con nombre locativa o temporal reciben la etiqueta de adjunto locativo (argM-loc) o temporal (argM-tmp. Respecto a los SPs, aquellos introducidos por una preposición especifica como *durante, tras, para* etc., se corrobora que dichas preposiciones apuntan a una determinada etiqueta argumental. También se ha comprobado que las preposiciones regidas de los complementos de régimen verbal no siempre se mantienen en el dominio nominal. En cuanto a los SAs relacionales, encontramos un 45 % que no eran argumentales. Parece confirmarse que que los adjetivos relaciones están sometidos al fenómeno de la co-ocurrencia léxica, es decir, que se anotan como argumentales o no argumentales dependiendo del nombre al que complementen. Los determinantes posesivos, por su parte, se interpretan mayoritariamente como el argumento correspondiente al sujeto verbal.

3.2 Sistemas automáticos

A continuación describimos las dos herramientas computacionales desarrolladas en esta tesis: el sistema RHN y el clasificador ADN.

El sistema de RHN está formado por 107 reglas heurísticas, cuyo objetivo es ligar un constituyente del SN del núcleo deverbal con un argumento y papel temático usando el léxico AnCora-Verb, el corpus AnCora-Es y una lista predefinida de adjetivos relacionales. Estas reglas se organizan en un forma-

to de lista de decisión y se aplican a un SN constituido por una nominalización (N) y un contexto que puede ser de uno, dos o tres constituyentes. Cada regla satisface una condición, una combinación lógica de predicados sobre N o sobre el contexto, y así, se asigna una etiqueta semántica. Hay dos tipos de reglas: (i) catorce reglas generales basadas en la información lingüística de AnCora-Es, y (ii) noventa y tres reglas específicas que también tienen en cuenta la información contenida en el léxico AnCora-Verb. RHN logra un 77 % de F1 (Peris y Taulé, 2011b).

El clasificador ADN clasifica automáticamente las nominalizaciones deverbales del español según su denotación sea de tipo eventivo, resultativo o subespecificado, o formen parte en construcciones lexicalizadas. Se desarrollaron una serie de experimentos para poner a prueba los diferentes modelos de clasificación de ADN y en diferentes escenarios y se han obtenido buenos resultados. Los modelos basados en rasgos del léxico AnCora-Nom superan a los modelos basados en rasgos del corpus. De la misma manera que los modelos que trabajan a nivel de sentido superan a los que trabajan a nivel de lema. ADN logra una mayor precisión en la detección de nominalizaciones resultativas que eventivas.(Peris, Taulé, y Rodríguez, 2009; Peris et al., 2010; Peris, Taulé, y Rodríguez, 2012)

3.3 Recursos Léxicos

Esta tesis ha dado lugar a dos nuevos recursos: se ha enriquecido la anotación del corpus AnCora-Es (Peris, Taulé, y Rodríguez, 2010) con la anotación de 23.431 ocurrencias de nominalizaciones deverbales con su denotación y su estructura argumental y se ha creado el léxico AnCora-Nom (Peris y Taulé, 2011a), con 1.655 entradas léxicas de nominalizaciones deverbales en español.

El enriquecimiento del corpus AnCora-Es se ha llevado a cabo en dos etapas: 1) se realizaron dos procesos automáticos de manera independiente, uno para la anotación de la denotación (con ADN) y otra para la estructura argumental (con RHN) y 2) se validaron manualmente estos dos tipos de informaciones. El corpus AnCora-Es es el único corpus del español anotado con este tipo de información.

El léxico AnCora-Nom, por su parte, fue creado automáticamente a partir de la información contenida en el corpus AnCora-Es. Incluye todos los lemas de las nominalizaciones del corpus con sus denotaciones y sus posibilidades de combinatoria de la estructura argumental.

Bibliografía

Peris, Aina y Mariona Taulé. 2009. Evaluación de los criterios lingüísticos para la distinción evento y resultado en los sustantivos deverbales. En *Proceedings of the 1st International Conference on Corpus Linguistics*, páginas 596–611, Murcia, España.

Peris, Aina y Mariona Taulé. 2011a. AnCora-Nom: A Spanish Lexicon of Deverbal Nominalizations. *Procesamiento del Lenguaje Natural.*, 46:11–19.

Peris, Aina y Mariona Taulé. 2011b. Annotating the argument structure of deverbal nominalizations in Spanish. doi: 10.1007/s10579-011-9172-x. *Language Resources and Evaluation.*

Peris, Aina, Mariona Taulé, Gemma Boleda, y Horacio Rodríguez. 2010. ADN-Classifier: Automatically Assigning Denotation Types to Nominalizations. En *Proceedings of the Language Resources and Evaluation Conference*, páginas 1422–1428, Valleta, Malta.

Peris, Aina, Mariona Taulé, y Horacio Rodríguez. 2009. Hacia un sistema de clasificación automática de sustantivos deverbales. *Procesamiento del Lenguaje Natural.*, 43:23–31.

Peris, Aina, Mariona Taulé, y Horacio Rodríguez. 2010. Semantic Annotation of Deverbal Nominalizations in the Spanish AnCora Corpus. En *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories*, páginas 187–198, Tartu, Estonia.

Peris, Aina, Mariona Taulé, y Horacio Rodríguez. 2012. Empirical methods for the study of denotation in nominalizations in Spanish. *Computational Linguistics*, 38(4):827–865.